Supplement of Hydrol. Earth Syst. Sci., 29, 6811–6828, 2025
https://doi.org/10.5194/hess-29-6811-2025-supplement

Hydrology and Earth System Sciences
Open Access

Supplement of

From RNNs to Transformers: benchmarking deep learning architectures for hydrologic prediction

Jiangtao Liu et al.

Correspondence to: Jiangtao Liu (jql6620@psu.edu)

S1. Model descriptions

CARDformer is a Transformer-based deep learning model for multivariate time series forecasting. It employs channel-aligned attention to model both temporal dependencies and inter-variable correlations, overcoming the limitations of channel-independent approaches. A token blend module combines information at multiple temporal resolutions to capture both local details and extended temporal dependencies.

Crossformer is a Transformer variant developed to capture dependencies across both time and variables in multivariate time series. It uses Dimension-Segment-Wise embedding to form a 2D time–variable representation and a Two-Stage Attention mechanism to model temporal patterns within each variable and relationships among variables, enhanced by a Hierarchical Encoder-Decoder for multi-scale forecasting.

ETSformer integrates exponential smoothing techniques into the Transformer framework to enhance time series forecasting. It employs Exponential Smoothing Attention (ESA) and Frequency Attention (FA) mechanisms. ESA prioritizes recent observations to capture trends, while FA captures repeating seasonal patterns using Fourier transforms. This structured attention enables ETSformer to decompose forecasts into interpretable trend and seasonal components, thereby improving both reliability and transparency of predictions.

Informer is designed to address the challenges in forecasting long time series. It employs a ProbSparse self-attention mechanism that selectively attends to the most important parts of the input, reducing computational complexity and memory usage compared to traditional self-attention. Informer also introduces self-attention distilling, which progressively shortens input sequences layer by layer while preserving essential information, alleviating memory limitations. Furthermore, its generative decoder produces complete forecast sequences in a single forward pass, substantially increasing inference speed and minimizing cumulative prediction errors.

iTransformer is a Transformer-based method that models multivariate time series by inverting the data processing dimensions. Specifically, rather than embedding all variables at each time step into a single temporal token, it treats the entire historical series of each individual variable independently as a token. This inversion enables the self-attention mechanism to capture correlations among different variables, while a shared feed-forward network separately learns representations from the full historical context of each variable.

The Non-stationary Transformer addresses non-stationarity, a common characteristic of real-world time series where statistical properties change over time. It employs a preprocessing step called Series Stationarization, which normalizes input data to improve predictability and ease model training. However, since this process may inadvertently eliminate valuable non-stationary dynamics, the model introduces De-stationary Attention, an attention mechanism designed to recover and reintegrate the intrinsic non-stationary temporal dependencies into the prediction.

PatchTST employs a patch-based approach inspired by techniques from fields such as computer vision, segmenting a long time series into smaller segments (patches). Each patch is then transformed into an input token for the Transformer model, which applies global self-attention across these patch tokens. This reduces the effective sequence length, allowing modeling of longer historical contexts.

Pyraformer introduces a pyramidal attention structure that models time series data at multiple resolutions, capturing local temporal dependencies through intra-scale connections at fine-grained scales and long-range dependencies via inter-scale connections at coarser scales. The method utilizes a hierarchical pyramidal attention module (PAM), combining inter-scale tree structures with intra-scale neighboring connections. This hierarchical approach shortens the path connecting distant time points, enabling modeling of long-range temporal patterns while maintaining linear computational complexity.

Reformer enhances traditional Transformers to handle long sequences through two key innovations. First, it introduces Locality-Sensitive Hashing (LSH) attention, which replaces standard dot-product attention by identifying the most relevant keys for each query, reducing computational complexity from quadratic (proportional to the square of sequence length) to near-linear (proportional to the sequence length multiplied by its logarithm). Second, Reformer employs reversible residual layers, which reduce memory usage by reconstructing intermediate activations during the backward pass rather than storing them at each layer.

TimesNet transforms one-dimensional time series data into two-dimensional representations, capturing both intra- and interperiod-variations. Leveraging the inherent multi-periodicity in real-world time series, TimesNet adaptively learns and identifies significant periodicities, reshaping data into structured 2D tensors. By doing so, TimesNet utilizes convolutional neural network modules (TimesBlock) to model complex temporal dependencies.

The standard Transformer is the foundational model upon which many subsequent variants were built. It models sequences through attention mechanisms, specifically multi-head self-attention and encoder-decoder attention, eliminating recurrent and convolutional structures. This attention mechanism allows the Transformer to simultaneously analyze global dependencies between all sequence elements, facilitating parallel computation and faster training. Additionally, positional encodings (to capture sequence order), residual connections, and layer normalization (to stabilize training) further enhance its effectiveness in learning from sequential data.

The Long Short-Term Memory (LSTM) network is a type of recurrent neural network (RNN) designed to handle data sequences where long-term dependencies are crucial. LSTMs use gating mechanisms (input, forget, and output gates) that dynamically control how information is maintained, discarded, or updated at each step in the sequence.

S2. Prompt example

You are an expert in streamflow forecasting, focusing on a basin with an area of {area} square kilometers located at latitude {lat} and longitude {lon}. Given the historical time series data: {string_data}, forecast the next {pred_len} values of the time series. Please ensure the output contains exactly {pred_len} forecasted values, and exclude any Python code and any text.

Table S1. Performance comparison of 11 models on five datasets for regression tasks: CAMELS (conterminous United States (CONUS)-scale streamflow), ISMN (global-scale soil moisture), SWE (snow water equivalent in the western United States), DO (dissolved oxygen at the CONUS–scale), and GLOBAL (global-scale streamflow). For each dataset, six evaluation metrics are presented: Kling–Gupta Efficiency (KGE), Nash–Sutcliffe Efficiency (NSE), unbiased Root Mean Square Error (ubRMSE), Low-Flow Volume Bias (FLV), High-Flow Volume Bias (FHV), and Coefficient of Determination (R^2). Each row corresponds to one model. Bold text indicates the best value for each metric.

CAMELS						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.42	0.34	2.01	-300.24	-38.61	0.39
LSTM	**0.80**	**0.80**	**1.10**	**3.44**	-10.32	**0.82**
Informer	0.63	0.69	1.33	-169.44	-14.89	0.74
Reformer	0.62	0.66	1.39	-226.98	-13.31	0.72
iTransformer	0.37	0.37	1.89	-147.66	-36.00	0.45
Transformer	0.60	0.70	1.29	-230.86	**-8.77**	0.76
ETSformer	0.68	0.67	1.41	-32.84	-19.19	0.70
Pyraformer	0.61	0.69	1.33	-98.85	-20.57	0.74
Crossformer	0.73	0.72	1.29	-42.39	-14.35	0.75
CARDformer	0.33	0.29	2.04	-183.48	-55.27	0.34
Non-stationary Transformer	0.61	0.72	1.24	-172.87	-15.52	0.77
Global Streamflow						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.24	0.14	1.53	-50.88	-31.02	0.36
LSTM	**0.75**	**0.71**	**0.88**	**-1.24**	-8.26	**0.76**
Informer	0.64	0.62	1.00	-8.55	-15.93	0.70
Reformer	0.62	0.57	1.10	-32.12	-16.40	0.65
iTransformer	0.32	0.16	1.58	-131.31	-12.34	0.37
Transformer	0.16	0.30	1.06	-355.33	-20.79	0.69
ETSformer	0.60	0.55	1.11	-45.96	-16.03	0.64
Pyraformer	0.62	0.60	1.03	9.04	-19.35	0.68
Crossformer	0.52	0.50	1.17	-30.92	-20.91	0.61
CARDformer	0.26	0.22	1.48	-58.71	-41.54	0.34
Non-stationary Transformer	0.57	0.57	1.03	-142.73	**-4.10**	0.71

Global Soil Moisture						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.31	-0.09	0.06	66.15	-11.24	0.55
LSTM	**0.71**	**0.61**	**0.04**	5.16	-3.13	**0.76**
Informer	0.64	0.47	0.05	**0.99**	-5.42	0.66
Reformer	0.69	0.54	0.04	10.86	-1.51	0.70
iTransformer	0.27	-0.45	0.07	7.32	0.93	0.31
Transformer	0.65	0.47	0.05	3.43	**0.22**	0.68
ETSformer	0.59	0.35	0.05	-7.77	0.27	0.63
Pyraformer	0.70	0.56	0.04	-0.83	-1.28	0.72
Crossformer	0.67	0.51	0.04	-7.80	-0.28	0.70
CARDformer	0.28	-0.16	0.06	48.95	-10.29	0.40
Non-stationary Transformer	0.66	0.48	0.05	13.13	-1.00	0.69
Snow Water Equivalent						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.53	0.60	0.11	-485734.04	-27.05	0.72
LSTM	0.87	**0.94**	**0.04**	**-18333.26**	-4.77	**0.96**
Informer	0.86	0.93	0.04	-50722.84	2.35	0.95
Reformer	0.80	0.85	0.06	-45488.46	-10.19	0.88
iTransformer	0.65	0.63	0.10	-158075.15	4.90	0.73
Transformer	0.85	0.92	0.04	-107073.25	1.86	0.95
ETSformer	0.84	0.91	0.05	-248400.38	-6.03	0.93
Pyraformer	0.86	0.89	0.05	-131439.12	-3.96	0.91
Crossformer	0.84	0.93	0.04	-140254.87	7.65	0.96
CARDformer	0.67	0.73	0.08	-204616.61	1.95	0.83
Non-stationary Transformer	**0.88**	0.94	0.04	-78699.35	**0.39**	0.95
DO						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.63	0.34	1.26	1.42	-14.11	0.68
LSTM	**0.70**	**0.53**	1.17	1.64	-10.08	0.70

Informer	0.70	0.51	1.20	3.43	-10.91	0.70
Reformer	0.69	0.46	1.21	2.27	-9.23	0.68
iTransformer	0.55	0.08	1.57	1.28	**-3.78**	0.45
Transformer	0.69	0.47	1.24	1.95	-10.05	0.68
ETSformer	0.68	0.46	**1.17**	3.28	-9.09	**0.70**
Pyraformer	0.67	0.46	1.22	3.16	-11.36	0.69
Crossformer	0.67	0.43	1.31	5.09	-11.04	0.67
CARDformer	0.57	0.20	1.43	**0.15**	-11.99	0.56
Non-stationary Transformer	0.69	0.40	1.23	-0.87	-5.74	0.67

Table S2. Performance comparison of 11 models under time-lagged forecasting scenarios using the CAMELS dataset. Results are reported for time lags of 1, 7, 30, and 60 days. For each time lag, six evaluation metrics are presented: Kling–Gupta Efficiency (KGE), Nash–Sutcliffe Efficiency (NSE), unbiased Root Mean Square Error (ubRMSE), Low-Flow Volume Bias (FLV), High-Flow Volume Bias (FHV), and Coefficient of Determination (R^2). Each row corresponds to one model. Bold text indicates the best value for each metric.

1 day						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.60	0.53	1.60	-254.22	-25.44	0.57
LSTM	**0.89**	**0.87**	**0.91**	-19.29	-4.15	**0.88**
Informer	0.61	0.75	1.07	-326.63	-5.87	0.83
Reformer	0.65	0.78	1.04	-248.23	-4.29	0.84
iTransformer	0.61	0.66	1.32	-228.32	-14.43	0.71
Transformer	0.65	0.77	1.07	-275.23	-9.05	0.83
ETSformer	0.81	0.79	1.09	-59.34	-6.59	0.82
Pyraformer	0.81	0.81	1.03	**-3.39**	-8.89	0.83
Crossformer	0.80	0.77	1.18	-33.48	-2.06	0.80
CARDformer	0.44	0.36	1.92	-209.11	-44.98	0.41
Non-stationary Transformer	0.79	0.82	1.02	-190.29	**-0.42**	0.85
7 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.50	0.38	1.95	-281.37	-36.79	0.41
LSTM	**0.85**	**0.83**	**1.03**	-4.60	-7.07	**0.84**
Informer	0.70	0.72	1.31	-80.25	-16.58	0.75
Reformer	0.78	0.76	1.24	-65.74	-13.23	0.78
iTransformer	0.68	0.61	1.52	-110.47	-14.38	0.63
Transformer	0.73	0.76	1.18	-187.64	-10.61	0.80
ETSformer	0.66	0.57	1.59	-24.63	-21.58	0.60
Pyraformer	0.77	0.75	1.25	**1.60**	-14.88	0.77
Crossformer	0.83	0.78	1.15	-29.71	**-3.33**	0.80
CARDformer	0.51	0.43	1.86	-119.18	-33.06	0.46
Non-stationary Transformer	0.81	0.79	1.13	-81.88	-7.61	0.81
30 days						

Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.40	0.30	2.06	-367.82	-40.11	0.36
LSTM	**0.83**	**0.81**	**1.07**	**-2.94**	-9.04	**0.83**
Informer	0.68	0.63	1.50	-25.01	-20.36	0.66
Reformer	0.72	0.69	1.37	-32.96	-14.06	0.72
iTransformer	0.64	0.53	1.67	-83.63	-13.03	0.57
Transformer	0.77	0.76	1.20	-28.79	-14.18	0.78
ETSformer	0.59	0.50	1.72	-84.85	-32.45	0.52
Pyraformer	0.76	0.74	1.26	27.90	-16.28	0.76
Crossformer	0.78	0.77	1.15	-57.35	-6.76	0.80
CARDformer	0.50	0.48	1.78	-87.59	-34.83	0.53
Non-stationary Transformer	0.82	0.77	1.20	-8.03	**-5.22**	0.79
60 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.43	0.34	2.02	-306.10	-37.61	0.40
LSTM	**0.82**	**0.81**	**1.09**	5.09	-11.00	**0.82**
Informer	0.67	0.60	1.57	-18.52	-20.87	0.63
Reformer	0.73	0.69	1.40	**4.03**	-14.40	0.71
iTransformer	0.63	0.50	1.73	-63.78	-11.65	0.55
Transformer	0.76	0.75	1.24	-15.63	-15.32	0.77
ETSformer	0.52	0.44	1.83	-123.40	-40.31	0.47
Pyraformer	0.73	0.71	1.32	55.49	-17.61	0.73
Crossformer	0.80	0.79	1.13	-30.50	-12.30	0.81
CARDformer	0.60	0.55	1.68	-91.73	-20.91	0.59
Non-stationary Transformer	0.81	0.75	1.28	-15.85	**0.34**	0.78

Table S3. Performance comparison of 13 models under autoregressive modeling scenarios using the CAMELS dataset. Results are reported for forecasting horizons of 1, 7, 30, and 60 days. For each horizon, six evaluation metrics are presented: Kling–Gupta Efficiency (KGE), Nash–Sutcliffe Efficiency (NSE), unbiased Root Mean Square Error (ubRMSE), Low-Flow Volume Bias (FLV), High-Flow Volume Bias (FHV), and Coefficient of Determination (R^2). Each row corresponds to one model. Bold text indicates the best value for each metric.

1 day						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.59	0.52	1.69	80.83	-28.33	0.54
LSTM	0.64	**0.58**	**1.57**	5.44	-27.26	**0.59**
Informer	0.59	0.53	1.61	-140.01	-24.41	0.58
Reformer	0.65	0.55	1.62	-94.83	**-19.16**	0.58
iTransformer	0.61	0.51	1.69	-15.25	-24.41	0.53
Transformer	0.61	0.54	1.62	-144.65	-20.28	0.57
ETSformer	0.52	0.47	1.67	-175.38	-25.16	0.52
Pyraformer	**0.65**	0.57	1.60	-29.59	-22.19	0.58
Crossformer	0.65	0.57	1.61	-28.62	-21.65	0.58
CARDformer	0.47	0.37	1.95	-3.41	-37.72	0.38
Non-stationary Transformer	0.65	0.55	1.65	-8.79	-23.92	0.55
PatchTST	0.64	0.53	1.65	13.24	-23.51	0.54
TimesNet	0.62	0.49	1.71	**1.24**	-23.19	0.51
7 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.27	0.19	2.20	132.74	-47.39	0.21
LSTM	0.15	0.10	2.35	35.29	-59.29	0.12
Informer	0.30	0.15	2.21	-74.95	-44.74	0.20
Reformer	0.29	0.18	2.17	-29.59	-49.18	0.21
iTransformer	0.29	0.16	2.24	-12.88	-43.11	0.20
Transformer	0.27	0.03	2.45	-109.68	**-31.85**	0.15
ETSformer	0.17	0.11	2.21	-231.34	-50.15	0.19
Pyraformer	**0.32**	**0.21**	**2.14**	27.96	-45.81	**0.23**
Crossformer	0.30	0.19	2.18	75.41	-45.03	0.23
CARDformer	0.25	0.11	2.30	8.50	-46.78	0.15

Non-stationary Transformer	0.27	0.11	2.29	**0.12**	-41.51	0.16
PatchTST	0.31	0.19	2.21	8.35	-44.96	0.22
TimesNet	0.30	0.09	2.31	-16.78	-36.59	0.17
30 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.06	0.06	2.44	242.67	-59.59	0.09
LSTM	-0.03	-0.01	2.54	62.28	-63.72	0.03
Informer	0.11	0.00	2.53	157.36	**-43.07**	0.08
Reformer	0.15	0.06	**2.40**	47.14	-53.37	**0.11**
iTransformer	0.11	0.02	2.48	**-0.46**	-51.84	0.08
Transformer	0.13	-0.05	2.62	-27.12	-39.46	0.07
ETSformer	0.11	0.02	2.46	-118.55	-56.95	0.09
Pyraformer	**0.15**	**0.06**	2.39	57.93	-53.11	0.11
Crossformer	0.11	0.04	2.45	49.02	-54.57	0.09
CARDformer	0.07	0.02	2.46	13.80	-59.74	0.07
Non-stationary Transformer	0.11	-0.01	2.51	15.25	-46.01	0.07
PatchTST	0.12	0.02	2.47	19.47	-54.44	0.08
TimesNet	0.18	-0.10	2.68	-14.10	-33.36	0.06
60 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	-0.06	0.01	2.49	297.49	-64.44	0.05
LSTM	-0.19	-0.07	2.58	85.40	-70.25	0.01
Informer	0.04	-0.03	2.56	175.41	-46.74	0.05
Reformer	0.07	0.01	2.47	38.11	-57.62	0.07
iTransformer	0.03	0.00	2.53	0.32	-57.63	0.05
Transformer	0.06	-0.04	2.63	37.50	-46.60	0.05
ETSformer	0.04	0.00	2.51	-4.35	-59.35	0.06
Pyraformer	0.06	0.02	2.46	89.20	-56.93	0.07
Crossformer	-0.02	0.00	2.52	85.80	-59.91	0.05
CARDformer	-0.02	-0.02	2.57	21.80	-61.59	0.03

Non-stationary Transformer	0.05	0.00	2.53	75.83	-52.74	0.06
PatchTST	0.01	-0.02	2.56	22.92	-58.62	0.04
TimesNet	0.10	-0.13	2.75	-8.11	-36.82	0.03

Table S4. Performance comparison of 13 models under autoregressive modeling scenarios using the CAMELS dataset with auxiliary data inputs (static attributes). Results are reported for forecasting horizons of 1, 7, 30, and 60 days. For each horizon, six evaluation metrics are presented: Kling–Gupta Efficiency (KGE), Nash–Sutcliffe Efficiency (NSE), unbiased Root Mean Square Error (ubRMSE), Low-Flow Volume Bias (FLV), High-Flow Volume Bias (FHV), and Coefficient of Determination (R^2). Each row corresponds to one model. Bold text indicates the best value for each metric.

1 day						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.68	0.59	1.54	-92.10	-23.90	0.60
LSTM	**0.81**	**0.78**	**1.11**	-14.86	-13.41	**0.79**
Informer	0.67	0.70	1.25	-115.75	-15.94	0.75
Reformer	0.64	0.72	1.21	-199.18	-11.50	0.76
iTransformer	0.62	0.56	1.55	-77.58	-17.91	0.59
Transformer	0.56	0.69	1.19	-259.77	-18.39	0.77
ETSformer	0.67	0.70	1.27	-79.62	-21.77	0.74
Pyraformer	0.74	0.73	1.25	48.02	-15.95	0.75
Crossformer	0.75	0.70	1.32	-53.81	-14.06	0.73
CARDformer	0.46	0.41	1.82	**-9.27**	-36.21	0.44
Non-stationary Transformer	0.74	0.73	1.23	-95.42	**-11.21**	0.75
PatchTST	0.63	0.57	1.56	-46.65	-22.44	0.59
TimesNet	0.66	0.64	1.43	-18.35	-24.34	0.66
7 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.33	0.23	2.13	-18.54	**-47.13**	0.25
LSTM	0.15	0.05	2.36	36.57	-50.61	0.11
Informer	0.28	0.24	2.07	-55.22	-49.85	0.28
Reformer	0.29	0.26	**2.00**	-46.98	-50.68	0.30
iTransformer	0.18	0.10	2.27	-82.36	-48.24	0.16
Transformer	0.28	0.23	2.09	-36.33	-48.32	0.26
ETSformer	0.31	0.24	2.04	-68.33	-49.50	0.27
Pyraformer	0.33	**0.29**	**2.01**	125.47	-48.21	**0.31**
Crossformer	**0.34**	0.29	2.03	**8.13**	-50.02	0.31
CARDformer	0.19	0.13	2.25	-17.87	-53.23	0.16

Non-stationary Transformer	0.21	0.19	2.12	-107.57	-51.97	0.23
PatchTST	0.17	0.14	2.22	-82.51	-54.36	0.18
TimesNet	0.22	0.18	2.17	-33.24	-53.83	0.21
30 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	-0.02	0.04	2.44	55.04	-70.75	0.07
LSTM	0.02	0.00	**2.47**	55.50	-55.59	0.06
Informer	0.14	0.07	2.37	13.51	-56.26	0.12
Reformer	**0.15**	0.10	2.30	95.32	-55.16	0.13
iTransformer	0.11	-0.01	2.50	-41.16	**-45.12**	0.07
Transformer	0.07	0.00	2.43	-77.34	-50.49	0.08
ETSformer	0.07	0.09	**2.30**	-39.08	-67.10	0.13
Pyraformer	0.13	**0.11**	2.31	26.53	-59.21	**0.14**
Crossformer	0.13	0.10	2.31	24.51	-60.17	0.13
CARDformer	0.02	0.02	2.44	**-5.39**	-61.80	0.06
Non-stationary Transformer	0.08	-0.03	2.45	-116.25	-48.17	0.08
PatchTST	0.07	0.02	2.44	-63.89	-60.47	0.08
TimesNet	0.07	0.02	2.41	-46.50	-59.04	0.08
60 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.00	0.06	2.44	82.13	-71.27	0.08
LSTM	-0.04	-0.02	2.48	54.07	-59.15	0.04
Informer	0.05	0.07	2.36	177.37	-61.66	0.09
Reformer	**0.07**	**0.08**	2.35	**191.37**	-65.14	**0.10**
iTransformer	-0.03	-0.06	2.57	-23.30	-55.12	0.03
Transformer	0.01	-0.04	2.52	-15.08	-53.36	0.05
ETSformer	0.05	0.06	2.34	-33.72	-68.92	0.10
Pyraformer	0.10	0.09	**2.33**	98.91	-63.74	0.11
Crossformer	0.07	0.06	2.40	33.65	-62.76	0.09
CARDformer	-0.04	0.01	2.48	37.33	-65.51	0.05

Non-stationary Transformer	0.03	-0.03	2.47	-111.29	**-51.37**	0.06
PatchTST	-0.01	0.00	2.51	**-2.62**	-64.39	0.05
TimesNet	-0.11	-0.05	2.52	-90.09	-66.94	0.03

Table S5. Performance comparison of 11 models using spatial cross-validation with the CAMELS dataset. Six evaluation metrics are presented: Kling–Gupta Efficiency (KGE), Nash–Sutcliffe Efficiency (NSE), unbiased Root Mean Square Error (ubRMSE), Low-Flow Volume Bias (FLV), High-Flow Volume Bias (FHV), and Coefficient of Determination (R^2). Each row corresponds to one model. Bold text indicates the best value for each metric.

Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
DLinear	0.43	0.33	2.01	-288.31	-37.27	0.39
LSTM	0.62	**0.67**	**1.38**	-32.72	-14.13	**0.74**
Informer	0.54	0.60	1.48	-136.05	-13.15	0.70
Reformer	0.56	0.61	1.49	-160.68	-15.16	0.70
iTransformer	0.32	0.34	1.97	-111.84	-40.96	0.43
Transformer	0.49	0.61	1.43	-202.44	-9.68	0.73
ETSformer	0.60	0.62	1.47	**-9.24**	-16.82	0.68
Pyraformer	0.60	0.64	1.41	-100.68	-16.23	0.72
Crossformer	**0.63**	0.62	1.49	-42.42	**-6.76**	0.71
CARDformer	0.34	0.29	2.05	-153.89	-48.91	0.35
Non-stationary Transformer	0.58	0.61	1.49	-130.38	-6.79	0.70

Table S6. Performance comparison of Large Language Models (LLMs) and Time Series Attention-based Models (TSAMs) for runoff forecasting under zero-shot forecasting scenarios. Results are reported for forecasting horizons of 7, 30, and 60 days. For each horizon, six evaluation metrics are presented: Kling–Gupta Efficiency (KGE), Nash–Sutcliffe Efficiency (NSE), unbiased Root Mean Square Error (ubRMSE), Low-Flow Volume Bias (FLV), High-Flow Volume Bias (FHV), and Coefficient of Determination (R^2). Each row corresponds to one model. Bold text indicates the best value for each metric.

Forecast 7 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
gemini-1.0-pro-latest	0.15	-0.01	**0.13**	-4.04	-20.69	0.03
gpt-3.5-turbo	0.53	0.00	0.19	-4.39	**-9.92**	0.31
gpt-4-turbo	0.46	0.26	0.21	-7.71	-16.40	0.40
Lag_Llama	0.39	0.22	0.15	-15.90	-47.26	0.39
LSTM (trained)	0.50	0.43	0.19	-8.58	-39.30	0.43
Meta-Llama-3-70B-Instruct	0.54	0.24	0.19	-6.27	-23.12	0.36
Meta-Llama-3-8B-Instruct	0.12	-0.02	0.16	-5.58	-35.61	0.10
TimeGPT	**0.68**	0.46	0.14	**-0.74**	-26.57	0.51
TTM	0.57	**0.52**	**0.13**	6.66	-30.35	**0.53**
Forecast 30 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
gemini-1.0-pro-latest	-0.05	-0.02	**0.16**	**-0.52**	-36.31	0.01
gpt-3.5-turbo	0.11	-0.18	0.29	-12.88	**-23.76**	0.01
gpt-4-turbo	0.01	-0.13	0.18	-26.19	-37.70	0.09
Lag_Llama	0.19	-0.03	0.18	-29.51	-60.54	0.10
LSTM (trained)	0.15	0.06	0.27	19.31	-55.63	0.14
Meta-Llama-3-70B-Instruct	0.09	-0.02	0.17	-22.51	-39.50	0.21
Meta-Llama-3-8B-Instruct	0.12	-0.22	0.23	-30.88	-40.70	0.04
TimeGPT	**0.33**	0.08	0.17	-5.10	-45.27	0.17
TTM	0.18	**0.20**	0.17	20.37	-66.38	**0.25**
Forecast 60 days						
Model Name	KGE	NSE	ubRMSE	FLV	FHV	R^2
gemini-1.0-pro-latest	0.02	-0.17	0.66	-11.16	-50.91	0.11
gpt-3.5-turbo	-0.04	-0.84	0.26	-48.44	3.28	0.09
gpt-4-turbo	-0.07	-0.48	0.29	-52.04	-52.25	0.01
Lag_Llama	0.01	-0.05	0.20	-26.88	-59.64	0.07

LSTM (trained)	-0.11	-0.02	0.30	36.71	-54.35	0.01
Meta-Llama-3-70B-Instruct	-0.31	-0.50	0.22	-39.16	-64.04	0.02
Meta-Llama-3-8B-Instruct	-0.08	-1.19	0.25	-31.51	8.90	0.00
TimeGPT	0.17	0.06	0.22	-9.96	-63.67	0.09
TTM	0.02	0.01	0.19	26.23	-68.46	0.08

Table S7. Comparison of model training time and energy consumption under identical training conditions (30 epochs). For each model, training time and carbon dioxide equivalent emissions (kg CO_2. eq) are reported.

	Hours (30 epochs)	Energy consumption (kg CO_2 eq.)
DLinear:	0.4	0.07
iTransformer:	0.7	0.12
LSTM:	1.0	0.17
Informer:	2.6	0.44
Transformer:	3.2	0.54
Non-stationary Transformer	3.6	0.61
Autoformer:	4.8	0.82
ETSformer:	6.3	1.07
Pyraformer:	8.7	1.48
Crossformer:	13.0	2.21

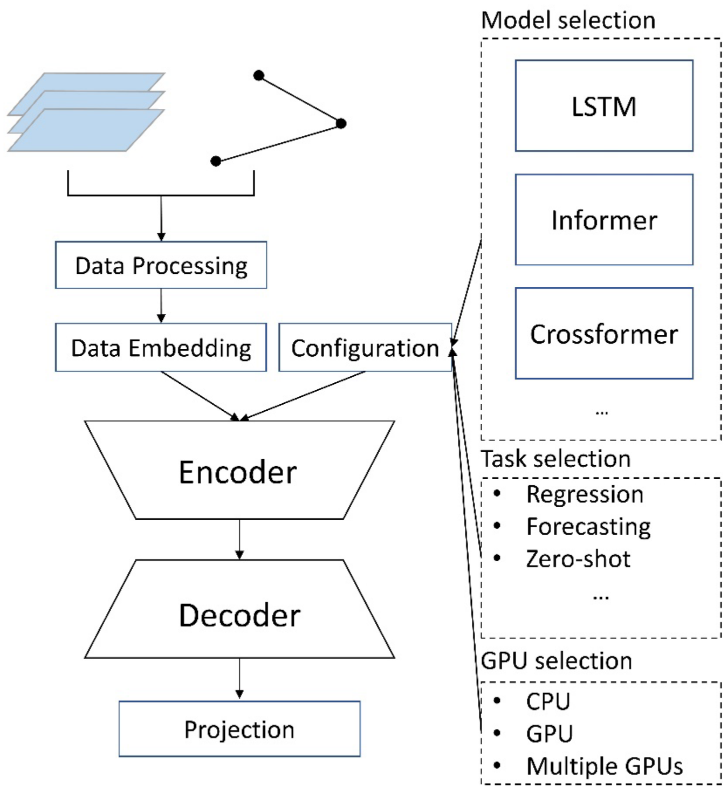

Figure S1. A schematic representation of the hydrologic modeling framework. The box on the right-hand side indicates selections of model, task, and GPU. Boxes in the main figure represent processing modules. Black arrows indicate the flow of data and information between processing steps.

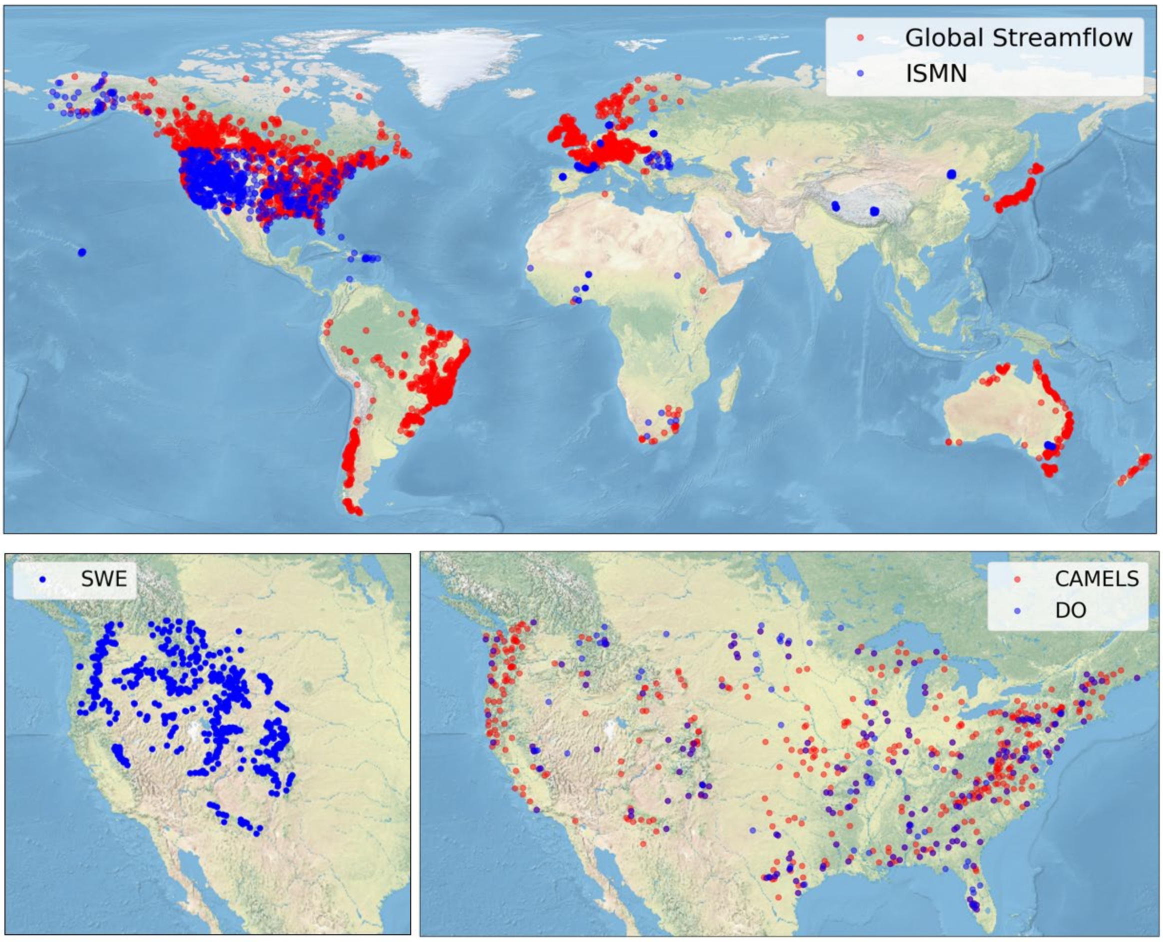

Figure S2. Overview of dataset distributions. CAMELS (Catchment Attributes and Meteorology for Large-sample Studies): 531 basins; Global streamflow dataset: 3,434 basins; SWE (Snow Water Equivalent) dataset: 525 sites; DO (Dissolved Oxygen) dataset: 236 basins; ISMN (International Soil Moisture Network) dataset: 1,317 sites. Maps are made with Natural Earth imagery, no permission needed (naturalearthdata.com).

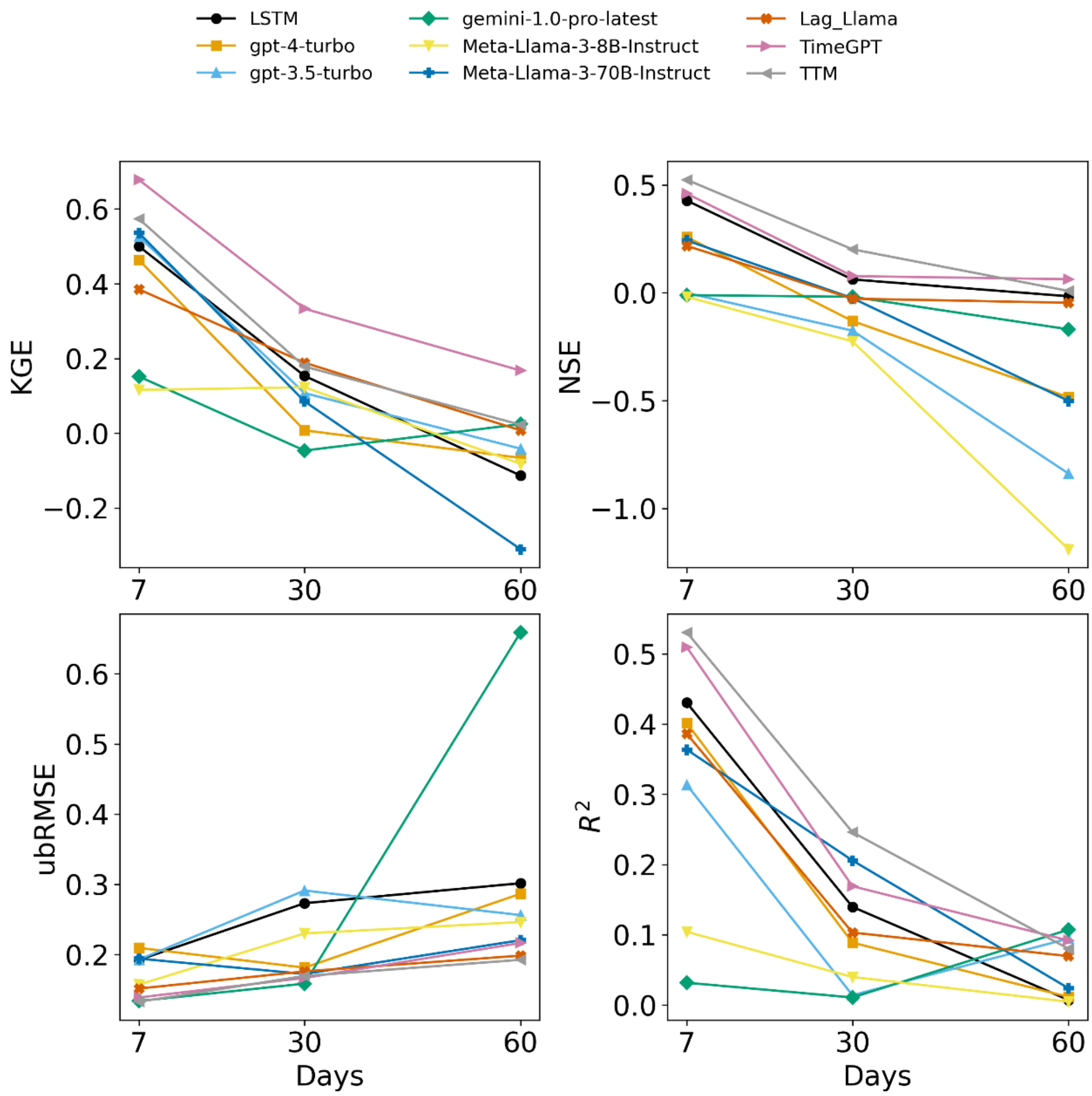

Figure S3. Comparison of Large Language Models (LLMs) and Time Series Attention-based Models (TSAMs) for runoff forecasting under zero-shot forecasting scenarios. Results are presented for 7-day, 30-day, and 60-day forecasting horizons based on a random selection of seven basins from the CAMELS dataset, with each basin containing 420 days of observations. A 90-day historical data window is used in all experiments to forecast runoff for the subsequent 7, 30, or 60 days.