Hydrological Prediction in Ungauged Basins

Ho Bao Thu, Ha Viet Khanh, Tran Kim Cuong Dang Thinh Tuong Minh, Nguyen Tat Thanh, Le Anh Son, Hoang Nguyen Long, Nguyen Trong Minh Phuong,

School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam

Abstract. River discharge observations are often sparse and discontinuous due to limited gauging networks and data gaps, particularly in ungauged and poorly gauged basins. To address this limitation, this study proposes a multi-source data integration framework for streamflow prediction that leverages complementary hydro-meteorological and geophysical information. Precipitation data from NOAA, meteorological variables from ERA5 reanalysis and TIGGE forecasts, and static basin attributes from HydroBASINS and HydroATLAS are systematically collected, preprocessed, and harmonized into a unified dataset aligned by station, time, and physical units. Exploratory and temporal analyses, including distributional diagnostics, hydrological memory assessment, and seasonal characterization, are conducted to ensure statistical consistency and to guide the selection of appropriate input window lengths. The resulting unified dataset is subsequently used to develop and evaluate streamflow prediction models in data-scarce and ungauged basins. Results demonstrate that the integrated multi-source dataset provides a physically consistent and effective basis for streamflow prediction in data-scarce and ungauged basins. aiming to support early warning and risk-informed planning in Vietnam.

1 Introduction

River discharge is a fundamental variable in hydrology, underpinning water resources management, ecosystem sustainability, and hydrological risk assessment. Accurate prediction of river discharge is essential for understanding basin-scale hydrological responses under changing climatic and land-use conditions. However, despite its importance, discharge observations remain spatially sparse and temporally incomplete across much of the global river network.

However, accurate prediction of these variables remains a formidable challenge due to the pervasive ungauged basin problem. While reliable in-situ gauges provide the gold standard for hydrological monitoring, their global distribution is critically sparse and declining in many regions [1, 2]. This lack of long-term, high-quality observational data often renders traditional time-series forecasting models ineffective for most of the world's river reaches. Therefore, there is an urgent need to pivot toward predictive frameworks that leverage widely accessible, cost-free, and remotely-sensed environmental datasets.

Traditionally, river forecasting has relied on physically-based hydrological models or numerical weather prediction (NWP) driven workflows. Although these models offer valuable process-based insights, they often suffer from heavy computational requirements, high sensitivity to parameter calibration, and the need for detailed catchment-specific physiographic data [3]. In recent years, data-driven approaches, particularly Machine Learning (ML), have emerged as a reliable and computationally efficient alternative. Unlike physical models, ML architectures can ingest vast amounts of multi-source data to learn complex, non-linear mappings without explicit prior knowledge of the underlying physical equations [4].

A fundamental complexity in hydrological modeling is the rainfall–runoff relationship, which is inherently non-linear and non-local in time [5]. The response of a river basin to a precipitation event is not solely determined by the rainfall volume but is significantly modulated by the antecedent state of the catchment, such as soil moisture and current water levels. Furthermore, the lack of gauge stations for water level monitoring limits the spatial scalability of local prediction models. This necessitates a more holistic approach that considers the synergistic effects of meteorological forcings, land surface characteristics, and the dynamical state of the river system.

Motivated by the need for accurate, generalizable predictions across data-sparse basins, we propose a framework that leverages readily available geographic, meteorological, and precipitation inputs to forecast river water levels at large scales. By integrating static basin attributes (e.g., terrain, land cover from HydroATLAS) with dynamic forcings (e.g., ERA5 reanalysis and GPM satellite rainfall), our regression-based machine learning models capture the non-linear dependencies modulated by antecedent river states—circumventing the scarcity of in-situ gauges. This approach enables scalable, transferrable predictions without long-term local observations.

2 Related Work

Recent advances in data-driven hydrological modeling have been strongly supported by the availability of large-scale, standardized datasets that integrate meteorological, hydrological, and geographic information. This

section reviews representative datasets and prior studies that motivate the data selection and methodological design of this work.

2.1 Global prediction of extreme floods in ungauged watersheds

In a seminal study on global hydrological intelligence, Nearing et al. (2024) [6] addressed the decadal challenge of "prediction in ungauged basins" (PUB) by leveraging large-scale artificial intelligence. By training Long Short-Term Memory (LSTM) networks on publicly available streamflow data from 5,680 watersheds, the authors established a forecasting framework capable of robust spatial extrapolation without the prerequisite of local model calibration. Their findings revealed that AI-based forecasting achieves a reliability at up to a five-day lead time that is comparable to or better than the zero-day nowcasts provided by the Copernicus Global Flood Awareness System (GloFAS), which represents the current state-of-the-art in global modeling. Furthermore, the model exhibited a significant capacity to predict extreme hydrological events, achieving accuracies for five-year return period events that met or exceeded the benchmark's performance for common one-year events. This system demonstrated improvements over the GloFAS benchmark in approximately 70 percentages of evaluated watersheds and has been successfully operationalized to provide real-time, public-access flood warnings in over 80 countries.

2.2 Large-Sample Hydrology Datasets

2.2.1 Caravan Dataset The Caravan dataset is a global large-sample hydrology dataset designed to support machine learning research across diverse hydroclimatic conditions. It integrates basin-scale streamflow observations with standardized meteorological forcings from ERA5-Land and static geographic attributes derived from HydroATLAS. A key contribution of Caravan is its consistent data structure across thousands of catchments worldwide, enabling fair comparison between different deep learning architectures without confounding effects from heterogeneous preprocessing.

Caravan has become a widely adopted benchmark for evaluating data-driven rainfall–runoff and streamflow forecasting models. Its emphasis on transferability to ungauged basins makes it particularly relevant for large-scale forecasting studies. In this work, Caravan serves as a methodological reference guiding the selection of input variables and basin attributes. [7]

2.3 Datasets Excluded Due to Conceptual or Scale Limitations

2.3.1 GRACE The GRACE satellite mission measures changes in terrestrial water storage at continental to sub-continental scales. Although valuable for large-scale water balance studies, its spatial resolution is too coarse to support validation of river discharge at individual catchment scales. Signals from neighboring basins often overlap, limiting its applicability for basin-level hydrological evaluation. [8]

2.3.2 FLUXNET2015 FLUXNET2015 provides high-quality point-scale measurements of land–atmosphere energy and water fluxes. However, these observations represent highly localized conditions and lack spatial representativeness for heterogeneous river basins that include mixed land cover types, urban areas, and surface water bodies. [9]

2.3.3 GLDAS and WFDEI GLDAS and WFDEI are land surface and forcing datasets derived from meteorological reanalysis and model-based assimilation systems. Using such datasets to validate ERA5-driven hydrological models introduces circular dependency, as they rely on similar atmospheric forcings. Additionally, WFDEI provides primarily monthly data and is based on the deprecated ERA-Interim reanalysis, limiting its suitability for daily flood forecasting. [10, 11]

2.4 Summary

In summary, this study follows recent trends in data-driven hydrology by leveraging globally consistent meteorological forcings, multi-source precipitation products, and basin-scale geographic attributes. At the same time, it explicitly avoids datasets that introduce spatial scale mismatches or model dependency biases, aligning the proposed framework with state-of-the-art approaches for ungauged basin streamflow prediction.

3 Data Collection & Processing

3.1 Data Collection & Preprocessing

3.1.1 Hydrological Observations River discharge observations used as ground truth in this study are obtained from the Global Runoff Data Centre (GRDC) [12]. GRDC is an internationally recognized repository that provides quality-controlled river discharge data collected from national hydrological services worldwide. The dataset is widely used in hydrological modeling and validation studies, and its standardized data management procedures ensure a high level of reliability and consistency across stations.

Discharge data are provided in the form of station-based daily time series measured at fixed gauging locations. Each GRDC station file contains discharge records along with associated station metadata. Only stations with discharge observations overlapping the study period are considered in order to ensure temporal consistency with other datasets used in this study. All data are obtained directly from the official GRDC archive to maintain data integrity.

Prior to analysis, preprocessing steps are applied to the raw GRDC files. Non-essential metadata fields contained in the original file headers are removed, retaining only information required for hydrological analysis, including station identifiers (GRDC-No.), catchment area, observation dates (`YYYY-MM-DD`), discharge values, and geographical coordinates.

Subsequently, records containing invalid or missing discharge values, indicated by the fill value `-999.000`, are identified and excluded from the dataset. This step ensures that only valid and physically meaningful discharge observations are used in subsequent data integration and modeling.

3.1.2 Precipitation Precipitation data are obtained from NOAA gauge-based and gridded products [13] and satellite-based precipitation estimates from NASA IMERG [14]. NOAA provides both gridded and gauge-based precipitation datasets with complementary characteristics. The gridded NOAA precipitation data offer complete spatial coverage over the study region and do not contain missing values; however, they are provided in a data-intensive format that is computationally more demanding to process. In contrast, gauge-based NOAA station data are spatially sparser and exhibit temporal gaps, but they are distributed as station-wise time series that are more straightforward to handle and integrate within data-driven analyses. To address the limited coverage and missing values in gauge-based observations, satellite-derived precipitation estimates from NASA IMERG are used to fill missing values in NOAA gauge records. Although satellite-based precipitation is generally less accurate than in-site gauge measurements, its continuous spatial coverage makes it suitable for gap-filling applications.

Both NOAA gridded precipitation data and NASA IMERG data are provided in NetCDF4 (`.nc4`) format, ensuring structural consistency between the two datasets. In contrast, NOAA gauge-based precipitation data are provided as station-wise CSV files containing time series observations at individual gauging locations. For gauge-based precipitation, approximately 6,000 NOAA stations located in South America with available precipitation records during the period 2015–2024 are selected. NASA IMERG precipitation data covering the same temporal range are collected to ensure temporal consistency.

For NOAA gauge data, redundant variables and records outside the study period are removed, while for NASA data only precipitation values within the South America bounding box are retained. South America is characterized by a predominantly tropical climate, with precipitation concentrated mainly in the wet season from October to April. During the dry season (May to September), many NOAA gauge stations exhibit missing precipitation records due to the inherently low rainfall amounts and the dominance of no-rain days. To account for this seasonal behavior, NOAA gauge data are analyzed separately for wet and dry seasons. For each station, the number of available records and statistical performance metrics, including correlation, root mean square error (RMSE), and mean bias error (MBE), are computed by comparing NOAA gauge observations with corresponding NASA IMERG precipitation estimates.

Only wet-season data are considered for precipitation gap filling. Stations are selected if they satisfy the following criteria: more than 200 available wet-season records, correlation greater than 0.6, and absolute bias less than 20 mm. Stations that do not meet these criteria are excluded from the filling process to avoid introducing noise into the dataset. For selected stations, missing NOAA gauge precipitation values during the wet season are filled using a linear adjustment model based on NASA IMERG precipitation, expressed as:

$$P_{\mathrm{NOAA}}(t) = a \cdot P_{\mathrm{NASA}}(t) + b,$$

where $P_{\mathrm{NASA}}(t)$ denotes the NASA IMERG precipitation on day t, a is the slope coefficient that adjusts precipitation amplitude, and b is the bias offset. The coefficients a and b are estimated using the Ordinary Least Squares (OLS) method based on days with overlapping NOAA and NASA observations during the wet season. A small precipitation threshold of 0.2 mm is applied to the filled values to remove spurious light rainfall events.

During the dry season, precipitation amounts are generally very low and strongly dominated by no-rain events. Missing NOAA gauge precipitation values are filled directly using raw NASA IMERG precipitation data, and the same 0.2 mm threshold is applied to suppress false precipitation signals.

3.1.3 Meteorological Datasets Meteorological forcing variables are used to characterize atmospheric conditions influencing river discharge dynamics at the basin scale.

Meteorological data are collected from two complementary sources:

- *Reanalysis*: ERA5, provided by the European Centre for Medium-Range Weather Forecasts (ECMWF), which offers physically consistent and temporally continuous meteorological fields through data assimilation of observations and numerical weather prediction models.
- *Forecast*: TIGGE (THORPEX Interactive Grand Global Ensemble), which provides ensemble-based numerical weather forecasts from multiple international meteorological centers.

Both ERA5 and TIGGE provide a wide range of atmospheric variables at high temporal resolution. However, only variables that are directly relevant to river discharge generation are downloaded and used in this study, following established practices in hydrological modeling literature (e.g., [15], [16]).

Meteorological variables are extracted over each river basin and spatially aggregated to basin-averaged time series to ensure consistency with river discharge observations.

Although six meteorological variables are initially extracted, snowfall exhibits near-zero or constant zero values across most South American basins due to predominantly tropical and subtropical climate conditions, resulting in negligible contribution to hydrological variability. Consequently, only five basin-averaged meteorological variables are retained for subsequent analysis and model training.

3.1.4 Geological Datasets Geological and physiographic characteristics of river basins provide essential static information that controls runoff generation, flow accumulation, and long-term hydrological response.

Geological and basin attribute data are obtained from two complementary global datasets:

- *HydroBASINS*, a global standardized watershed delineation dataset derived from the HydroSHEDS framework, which provides hierarchical basin boundaries and drainage network topology.
- *HydroATLAS*, a comprehensive global database of hydro-environmental attributes linked to HydroBASINS, offering basin-level information on topography, geology, soil properties, land cover, climate indices, and upstream characteristics.

The joint use of HydroBASINS and HydroATLAS is necessary to ensure both geometric consistency and attribute completeness. HydroBASINS defines hydrologically consistent basin boundaries and upstream–downstream relationships, while HydroATLAS enriches these basins with a wide range of geological and environmental descriptors that cannot be derived from basin geometry alone. Using both datasets allows static basin characteristics to be accurately aligned with river discharge observations and meteorological forcings.

After downloading both datasets, a validation procedure is conducted to ensure consistency and completeness across basin representations. Basin identifiers, spatial extents, and hierarchical relationships are cross-checked between HydroBASINS and HydroATLAS to confirm one-to-one correspondence. Basins with missing attributes, inconsistent upstream area values, or mismatched identifiers are excluded from further analysis to avoid introducing structural biases into the modeling framework.

3.2 Data Integration

Data integration was designed as a multi-step pipeline to construct a consistent spatio-temporal dataset linking river discharge observations with upstream basin characteristics, precipitation forcing, and meteorological variables. Each step in the pipeline serves a specific purpose and progressively enriches the dataset used for model training and evaluation. The overall pipeline is illustrated in Figure 1

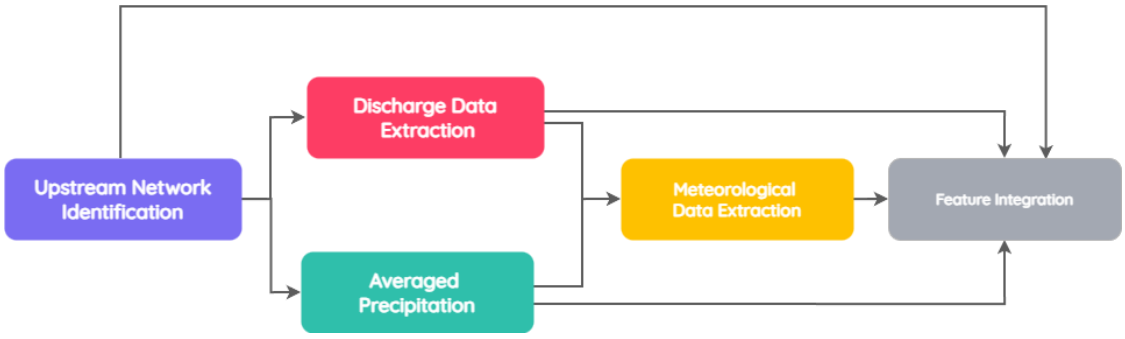


Fig. 1: Overview of the multi-source data integration pipeline.

Step 1: Upstream Network Identification This step assigns each hydrological observation station to its corresponding local river basin and reconstructs the complete upstream basin network.

First, each station is matched to its local basin based on geographical coordinates (latitude and longitude). To ensure spatial consistency, the reported discharge observation is validated by comparing the catchment area of every station with the upstream area provided in the basin topology data. Stations with missing values or with discrepancies exceeding approximately 20% are excluded. This filtering step ensures that the observed discharge corresponds to a hydrologically consistent drainage area aligned with the associated static attributes and precipitation inputs.

For the remaining valid stations, the full upstream basin network is traced starting from the identified local basin. All contributing upstream sub-basins are recursively identified, and their total number and cumulative contributing area are computed to verify the completeness and accuracy of the reconstructed drainage network.

Finally, basin-scale static attributes are aggregated over the entire upstream network using area-weighted averaging, yielding representative long-term geographical and physiographic characteristics for each observation station.

Step 2: Discharge Data Extraction After identifying valid stations, daily river discharge time series are extracted for each selected station and synchronized over the period of 2019–2025. These observations serve as the reference output (ground truth) for subsequent data integration and modeling.

Step 3: Upstream Precipitation Averaging In this step, precipitation data are aggregated over the upstream basin network associated with each station. Using the mapped topology between stations and their contributing upstream areas, basin boundaries defined in HydroBASINS are employed to spatially clip NOAA precipitation data over the same observation period. Basin-scale precipitation time series are then computed as area-weighted averages of the gridded precipitation fields, ensuring that the resulting precipitation forcing represents the integrated rainfall contribution across the entire upstream drainage area rather than point-based measurements.

Step 4a: Meteorological Reanalysis Data Extraction Meteorological variables are extracted from ERA5 reanalysis products following [17].

Given the large spatial and temporal resolution of ERA5 reanalysis products, the extraction procedure is designed to minimize redundant computation and memory overhead. Rather than processing data independently for each basin, meteorological fields are first spatially restricted to a unified bounding box covering all valid upstream basins identified in Step 1. This preliminary clipping substantially reduces the size of the gridded domain prior to basin-level aggregation.

To further improve efficiency, basin-level statistics are computed in a hierarchical manner. For each meteorological variable and each daily time step, gridded values are first aggregated at the HydroBASINS sub-basin level using a precomputed region mask. These sub-basin means are calculated only once per day and reused across all downstream stations sharing the same contributing basins. Station-level meteorological time series are then derived through area-weighted aggregation of the corresponding upstream sub-basins, avoiding repeated reprocessing of the same ERA5 grids.

Step 4b: Meteorological Forecast Data Extraction The same meteorological variables are extracted from the TIGGE dataset to represent forecast numerical predictions. The ERA5-based spatial clipping and grid aggregation strategy is reused for TIGGE forecast fields to generate time series of meteorological variables for each basin.

Basin-level meteorological time series are computed using vectorized region masking, where gridded forecast values are aggregated only once at the HydroBASINS level for each forecast time step. Station-level forecast inputs are subsequently derived through area-weighted aggregation of the corresponding upstream sub-basins, ensuring consistency with the discharge–precipitation spatial mapping. The resulting daily forecast meteorological time series are temporally aligned with the modeling period and used as predictive inputs for downstream hydrological forecasting.

Step 5: Data Unification and Physical Normalization In this step, precipitation, discharge, and meteorological variables are consolidated into a unified dataset to ensure consistent temporal alignment across all data sources. Specifically, all variables are merged based on the common `(grdc_no, date)` key, guaranteeing that each record corresponds to the same station and observation date.

Due to differences in variable definitions and physical units between ERA5 reanalysis and TIGGE forecast products, meteorological variables are standardized to consistent units prior to further analysis. This harmonization step ensures comparability between historical and forecast inputs and prevents scale-induced biases during model training.

In addition, physically plausible bounds are applied to selected meteorological variables to enforce physical and climatological consistency. Near-surface air temperature ($t2m$) is constrained to the range [−89.2°C, 56.7°C], corresponding to the lowest and highest reliably observed surface air temperatures on Earth. Surface pressure (sp) is restricted to the interval [30, 108.5] kPa, reflecting physically attainable atmospheric pressure values from high-altitude mountainous regions to extreme high-pressure systems. Values falling outside these ranges are treated as physically implausible and corrected or excluded during preprocessing, thereby reducing numerical artifacts and improving the physical realism of the final input dataset.

4 Data analysis

4.1 Temporal structure of the target variable

To establish a rigorous foundation for the predictive modeling of river discharge, the temporal dynamics of the target variable were characterized through lag-dependency and seasonal variability analyses. The primary objective was to determine whether the discharge series exhibits a deterministic structure—specifically "hydrological memory"—and to identify periodic patterns that may be obscured by spatial aggregation.

4.1.1 Temporal lag dependence The temporal dependence of the global mean discharge was first evaluated using the Autocorrelation Function (ACF) over a 400-day lag period. As illustrated in Figure 2, the ACF does not exhibit the rapid decay characteristic of white noise; instead, it demonstrates a slow, persistent decline, remaining statistically significant over an extended duration. This decay pattern is indicative of long-range dependence, confirming the existence of a substantial hydrological memory within the system. From a physical perspective, such memory implies that the discharge is not merely a stochastic response to instantaneous meteorological forcing but is heavily regulated by catchment storage and routing processes, such as groundwater fluctuations and soil moisture retention. The identification of this persistent structure justifies the application of time-series architectures capable of capturing long-term temporal dependencies.

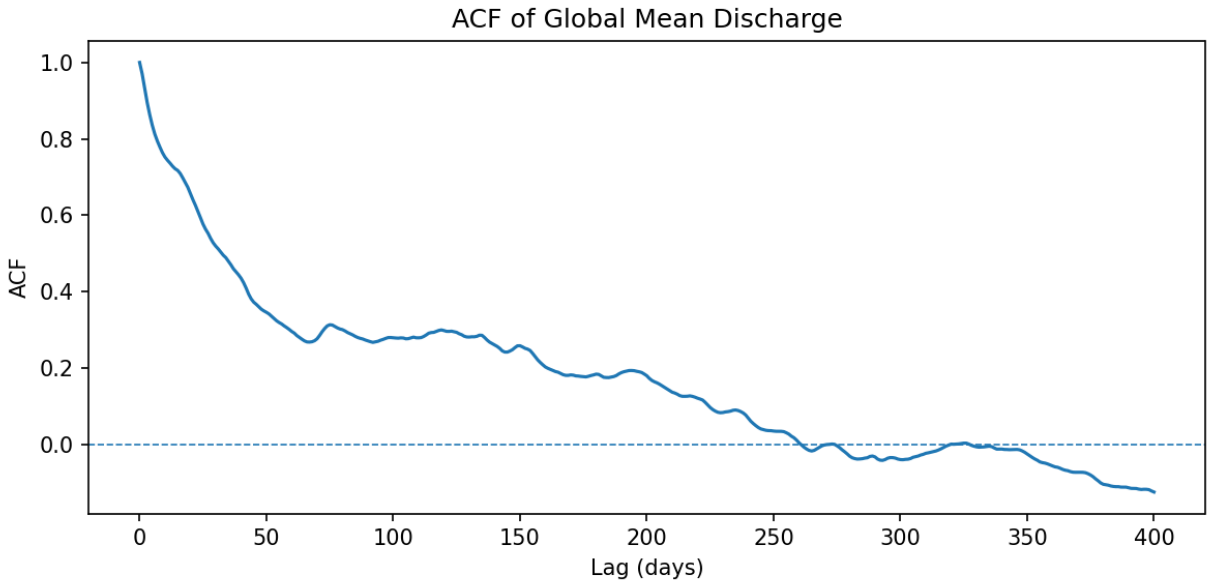


Fig. 2: Autocorrelation Function (ACF) of global mean discharge, illustrating the slow decay characteristic of long-range hydrological memory.

4.1.2 Seasonal discharge patterns The hydrological regime of the study area is characterized by a distinct seasonal structure, necessitating a rigorous evaluation of temporal patterns to optimize model training. Preliminary analysis of the discharge data confirms a robust annual cycle, with recurrent peaks and troughs aligned with regional climatic drivers. This intra-annual variability presents a significant challenge for short-term modeling; specifically, adopting a monthly training window would introduce substantial sampling bias due to the high discharge distinction between wet and dry months. To mitigate this and ensure the model captures the full spectrum of hydrological behavior, we established an annual training window, which provides a holistic representation of the yearly discharge distribution and prevents the model from over-fitting to transient monthly fluctuations.

Furthermore, we identify a critical scale-dependency in the seasonal signal. Aggregating data through a global mean across all basins tends to obscure localized patterns due to spatial-temporal asynchrony. Since different geographical regions experience peak precipitation at varying times of the year, global averaging results in signal attenuation, masking the true intensity of the seasonal cycle. To demonstrate this phenomenon, we selected two representative basins at random for granular inspection: Basin 3107010 and Basin 3107015 (Figure 3).

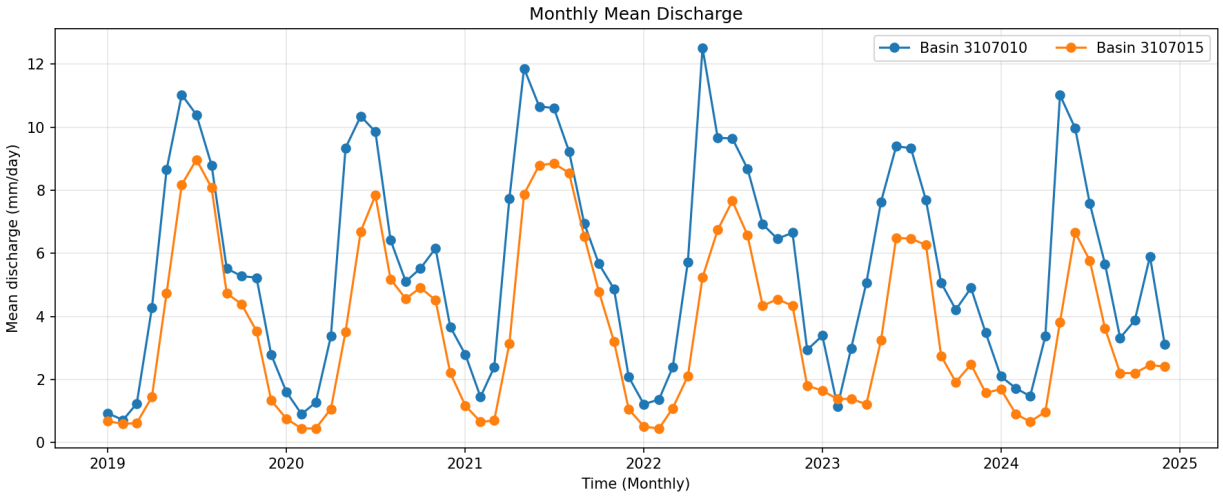


Fig. 3: Monthly mean discharge for two representative basins, demonstrating consistent annual periodicity and inter-basin temporal shifts.

As illustrated, while both basins exhibit clear annual periodicity, their respective hydrographs reveal significant temporal shifts in peak discharge. This divergence underscores the necessity of incorporating basin-specific features to preserve the integrity of the seasonal signal in multi-scale forecasting.

4.2 Precipitation forcing data quality and coverage

This section analyzes the distributional properties and availability of station-based precipitation data from NOAA to assess its suitability for basin-scale streamflow modeling. Exploratory analysis shows that, where observations are available, NOAA gauge precipitation exhibits physically reasonable distributions and temporal behavior. To address temporal gaps in individual station records, missing values are supplemented using gridded precipitation products from NASA. Correlation analysis over overlapping periods indicates strong agreement between NOAA gauge observations and NASA-based estimates (Figure 4), suggesting that gap-filling is an effective approach for mitigating temporal discontinuities at the station level.

However, while gap-filling improves temporal completeness, it does not resolve the more fundamental limitation of spatial data sparsity. Rain gauge coverage is extremely sparse relative to both the number of river discharge stations and the spatial extent of HydroBASINS catchments. As shown in Table 1, 4,414 out of 4,436 HydroBASINS catchments (99.5%) contain no in-basin precipitation station, and 39 out of 52 GRDC discharge stations (75.0%) lack colocated rain gauge support. This severe spatial mismatch prevents station-based precipitation from providing consistent and representative basin-scale forcing across the study domain.

Based on these findings, station-based NOAA precipitation data are not used as direct inputs for model training. Although enhanced through gap-filling, the limited spatial coverage of rain gauges renders station-based precipitation unsuitable for large-scale and ungauged basin applications. Consequently, gridded precipitation products are adopted for model training to ensure spatial completeness, uniform forcing availability, and consistency across both gauged and ungauged basins.

Category	Total	Without rain station
HydroBASINS catchments	4,436	4,414
GRDC discharge stations	52	39

Table 1: Precipitation station availability.

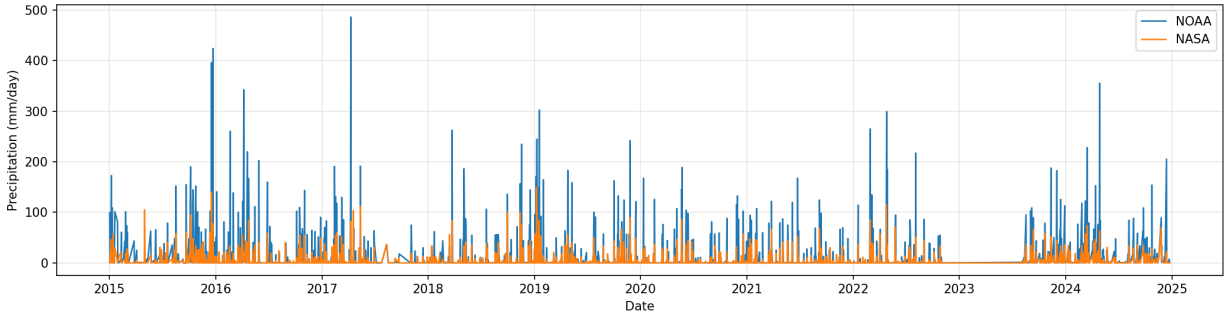


Fig. 4: Correlation between NOAA gauge data and NASA satelite precipitaion

4.3 Precipitation–Discharge dependence

4.3.1 Pearson x Spearman To evaluate the predictive potential of precipitation as a primary driver of river discharge, the statistical dependence between these variables was examined through both linear and non-linear lenses. Pearson correlation was utilized to establish a baseline for instantaneous linear relationships, while Spearman rank correlation was employed to capture monotonic, non-linear dependencies. The latter is particularly critical in hydrological modeling, as runoff mechanisms typically involve non-linear thresholds where significant discharge only occurs after specific soil moisture or precipitation limits are surpassed.

The results of the lagged correlation analysis, presented in Table 3, reveal a distinct discrepancy between the two metrics across all temporal lags. Specifically, the Pearson coefficients remain relatively low, hovering around 0.22–0.26, whereas the Spearman coefficients are markedly higher, peaking at approximately 0.44. This significant difference confirms that the relationship between precipitation and discharge is inherently non-linear rather than a simple proportional response. Furthermore, the analysis indicates that the hydrological response is not instantaneous; the peak correlation does not occur at lag 0 but is instead reached at a 2-day lag for Pearson ($r = 0.264$) and a 3-day lag for Spearman ($\rho = 0.437$). This temporal shift reflects the physical delay required for precipitation to infiltrate the catchment and transition through routing processes before contributing to measurable streamflow.

Lag (days)	Pearson (r)	Spearman (ρ)
0	0.222380	0.399997
1	0.252133	0.421927
2	**0.264080**	0.434700
3	0.251517	**0.437751**
4	0.239658	0.433852
5	0.227008	0.427548
6	0.216791	0.421088
7	0.210500	0.414148

Table 2: Pearson (r) and Spearman (ρ) correlation coefficients between precipitation $P(t-k)$ and discharge $Q(t)$ for lags $k = 0$ to 7 days.

Despite the statistical significance of these correlations, it is important to note that the peak values remain moderate, suggesting that individual precipitation events alone are insufficient to fully explain the variance in river discharge. The moderate magnitude of the coefficients indicates that discharge is a multi-factorial process influenced by additional variables such as antecedent moisture conditions and evapotranspiration. Consequently, while precipitation serves as a vital input, the modeling framework must account for its non-linear nature and temporal lag to achieve reliable forecasting accuracy without overestimating the influence of isolated rainfall events.

4.3.2 Dual Axis Time Series Analysis The interaction between meteorological inputs and hydrological outputs is fundamentally characterized by a temporal offset in the catchment's response. Observation of the dual-axis temporal distribution reveals a distinct delay response of discharge relative to precipitation events (Figure 5). This lag indicates that fluctuations in discharge are physically driven by preceding rainfall, which justifies the use of precipitation as a primary predictor in the discharge forecasting model.

However, the implementation of long-term forecasting faces practical limitations. Due to significant challenges in downloading and processing high-resolution meteorological data, this study focuses on a 1-day lead time for the predictive task. This constraint ensures a balance between the physical relevance of the inputs and the operational feasibility of the data pipeline.

Furthermore, the empirical evidence presented in the figure confirms that while precipitation contributes significantly to discharge, the relationship between these two variables is non-linear. The magnitude of discharge peaks does not always scale proportionally with rainfall intensity across different periods. This non-linearity implies that precipitation alone is insufficient to capture the full complexity of the hydrological process, thereby highlighting the necessity of incorporating additional environmental variables to improve the accuracy of the discharge predictions.

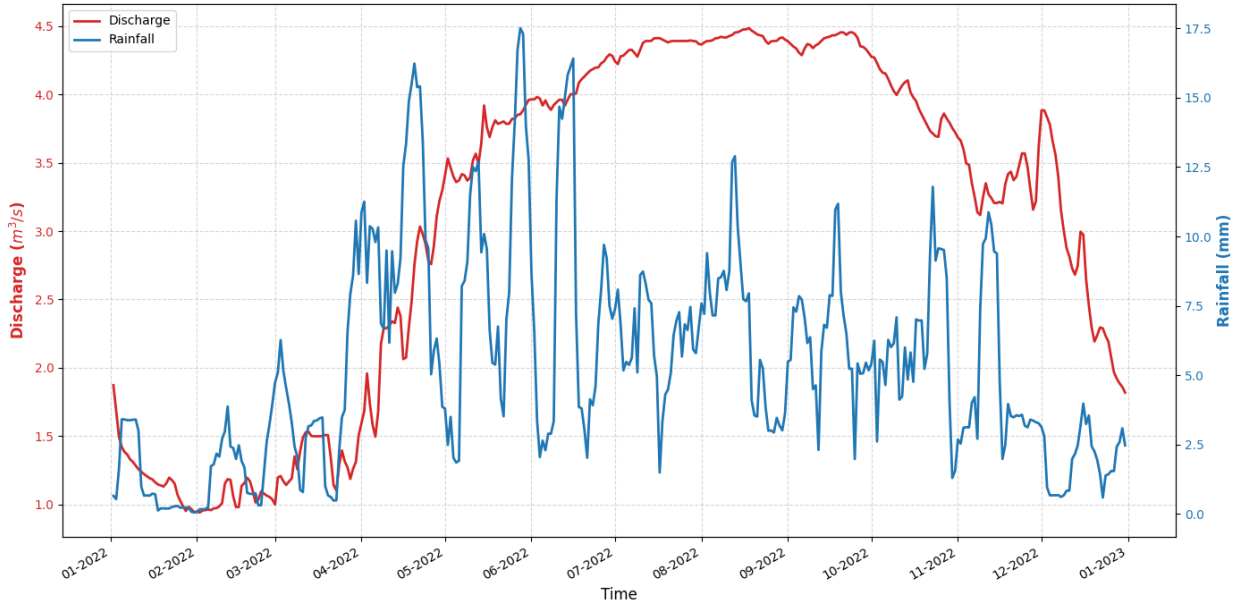


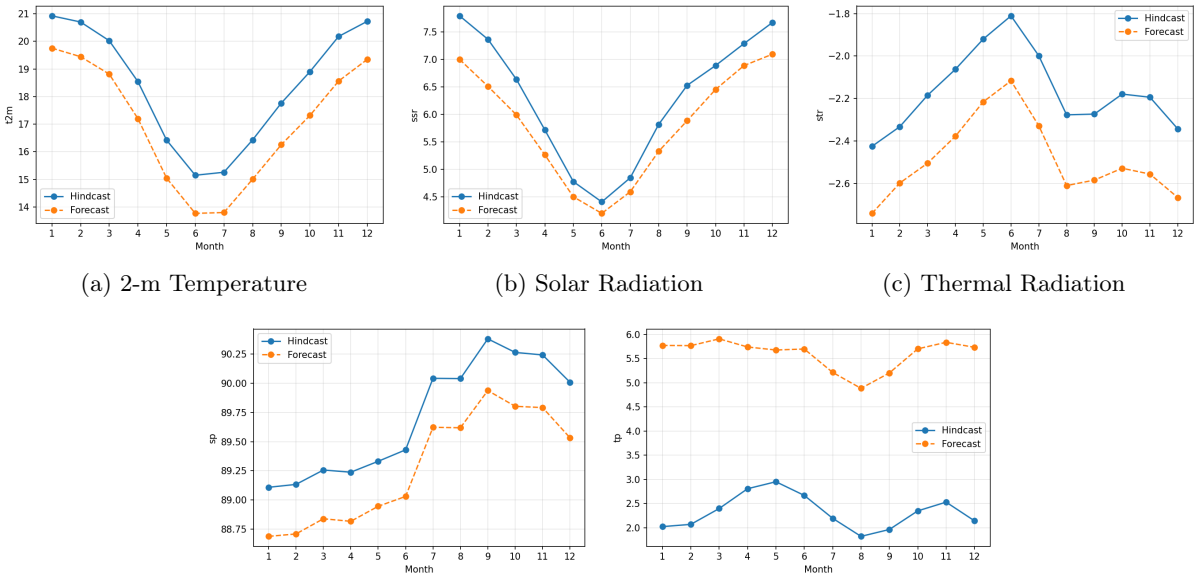
Fig. 5: Dual-axis time series illustrating the temporal relationship between discharge (red line, left axis, m^3/s) and the inverted 7-day moving average of precipitation (blue line, right axis, mm) over the study period.

4.4 Statistical consistency of meteorological forcing data

4.4.1 Comparison of ERA5 and TIGGE data regimes To ensure the reliability of the forecasting framework, it is essential to verify the consistency between the ERA5 hindcast data and the TIGGE (ECMWF) forecast products. This verification process aims to determine if both datasets share a similar climatological regime, which is a prerequisite for robust model training and seamless transitions during real-time inference. The seasonal cycles of five key meteorological variables—total precipitation (tp), surface thermal radiation (str), surface solar radiation (ssr), surface pressure (sp), and 2-m temperature (t2m)—were analyzed and compared based on their mean monthly values.

	t2m	**tp**	**sp**	**ssr**	**str**
Pearson	0.998	0.536	0.999	0.993	0.991

Table 3: Pearson (r) and Spearman (ρ) correlation coefficients between precipitation $P(t-k)$ and discharge $Q(t)$ for lags $k = 0$ to 7 days.



(a) 2-m Temperature (b) Solar Radiation (c) Thermal Radiation

(d) Surface Pressure (e) Total Precipitation

Fig. 6: Comparison of seasonal cycles between ERA5 hindcast and ECMWF forecast datasets for various meteorological variables.

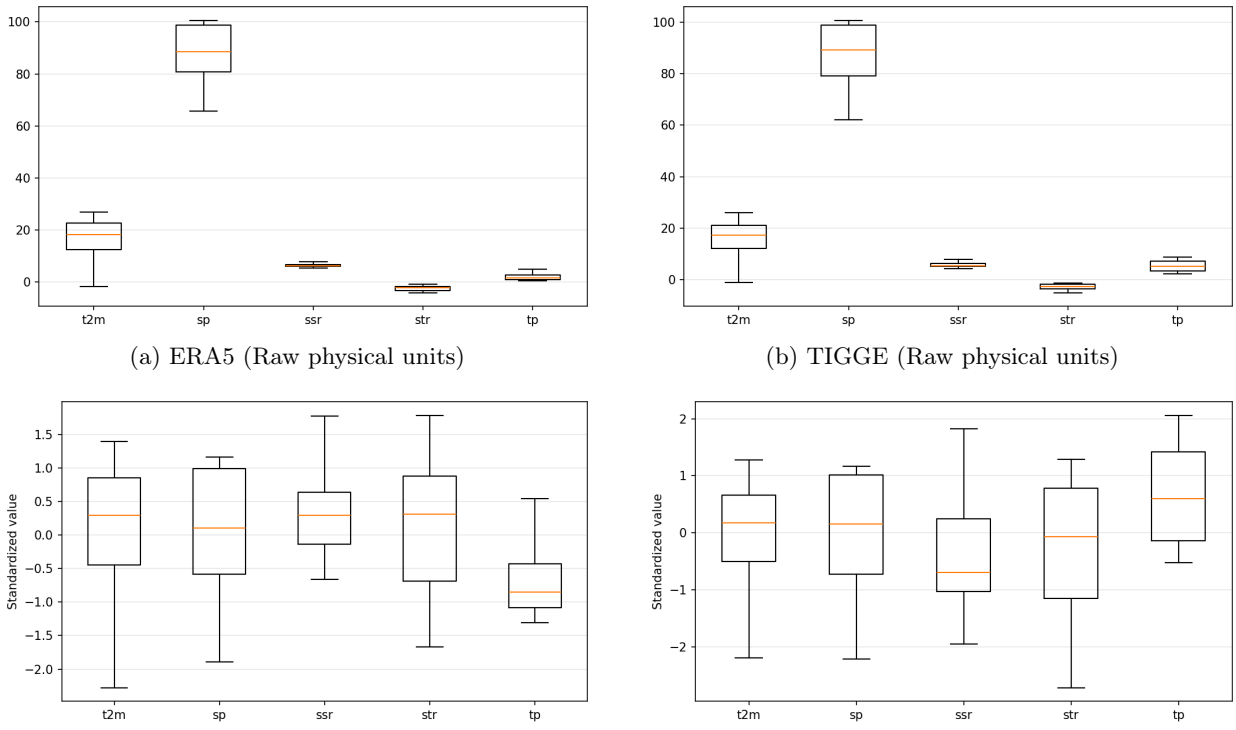
The analysis indicates a high degree of climatological agreement for the majority of the selected variables. As shown in Figure 6(a-d), the variables *t2m*, *ssr*, *str*, and *sp* exhibit nearly identical seasonal trajectories with Pearson correlation coefficients ranging from 0.98 to 0.99. Although minor systematic biases in magnitude are observed—specifically in radiation and pressure—the underlying seasonal regimes are highly synchronized, suggesting that these forecast variables are well-calibrated with the historical reanalysis data.

However, a significant discrepancy was identified in the total precipitation (*tp*) variable. As the correlation coefficient dropping to the range of 0.536 combine with a profound magnitude disparity in Figure 6(e) reveals a severe mismatch in magnitude and variance between the two datasets, with the ERA5 hindcast values appearing near zero relative to the TIGGE forecast cycle. This fundamental regime inconsistency indicates that the *tp* variable from these sources cannot be used interchangeably without extensive bias correction. Consequently, the precipitation variable was excluded from the current verification set, while the remaining variables were deemed suitable for the modeling process due to their consistent seasonal behavior.

4.4.2 Scale and variability verification Beyond regime consistency, it is imperative to assess the distribution and numerical scales of the meteorological features. Figures 7(a) and 7(b) illustrate the basin-mean distributions of the raw physical variables for the ERA5 and TIGGE datasets, respectively. A key observation is the pronounced difference in magnitude between various features; for example, surface pressure (*sp*) typically falls within roughly 60 to 100 units, while surface thermal radiation (*str*) is concentrated in the narrower interval from -5 to 0.

These variations are physically reasonable and reflect the inherent natural characteristics of the variables, such as the high absolute values of atmospheric pressure compared to heat flux or temperature. However, from a computational modeling perspective, such large differences in scale can introduce numerical instability and cause the model to be disproportionately biased toward features with larger absolute magnitudes, potentially masking the predictive signals of variables with smaller scales like *str* or *ssr*.

To mitigate these effects and ensure a balanced contribution of all meteorological drivers to the predictive framework, Z-score normalization was applied to both datasets. As shown in the lower panels of Figure 7(c) and 7(d), the transformation successfully maps all variables—*t2m*, *sp*, *ssr*, and *str*—onto a comparable dimensionless scale, with the majority of values now distributed within a consistent range between approximately -2.5 and 1.5. This normalization process preserves the underlying variability and outliers of each feature while establishing a uniform feature space, which is essential for the convergence and performance of gradient-based optimization algorithms.



(a) ERA5 (Raw physical units)

(b) TIGGE (Raw physical units)

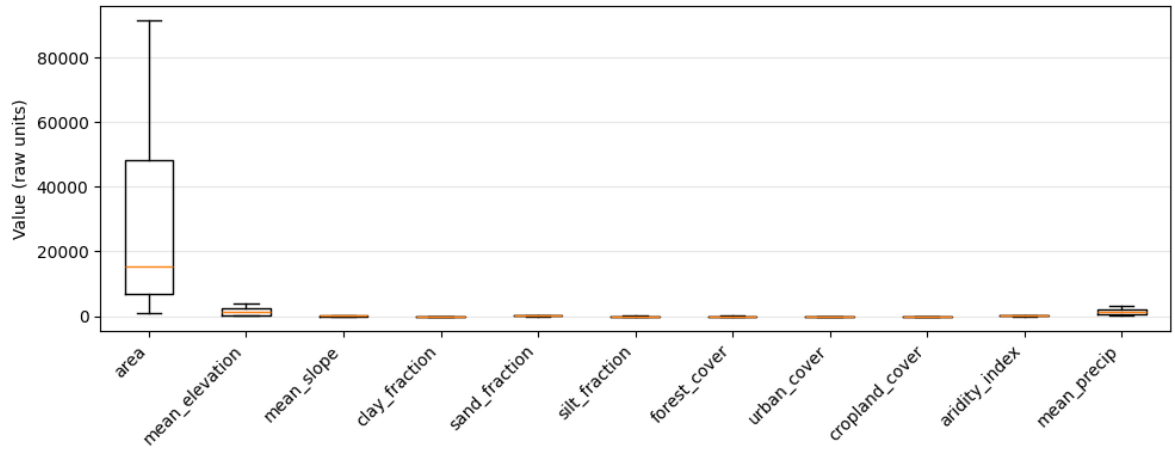
(c) ERA5 (Z-score normalized)

(d) TIGGE (Z-score normalized)

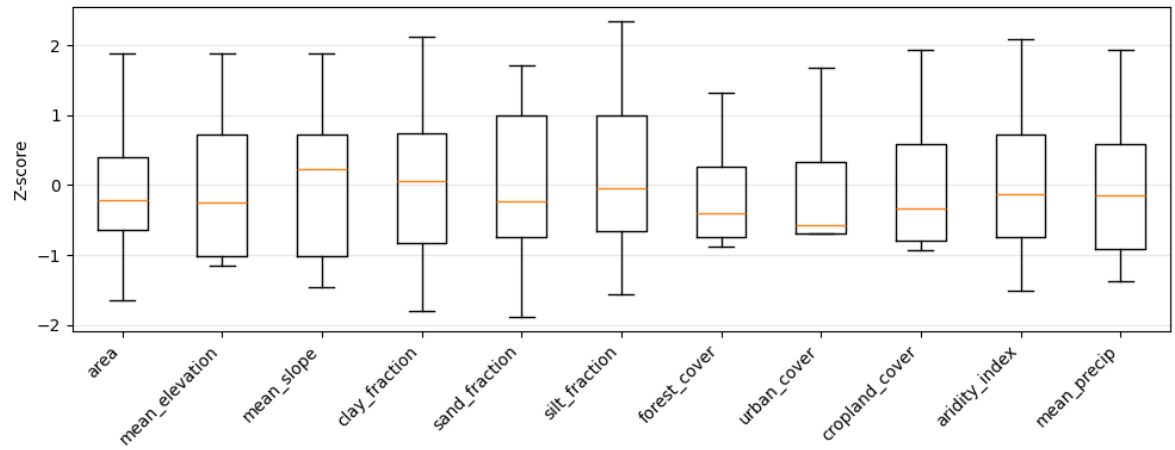
Fig. 7: Comparison of basin-mean distributions for meteorological variables before and after normalization for the ERA5 hindcast and TIGGE forecast datasets.

4.5 Static Catchment Attributes

The static attributes reveal significant geophysical and environmental diversity across the studied catchments. Analysis of the raw distributions in Figure 8(a) shows a high degree of landscape heterogeneity. The *area* attribute represents the primary factor of scale, spanning from small catchments to large-scale river systems. Topographical features, notably *mean_elevation* and *mean_slope*, exhibit substantial variance and numerous outliers, indicating a study area that encompasses diverse terrains from plains to high-altitude regions. Furthermore, the *urban_cover* is consistently low, suggesting the dataset primarily comprises natural or semi-natural landscapes, while the varying distributions of *forest_cover* and *cropland_cover* reflect a broad spectrum of land usage. As shown in Figure 8(b), Z-score normalization preserves these relative variances while aligning the attributes onto a consistent dimensionless scale for modeling.



(a) Raw distributions showing physical diversity



(b) Standardized distributions for cross-attribute comparison

Fig. 8: Basin-wise distributions of static features in raw (a) and normalized (b) scales.

A granular analysis of the *area* attribute in Figure 9 reveals a heavily right-skewed distribution in its raw state, where a vast majority of small catchments are contrasted by a few "mega-basins" in the long tail. This skewness indicates that a linear scale would over-represent extreme outliers. The *log1p* transformation successfully redistributes the data into a more balanced, log-normal form, uncovering a continuous spectrum of basin sizes. This ensures the model effectively interprets catchment size as a continuous physical gradient.

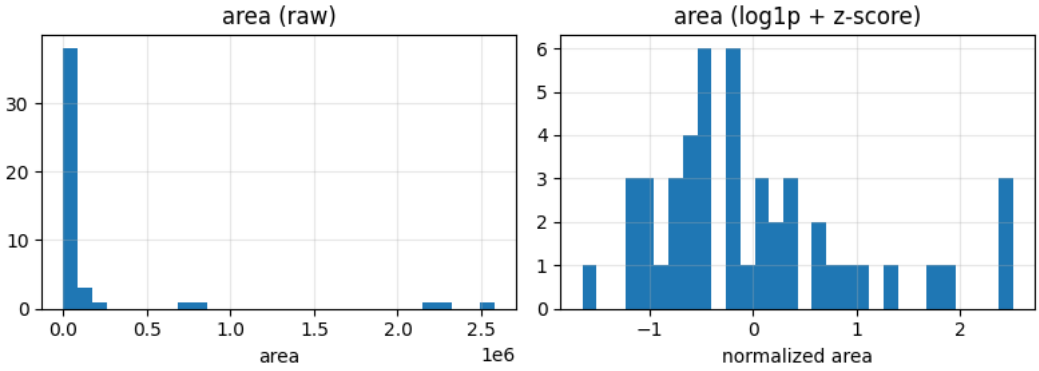


Fig. 9: Area distribution showing the transition from a skewed raw state to a balanced log-transformed state.

5 Forecasting Models

In this study, river discharge forecasting is performed using a deep learning model inspired by the global streamflow forecasting framework proposed by Nearing et al. [6]. The model is designed to predict short-term river discharge in both gauged and ungauged basins by leveraging both dynamic and static features.

Group	Variable	Description
Dynamic	`t2m`	Near-surface air temperature representative of basin-mean conditions.
	`tp`	Total precipitation aggregated over the basin.
	`sp`	Surface pressure (basin average).
	`ssr`	Surface net solar radiation (basin average).
	`str`	Surface net thermal (longwave) radiation (basin average).
	`prep_NOAA`	Basin-aggregated precipitation product used as an additional meteorological forcing.
Static	`area`	Drainage area of the basin.
	`mean_elevation`	Basin-average elevation above sea level.
	`mean_slope`	Basin-average terrain slope.
	`clay_fraction`	Proportion of soil classified as clay.
	`sand_fraction`	Proportion of soil classified as sand.
	`silt_fraction`	Proportion of soil classified as silt.
	`forest_cover`	Proportion of basin area covered by forest.
	`urban_cover`	Proportion of basin area covered by urban/built-up land.
	`cropland_cover`	Proportion of basin area covered by cropland/agriculture.
	`aridity_index`	Long-term climatic aridity indicator.
	`mean_precip`	Long-term mean precipitation over the basin.

Table 4: Input features used by the forecasting model.

5.1 Problem Formulation

Let $b \in \mathcal{B}$ denote a river basin and t a daily time index. The objective is to predict next-day river discharge using basin-aggregated meteorological forcings and static basin attributes. For each basin, we construct samples using a sliding-window scheme with a historical context of length L days.

Let $\mathbf{x}_t^{(b)} \in \mathbb{R}^d$ be the vector of dynamic meteorological variables for basin b on day t, and let $\mathbf{s}^{(b)} \in \mathbb{R}^m$ denote the static attribute vector of basin b. Given the historical forcing sequence up to day $t-1$ and the forcing for day t, the model predicts the discharge on day t:

$$\hat{q}_t^{(b)} = f_\theta\Big(\mathbf{x}_{t-L:t-1}^{(b)},\ \mathbf{x}_t^{(b)},\ \mathbf{s}^{(b)}\Big), \tag{1}$$

where $\mathbf{x}_{t-L:t-1}^{(b)} = \{\mathbf{x}_{t-L}^{(b)}, \ldots, \mathbf{x}_{t-1}^{(b)}\}$ is the L-day historical context, $\mathbf{x}_t^{(b)}$ is the dynamic forcing on the forecast day, and θ denotes the learnable parameters.

Training is posed as supervised regression over all valid time indices and basins. Each training example corresponds to one basin-day pair (b, t), and the model parameters θ are learned to minimize the discrepancy between the predicted discharge $\hat{q}_t^{(b)}$ and the observed discharge $q_t^{(b)}$ across the training set.

5.2 Input Features

The forecasting model uses two groups of predictors: (i) **dynamic** basin-aggregated meteorological variables that vary daily and (ii) **static** basin attributes that describe time-invariant catchment characteristics. Dynamic forcings are provided as a sequence over the historical window of length L (together with the forcing variables for the forecast day), while static attributes are provided as a fixed vector for each basin. Table 4 summarizes the input variables used in this study.

5.3 Model Architecture

The forecasting model adopts a sequence-to-sequence Long Short-Term Memory (LSTM) architecture with an encoder–decoder structure. This design allows the network to summarize long-range hydrometeorological context from past observations and use it to support short-term discharge forecasting. Importantly, the model does not require contemporaneous discharge measurements as inputs, enabling application to ungauged basins where streamflow observations may be unavailable.

Encoder. The encoder receives a historical sequence of basin-aggregated dynamic meteorological variables over a fixed history length of $L = 365$ days, together with static basin attributes. The historical sequence is processed by an LSTM to produce a compact representation of the basin state. The final hidden and cell states of the encoder are interpreted as a latent summary of antecedent conditions, capturing seasonal cycles and longer-term memory effects in the hydrological response.

Decoder. The decoder is initialized using the encoder hidden and cell states and produces discharge predictions for the next day. The decoder takes as input the dynamic meteorological variables corresponding to the forecast day (i.e., the first day ahead) and outputs the predicted discharge. This encoder-to-decoder state transfer enables information learned from the historical context to influence the short-term forecast.

State transfer and output layer. To facilitate effective information flow between the encoder and decoder, the encoder states are passed through learnable transformation networks before being used to initialize the decoder. The decoder hidden state is then mapped to the final discharge prediction through a linear output layer.

Both encoder and decoder use LSTM layers with 128 hidden units, consistent with the configuration adopted in the reference study.

5.4 Model Inputs and Data Flow

For each basin-day sample (b, t), the model consumes (i) a historical sequence of dynamic meteorological forcings, (ii) the dynamic forcings on the forecast day, and (iii) a static attribute vector describing basin characteristics. Let $\mathbf{x}_\tau^{(b)} \in \mathbb{R}^d$ denote the basin-aggregated dynamic feature vector on day τ, and let $\mathbf{s}^{(b)} \in \mathbb{R}^m$ denote the static attribute vector.

Encoder input. The encoder receives the historical forcing sequence of length L,

$$\mathbf{X}_{\text{enc}}^{(b)}(t) = \left[\mathbf{x}_{t-L}^{(b)}, \mathbf{x}_{t-L+1}^{(b)}, \ldots, \mathbf{x}_{t-1}^{(b)}\right] \in \mathbb{R}^{L\times d}, \tag{2}$$

which summarizes antecedent atmospheric conditions prior to the forecast day.

Decoder input. For one-day-ahead forecasting ($H = 1$), the decoder consumes the dynamic forcing vector on the forecast day,

$$\mathbf{X}_{\text{dec}}^{(b)}(t) = \mathbf{x}_t^{(b)} \in \mathbb{R}^d. \tag{3}$$

Static conditioning. The static attribute vector $\mathbf{s}^{(b)}$ is used to condition the sequence-to-sequence model, providing basin-specific context that modulates the rainfall–runoff relationship. In practice, $\mathbf{s}^{(b)}$ is combined with the dynamic inputs through a learned embedding and supplied to the encoder–decoder network as an additional conditioning signal.

Overall, the model predicts the next-day discharge $\hat{q}_t^{(b)}$ given the triplet $\left(\mathbf{X}_{\text{enc}}^{(b)}(t), \mathbf{X}_{\text{dec}}^{(b)}(t), \mathbf{s}^{(b)}\right)$.

5.5 Training Configuration

Data splitting protocol. To evaluate generalization across catchments, data are split at the *basin level* rather than by individual time steps. Basins are partitioned into training, validation, and test sets using an 80/10/10 split, such that all samples from a given basin belong to exactly one split. The training set is used to fit model parameters, the validation set is used for model selection and early stopping, and the test set is held out for final evaluation. This basin-wise splitting strategy prevents information leakage across splits and provides a more realistic assessment of performance in ungauged basins.

Loss function. Model parameters are learned by minimizing the mean squared error (MSE) between predicted and observed discharge values over the training set:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{q}_i - q_i\right)^2, \tag{4}$$

where q_i and $\hat{q}_i$ denote the observed and predicted discharge for sample i, respectively, and N is the number of training samples.

Optimization. Training is performed using the Adam optimizer with mini-batch gradient descent. The learning rate, batch size, and regularization settings are selected based on validation performance. To prevent overfitting, early stopping is applied by monitoring the validation loss and retaining the model checkpoint with the lowest validation error.

5.6 Evaluation Metrics

Model performance is assessed using standard regression metrics that quantify overall error magnitude and goodness-of-fit between observed discharge q_t and predicted discharge $\hat{q}_t$ over a set of N samples. Let

$$\bar{q} = \frac{1}{N} \sum_{t=1}^{N} q_t$$

denote the sample mean of observations.

Mean Squared Error (MSE). MSE measures the average squared prediction error, while RMSE reports error in the same unit as discharge:

$$\text{MSE} = \frac{1}{N} \sum_{t=1}^{N} \left(\hat{q}_t - q_t\right)^2 . \tag{5}$$

Bias. Bias indicates systematic over- or under-prediction:

$$\text{Bias} = \frac{1}{N} \sum_{t=1}^{N} \left(\hat{q}_t - q_t\right) . \tag{6}$$

Pearson Correlation Coefficient (r). To quantify the linear association between observed discharge q_t and predicted discharge $\hat{q}t$, we report the Pearson correlation coefficient:

$$r = \frac{\sum t = 1^N \left(q_t - \bar{q}\right) \left(\hat{q}t - \bar{\hat{q}}\right)}{\sqrt{\sum t = 1^N \left(q_t - \bar{q}\right)^2}; \sqrt{\sum_{t=1}^{N} \left(\hat{q}t - \bar{\hat{q}}\right)^2}}, \tag{7}$$

where $\bar{q} = \frac{1}{N} \sum t = 1^N q_t$ and $\bar{\hat{q}} = \frac{1}{N} \sum_{t=1}^{N} \hat{q}_t$ denote the sample means of the observations and predictions, respectively.

Nash–Sutcliffe Efficiency (NSE). NSE is a widely used hydrological metric comparing the predictive skill of the model to the mean-flow baseline:

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{N} \left(\hat{q}_t - q_t\right)^2}{\sum_{t=1}^{N} \left(q_t - \bar{q}\right)^2}. \tag{8}$$

NSE values close to 1 indicate high predictive skill, while NSE = 0 corresponds to a model that performs comparably to using the mean observed discharge as prediction.

5.7 Rationale for Model Selection

The encoder–decoder LSTM architecture is particularly well suited for large-scale hydrological forecasting due to its ability to capture nonlinear temporal dependencies, operate without local calibration, and generalize to ungauged basins. Compared to traditional models based on physical bases, this data-driven approach significantly reduces calibration requirements while maintaining competitive predictive skill, especially for extreme flood events.

6 Implementation Results

6.1 Baseline Forecasting Performance

We first evaluate the baseline encoder–decoder LSTM model described in Section 5 for one-day-ahead discharge forecasting. Overall, the baseline model achieves strong agreement with observed discharge dynamics, reflected by a high correlation and positive NSE on the test set. The near-zero Bias indicates that the baseline configuration produces well-calibrated predictions without systematic over- or under-estimation. Detailed results are reported in Table 5.

In comparison, the Caravan baseline attains a lower RMSE but exhibits near-zero correlation and a strongly negative NSE, indicating that it fails to capture temporal discharge variability and performs worse than a mean-flow predictor under the constructed test set. This contrast highlights that minimizing absolute error alone is insufficient for hydrological skill and underscores the importance of preserving temporal dynamics when evaluating streamflow prediction performance.

Setting	**RMSE** ↓	**Bias** ↓	r ↑	**NSE** ↑
Caravan	**1.839**	-0.916	0.044	-1.093
Ours	2.094	**-0.001**	**0.806**	**0.634**

Table 5: One-day-ahead streamflow prediction results on the constructed test set.

6.2 Ablation Study

An ablation study is conducted to assess the contribution of individual input components in the constructed dataset. All ablation configurations are evaluated using the same modeling setup adopted from existing studies to ensure a consistent and fair comparison. The ablation results are summarized in Table 6.

The first ablation examines the role of meteorological precipitation (`tp`). This experiment is motivated by the observation that precipitation forcing is already explicitly provided through NOAA-based rainfall products, as well as by the identified mismatch between hindcast and forecast representations of `tp`. As shown in Table 6, excluding `tp` from the meteorological inputs leads to an approximately 3.3% reduction in RMSE, together with 3.0% and 3.8% increases in correlation and NSE, respectively, indicating more accurate and skillful discharge predictions. Overall, these results suggest that, in the presence of NOAA-derived precipitation, meteorological precipitation `tp` provides limited additional information and may introduce redundancy due to inconsistencies between data regimes.

Secondly, the contribution of NOAA-based precipitation is then assessed by removing it from the full feature set. Compared to the full-variable setting, RMSE increases by approximately 5%, bias deteriorates markedly, while both correlation and NSE decrease by over 30%. These results show that meteorological and static variables alone cannot adequately capture discharge dynamics, confirming NOAA-based precipitation as an essential and non-substitutable forcing component for robust streamflow modeling.

The final ablation examines the effect of removing static catchment attributes while retaining all dynamic forcing variables. Excluding static attributes results in a substantial performance degradation, with RMSE increasing by approximately 36.3% and NSE decreasing by 49.5% relative to the baseline configuration. This degradation is further reflected by a 5.3% reduction in explained variance and a marked increase in bias. These results demonstrate that static catchment attributes play a critical role in constraining hydrological responses by linking meteorological forcing to basin-scale discharge behavior. Without this contextual information, the model struggles to generalize across basins with differing geomorphological and hydro-climatic properties, underscoring the necessity of including static attributes in the constructed dataset.

Setting	**RMSE** ↓	**Bias** ↓	r ↑	**NSE** ↑
All inputs	2.094	**-0.001**	0.806	0.634
w/o `tp`	**2.025**	-0.173	**0.830**	**0.658**
w/o `precip_NOAA`	2.899	-0.421	0.775	0.597
w/o static attributes	2.853	0.925	0.784	0.320

Table 6: Ablation study of input features for one-day-ahead streamflow prediction.

7 Discussion

This section discusses the key insights derived from the experimental results and reflects on the main challenges encountered during dataset construction and analysis.

A central insight from this study is that reliable streamflow prediction in ungauged basins critically depends on the integration of multiple complementary data sources, including dynamic meteorological variables, precipitation forcing, and static catchment attributes. The ablation results demonstrate that each of these components plays a distinct role in constraining hydrological responses, and that excluding essential inputs—such as NOAA-based precipitation or static basin characteristics—leads to substantial performance degradation. At the same time, the analysis reveals that not all precipitation-related variables contribute equally. In particular, the simultaneous inclusion of meteorological precipitation (`tp`) alongside explicitly provided precipitation forcing introduces redundancy and, in some cases, degrades predictive performance. This finding highlights the importance of careful feature selection in multi-source hydrological datasets, as redundant inputs may introduce noise or inconsistencies rather than additional information.

Several practical challenges were encountered during the construction of the dataset. Processing large volumes of raw meteorological data stored in NetCDF (`.nc`) and GRIB formats proved computationally intensive and time-consuming, particularly for high-resolution reanalysis and numerical weather prediction products. These formats store multi-dimensional spatio-temporal fields, resulting in large file sizes and non-trivial input–output and memory overhead during data access and processing. To maintain feasibility within available computational resources, the spatial and temporal scope of the dataset was therefore constrained to South America for the period 2019–2025.

In addition, substantial effort was devoted to optimizing data extraction and preprocessing workflows, including spatial subsetting, basin-level aggregation, and temporal alignment, in order to reduce computational cost while preserving physical consistency and data integrity. These design choices represent a trade-off between dataset coverage and practical scalability, and they establish a foundation for future extensions of the dataset as computational resources and processing capacity permit.

References

1. Günter Blöschl et al. Twenty-three unsolved problems in hydrology (uph)—a community perspective. *Hydrological Sciences Journal*, 64(10):1141–1158, 2019.
2. Hylke E. Beck et al. Global evaluation of runoff simulations and data scarcity in ungauged basins. *Nature*, 2020.
3. Markus Reichstein et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 2019.
4. Chaopeng Shen. A transdisciplinary review of deep learning in water resources scientists. *Water Resources Research*, 54(11):8565–8598, 2018.
5. Keith J Beven. *Rainfall-runoff modelling: the primer*. John Wiley Sons, 2011.
6. Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Daniel Klotz, Frederik Kratzert, Asher Metzger, Sella Nevo, Florian Pappenberger, Christel Prudhomme, Guy Shalev, and Shlomo Shenzis. Global prediction of extreme floods in ungauged watersheds. *Nature*, 627(8004):559–563, 2024.
7. Frederik Kratzert et al. Caravan: A global community dataset for large-sample hydrology. *Scientific Data*, 10:61, 2023.
8. Felix W. Landerer and Sean C. Swenson. Accuracy of scaled grace terrestrial water storage estimates. *Water Resources Research*, 48, 2012.
9. G. Pastorello et al. The fluxnet2015 dataset and the oneflux processing pipeline. *Scientific Data*, 7:225, 2020.
10. M. Rodell et al. The global land data assimilation system. *Bulletin of the American Meteorological Society*, 85:381–394, 2004.
11. G. P. Weedon et al. The wfdei meteorological forcing data set: Watch forcing data methodology applied to era-interim. *Water Resources Research*, 50:7505–7514, 2014.
12. Global Runoff Data Centre. Grdc data download: River discharge data online. https://waterandchange.org/en/grdc-data-download-provides-river-discharge-data-online/, 2024.
13. National Oceanic and Atmospheric Administration. Noaa precipitation data products. https://www.ncei.noaa.gov, 2024. Accessed 2024.
14. George J. Huffman, Emily F. Stocker, David T. Bolvin, Eric J. Nelkin, and Jackson Tan. Integrated multi-satellite retrievals for the global precipitation measurement (gpm) mission (imerg). *Journal of Hydrometeorology*, 20(4):729–751, 2019.
15. Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis, 2020.
16. Philippe Bougeault, Zoltan Toth, Charles Bishop, Brett Brown, David Burridge, Dongmei Chen, Elizabeth Ebert, Mariano Fuentes, Thomas M Hamill, Kevin Mylne, et al. The THORPEX interactive grand global ensemble (TIGGE), 2010.
17. H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut. Era5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2023. Accessed on 06-Jan-2026.