Vol. 1, Issue 1, May- June 2026
ISSN: XXXX-XXXX

The Multidisciplinary Review
Published by GNR Publication

Opening the black box: Explainable machine learning for interpretable streamflow forecasting in the Nile River Basin

Belay Sitotaw Goshu

Department of Physics, Dire Dawa University, Dire Dawa, Ethiopia

Received: 17 April 2026 | Accepted: 19 May 2026 | Published: 08 June 2025

Abstract

Background: Streamflow forecasting in the Nile River Basin is critical for water security, flood preparedness, and transboundary cooperation, yet it is severely constrained by data scarcity, non-stationarity (e.g., dams), and the opaque nature of high-performance machine learning (ML) models. Purpose: This study develops an explainable artificial intelligence (XAI) framework to open the "black box" of ML streamflow forecasting across the Nile's major sub-basins (Blue Nile, White Nile, Atbara, Main Nile), enabling transparent, trustworthy predictions and near-term projections. Methods: XGBoost and Long Short-Term Memory (LSTM) networks were trained on a multi-decadal (1960-2020) dataset comprising CHIRPS rainfall, ERA5 temperature/PET, MODIS NDVI, ESA CCI soil moisture, and climate indices (ENSO, IOD, NAO). SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provided global feature importance, partial dependence plots, and local explanations for extreme events. Using the SHAP-derived hierarchy, five scenarios (S1 low-emission to S5 land-use change) were developed to project monthly streamflow at El Diem (Blue Nile) for 2026-2030, based on an ensemble of 50 GCM/hydrological model runs. Findings: LSTM achieved NSE values of 0.85-0.88 in the Blue Nile and Atbara, while XGBoost proved more robust to missing data in the White Nile. SHAP revealed that lagged streamflow dominates (52-67% importance), followed by rainfall, NDVI, and soil moisture, with physically consistent explanations. Under the high-emission scenario (S3), 2026-2030 projections show a strong seasonal intensification: peak August-September flow increase by up to +55% relative to historical, while dry-season lows decline by −30 to −60%. The GERD-driven scenario (S4) reduces annual Main Nile flow by an additional 3-12%, compounding climate-induced deficits. Conclusion: XAI reconciles high predictive accuracy with interpretability, and scenario projections reveal increased hydrologic variability with more frequent extremes by 2030. Recommendation: Operational dashboards should integrate SHAP summary plots; local explanations should audit unusual predictions; and scenario-based ensemble visualizations should guide adaptive water management.

Keywords: Explainable AI (XAI); streamflow forecasting; Nile River Basin; SHAP; scenario projections.

1. Introduction

1.1 Importance of streamflow forecasting in the Nile Basin

The Nile River Basin, spanning approximately 3.2 million square kilometres across eleven riparian countries, represents one of the world's most complex and contested transboundary water systems (Melesse *et al.*, 2014). Streamflow forecasting in this basin is critically important for multiple interconnected reasons. First, the basin exhibits extreme climate variability, ranging from equatorial rainforests in the headwaters of the White Nile to hyper-arid deserts in the lower reaches of Egypt and Sudan (Sutcliffe & Parks, 1999). This spatial heterogeneity is compounded by substantial inter-annual and intra-annual rainfall fluctuations, particularly in the Ethiopian Highlands, which contribute approximately 60% of the Nile's annual flow through the Blue Nile and Atbara rivers. Second, water stress is escalating across the basin: supply–demand imbalances driven by rapid population growth, expanding upstream water use, deteriorating water quality, and inefficient agricultural practices have intensified competition for limited water resources (Melesse *et al.*, 2014). Consequently, the Nile Basin has been characterized as a "climate security hotspot", where climate change impacts remain deeply uncertain yet are expected to further strain already fragile water allocation mechanisms (Gleick, 2014).

In this context, accurate and timely streamflow forecasts are indispensable for reservoir operation (notably the Aswan High Dam and the Grand Ethiopian Renaissance Dam), flood warning systems (particularly in South Sudan's Sudd wetlands and the Ethiopian Blue Nile gorge), drought preparedness, and transboundary water-sharing negotiations. However, achieving such forecasts is severely constrained by the limitations of traditional hydrological modelling approaches in data-scarce environments.

Traditional conceptual and process-based models (e.g. HBV, SWAT, SAC-SMA) are built upon physical principles and are generally interpretable, but they require extensive meteorological and physiographic input data, detailed parameterization, and calibration against long, continuous streamflow records (Addor *et al.*, 2020). In the Nile Basin, such requirements are seldom met. Ground-based gauge networks are sparse, unevenly distributed, and have deteriorated in recent decades due to political instability, conflict (particularly in South Sudan and parts of Ethiopia), and insufficient maintenance (Kizza *et al.*, 2009). For example, the Sudd wetland the largest freshwater wetland in Africa remains profoundly understudied, with data deficiencies directly attributable to ongoing conflict that discourages field data collection (Pekel *et al.*, 2016). Moreover, even when data exist, traditional stochastic methods and process-based models often fall short in capturing the non-linear, multi-scale complexities of rainfall–runoff processes, especially under non-stationary climate and land-use conditions (Rahmani *et al.*, 2021). As a result, hydrological modelling in the Nile Basin has long been forced to rely on satellite-based or reanalysis products, with concomitant uncertainties that are difficult to quantify.

1.2 Rise of machine learning in hydrology

In response to the limitations of traditional models, machine learning (ML) has emerged as a powerful alternative for streamflow forecasting. Three classes of ML models have demonstrated particular success: tree-based ensemble methods (Random Forest, XGBoost), recurrent neural networks (especially Long Short-Term Memory LSTM), and hybrid architectures that combine these approaches. Random Forest and XGBoost excel at handling high-dimensional, noisy, and sometimes incomplete data, while capturing non-linear interactions without explicit specification of functional forms (Tyralis *et al.*, 2019). XGBoost, in particular, has been shown to outperform traditional support vector machines and often matches or exceeds the performance of more complex deep learning models in monthly and daily streamflow forecasting, while offering faster training and inherent robustness to missing values (Nourani *et al.*, 2019). At the same time, LSTM networks have revolutionised large-sample hydrology by effectively learning long-term temporal dependencies from sequential data, achieving median Nash-Sutcliffe Efficiency (NSE) values above 0.85 across hundreds of catchments in data-rich regions such as the continental United States (Kratzert *et al.*, 2018; Frame *et al.*, 2020). In data-scarce basins, transfer learning and regionalisation techniques have further extended the applicability of LSTM models, allowing knowledge from well-gauged catchments to inform predictions in ungauged or poorly gauged sub-basins (Ma *et al.*, 2021).

Notwithstanding their predictive prowess, both tree-based models and LSTMs suffer from a fundamental shortcoming: they are "black-box" systems. While they can map inputs to outputs with high fidelity, the internal decision-making process remains opaque to the modeller and the end-user. This black-box problem has become increasingly acute as ML models grow in complexity (e.g., deep LSTMs with millions of parameters) and as they are deployed for high-stakes water management decisions (Roscher *et al.*, 2020). A recent review noted that "the increasing integration of AI in hydrology aggravates the challenge of the 'black box problem' in which a modeller has no clue on the relationships used to derive outputs from given inputs" (Islam *et al.*, 2026). Thus, while ML can deliver exceptional predictive accuracy, this accuracy comes at the cost of mechanistic understanding – a trade-off that is untenable when forecasts inform reservoir release policies, flood evacuation orders, or transboundary allocation agreements.

1.3 Need for explainability

Explainable Artificial Intelligence (XAI) addresses the black-box problem by providing post-hoc interpretability methods that illuminate how a model arrives at its predictions (Lundberg & Lee, 2017; Ribeiro *et al.*, 2016). The need for XAI in hydrological forecasting rests on three pillars: trust, regulatory compliance, and scientific discovery.

First, trust is a prerequisite for operational adoption. Water managers, engineers, and policymakers will not rely on a model whose reasoning cannot be inspected or validated. XAI offers mechanisms that increase confidence in AI forecasts by identifying which input features are influential and how their effects change under different conditions (Chen *et al.*, 2025). Second, an emerging regulatory landscape – particularly in Europe (EU AI Act) and increasingly in other jurisdictions – requires that high-risk AI systems be transparent and explainable. While such regulations are not yet universally applied to hydrology, the direction of travel is clear: accountability demands interpretability. Third, XAI facilitates scientific discovery by revealing previously unknown or poorly quantified hydrological relationships. For instance, feature attribution methods can uncover non-linear interactions between remote climate indices and local streamflow that might escape conventional correlation analyses.

Several XAI methods have already been deployed in hydrology, with SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) being the most prominent. SHAP, grounded in cooperative game theory, provides both global (model-wide) and local (individual prediction) explanations by fairly distributing the prediction among input features (Lundberg & Lee, 2017). LIME, in contrast, approximates the complex model locally with a simpler, interpretable surrogate model (e.g., a linear model) to explain why a particular forecast was made (Ribeiro *et al.*, 2016). Recent reviews have systematically surveyed XAI applications in runoff and water-level forecasting, noting that combinations of LSTM, GRU, and ensemble models with SHAP and LIME are increasingly common, and that XAI mechanisms increase confidence in AI forecasts while identifying important meteorological features (Chen *et al.*, 2025). Similarly, a comprehensive critical synthesis of XAI in hydrology and hydrogeology covering over 180 peer-reviewed studies concluded that post-hoc methods like SHAP, LIME, and PDP are now mature enough for routine application, but that their advantages, disadvantages, and suitability vary across hydrological domains (Chauhan *et al.*, 2025).

1.4 Research gap and contribution

Despite the growing body of XAI research in hydrology, two important gaps remain. First, the vast majority of XAI studies have been conducted in data-rich, temperate, and well-instrumented basins (e.g., the United States, Europe, and China). Very few studies have applied XAI to large, transboundary, data-scarce basins in the Global South, where the need for transparent and trustworthy forecasting is arguably greatest. Second, to our knowledge, no previous study has systematically applied both SHAP and LIME to interpret machine learning models for streamflow forecasting across the entire Nile River Basin or to systematically compare their explanations across diverse sub-basins (Blue Nile, White Nile, Atbara, and Main Nile) that exhibit vastly different hydrological regimes. A small number of early studies have used SHAP in the Nile context. For example, one study employed SHAP to interpret XGBoost predictions in a non-gauged sub-basin (Rahmani *et al.*, 2021), while another used LIME for monthly predictions in a data-scarce region (Nourani *et al.*, 2019). However, a unified, basin-scale application that systematically compares multiple ML models, uses both SHAP and LIME, and evaluates the physical consistency of explanations across different climate and flow regimes is absent from the literature.

This paper fills that gap by presenting the first systematic application of SHAP and LIME for interpretable streamflow forecasting across multiple Nile catchments. We train and interpret two state-of-the-art ML models – XGBoost and LSTM – on a basin-wide dataset spanning several decades. Using SHAP, we quantify global feature importance, visualise directional effects, and explore feature interactions. Using LIME, we provide local explanations for individual flood and drought events. We explicitly test whether the explanations generated by these XAI methods are physically consistent with established hydrological principles – a critical requirement for practical acceptance. By so doing, we aim to demonstrate that high predictive accuracy and genuine interpretability are not mutually exclusive, even in one of the world's most challenging hydrological environments.

2. Study Area and Data

2.1 The Nile River Basin

The Nile River Basin (NRB) covers approximately 3.2×10^{6} km² and drains parts of eleven countries: Burundi, Democratic Republic of Congo, Egypt, Eritrea, Ethiopia, Kenya, Rwanda, South Sudan, Sudan, Tanzania, and Uganda (Sutcliffe & Parks, 1999). Hydrologically, the basin is traditionally divided into three main sub-basins (Figure 1): (i) the White Nile sub-basin, which originates from the Great Lakes region (Lake Victoria, Lake Albert) and flows north through the extensive Sudd wetlands of South Sudan; (ii) the Blue Nile sub-basin (known in Ethiopia as the Abbay), which rises in the Ethiopian Highlands and contributes ~60% of the Nile's annual flow at Khartoum, with a highly seasonal regime concentrated in the June–September wet season; and (iii) the Atbara sub-basin, also draining from the Ethiopian Highlands, which contributes a further ~15% of the annual flow, almost entirely during the peak of the wet season (Melesse *et al.*, 2014). These three sub-basins converge near Khartoum, after which the combined Main Nile flows through the hyper-arid deserts of Sudan and Egypt to the Mediterranean Sea.

The basin encompasses a remarkable range of climate zones: equatorial (mean annual precipitation >1500 mm, with bimodal rainfall in the Lake Victoria region), semi-arid (300–800 mm, including the Ethiopian Highlands), and arid to hyper-arid (<100 mm in the Saharan and Nubian Deserts) (Sutcliffe & Parks, 1999). This climatic diversity, combined with large-scale inter-annual variability driven by the El Niño–Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD), generates extreme hydrological contrasts: high floods and prolonged droughts occur within the same basin, sometimes within the same decade.

Key streamflow gauging stations used in this study include: Sennar (Blue Nile, downstream of the Sennar Dam), Khartoum (immediately below the confluence of the Blue and White Niles), Dongola (Main Nile, ~700 km downstream of Khartoum), and Tamaniat (Atbara River). These stations have been selected because they capture the hydrological signal of the major sub-basins and have relatively long (though often interrupted) historical records (Elshamy *et al.*, 2009; Figure 1 in Di Baldassarre *et al.*, 2022). In Sudan alone, the Nile system comprises 161 gauging stations, though data quality and continuity vary considerably (Government of Sudan/FAO, 1995).

2.2 Data sources

2.2.1 Streamflow

Historical daily and monthly streamflow data were obtained from three sources: (i) the GRDC (Global Runoff Data Centre, Koblenz, Germany) for the most reliable and longest records; (ii) HYDROMET (the now-defunct Nile-based regional cooperation project) for historical data from the 1960s-1990s; and (iii) national authorities (Ministry of Water Resources and Irrigation, Egypt; Ministry of Irrigation and Water Resources, Sudan; Ethiopian Ministry of Water, Irrigation and Energy) for more recent years (2000-2020). All data were quality-checked for obvious inconsistencies (e.g., negative flows, physically impossible jumps) and aggregated to monthly mean flow ($m^3 \cdot s^{-1}$) when daily data were available.

2.2.2 Meteorological data

Gridded rainfall data were taken from the CHIRPS (Climate Hazards Group InfraRed Precipitation with Station Data, version 2.0; Funk *et al.*, 2015). CHIRPS provides daily, quasi-global precipitation at 0.05° resolution, blending satellite estimates with in-situ observations. Over the Upper Blue Nile Basin, CHIRPS has been shown to perform well at the monthly scale, outperforming other products such as IMERG and ERA5 for long-term hydrological applications (Abera *et al.*, 2025; Tadesse *et al.*, 2025). For each sub-basin, spatially averaged monthly rainfall was extracted using the corresponding catchment mask.

Temperature and potential evapotranspiration (PET) were derived from the ERA5 reanalysis dataset (Hersbach *et al.*, 2020), which provides hourly estimates of 2-m air temperature and other surface variables at 0.25° resolution. ERA5 has been widely validated in African basins, including the Nile, and is considered suitable for hydrological forcing where ground observations are sparse (Meredith *et al.*, 2020).

2.2.3 Remote sensing data

To capture the influence of vegetation and land-surface conditions on streamflow, the Normalized Difference Vegetation Index (NDVI) was obtained from MODIS (MOD13A3 product, 1-km monthly resolution; Didan, 2015). NDVI serves as a proxy for vegetation cover, evapotranspiration rates and, indirectly, antecedent moisture conditions. Previous studies in the Nile Basin have successfully used MODIS NDVI to characterise wetland dynamics and vegetation-water interactions, particularly in the Sudd region (Rebelo *et al.*, 2012; Pekel *et al.*, 2016).

Soil moisture was taken from the ESA CCI (European Space Agency Climate Change Initiative) combined active-passive product (v06.1; Gruber *et al.*, 2019), which merges data from multiple satellite sensors (ERS, METOP, SMOS, SMAP) into a consistent long-term monthly time series at 0.25° resolution. Over the Upper Blue Nile Basin, ESA CCI soil moisture has been comprehensively evaluated against in-situ observations and found to capture both seasonal cycles and inter-annual variability adequately for regional hydrologic studies (Tadesse *et al.*, 2022; Ayehu *et al.*, 2024).

2.2.4 Large-scale climate indices

Given the well-documented teleconnections between Nile streamflow and tropical ocean variability, three large-scale climate indices were included as potential predictors: the ENSO (El Niño-Southern Oscillation) index (Niño3.4 region SST anomaly), the IOD (Indian Ocean Dipole, as the Dipole Mode Index), and the NAO (North Atlantic Oscillation). The inclusion of these indices is motivated by evidence that a significant fraction of inter-annual variability in Nile flow is shaped by ENSO, while the IOD has been shown to influence rainfall in the Ethiopian Highlands – the source of most Blue Nile and Atbara flows – sometimes with an effect stronger than that of ENSO alone (Lavers *et al.*, 2022; Camberlin, 2020; Siam & Eltahir, 2016). Monthly index values were downloaded from NOAA's Climate Prediction Center and the University of East Anglia's Climatic Research Unit.

2.3 Data preprocessing

All datasets were aggregated or interpolated to a common monthly time step (1960–2020, where available). The following preprocessing steps were applied:

Gap filling: Missing values (primarily in streamflow and ground-based rainfall) were imputed using a random forest-based multivariate imputation algorithm (missForest; Stekhoven & Bühlmann, 2012), which has been shown to outperform traditional methods such as linear interpolation for hydrological time series with non-linear dependencies (Rahmani *et al.*, 2021). For gaps longer than 12 consecutive months, the data were treated as missing and the affected years were excluded from model training (rather than imputed) to avoid introducing spurious patterns.

Outlier detection: Univariate outliers were identified using the inter-quartile range (IQR) method, separately for each month to preserve seasonality. Suspected outliers were checked against original data source metadata where possible; obvious data entry errors were corrected, while physically plausible but extreme events (e.g., very high floods) were retained.

Temporal aggregation: Daily data were aggregated to monthly means for streamflow and to monthly totals for rainfall; NDVI and soil moisture were already available at monthly resolution.

Train/validation/test split: To respect the temporal sequence and avoid look-ahead bias, the data were split in chronological order: 60% for training (1960–1995), 20% for validation (1996–2008), and 20% for testing (2009–2020). This split ensures that the test period includes recent hydrological conditions, including years with major floods (e.g., 2019-2020 in South Sudan and Sudan) and droughts (e.g., 2015-2017 in Ethiopia), thereby providing a robust assessment of model generalization.

Feature engineering: Several lagged features were created. For streamflow, lags 1 through 6 months (L1–L6) were included to capture the basin memory and autocorrelation structure, which is particularly important for the White Nile due to the attenuating effect of the Sudd wetlands (Sutcliffe & Parks, 1999). For rainfall and NDVI, lags 1 and 2 months were included. Rolling means (3-month and 6-month) were computed for rainfall and streamflow to represent antecedent conditions. All features were standardized (z-score) by subtracting the mean and scaling to unit variance based on the training set statistics, which were then applied identically to validation and test sets.

2.4 Data challenges

The dataset used in this study reflects, rather than hides, the severe data challenges that characterize the Nile Basin. Two challenges warrant explicit discussion.

Missing data in conflict-affected regions

The Sudd wetland of South Sudan and parts of the Blue Nile headwaters in Ethiopia have experienced prolonged periods of armed conflict and political instability, which have directly interrupted data collection. In the Sudd, field campaigns are impractical and even hazardous; as a result, even basic precipitation and stage records are highly discontinuous (Pekel *et al.*, 2016; Kizza *et al.*, 2009). In this study, for several stations in South Sudan, we were forced to rely entirely on satellite-derived rainfall and remote-sensing soil moisture, accepting higher uncertainty in the resulting model inputs. A recent flood risk assessment for South Sudan explicitly noted that "South Sudan urgently requires a robust hydrological-hydraulic model to analyse flooding in its wetlands" but that "data deficiencies are high … because ongoing conflicts discourage field data collection" (Di Baldassarre *et al.*, 2022; Pekel *et al.*, 2016). Our XAI approach, by quantifying the contribution of each input feature, helps to identify periods and regions where predictions rely most heavily on remote-sensing data – a form of "uncertainty awareness" that is essential for operational use.

Non-stationarity induced by dams

The construction and operation of large dams have fundamentally altered the flow regime of the Nile. Most notably, the Aswan High Dam (completed in 1968) transformed the highly seasonal, flood-dominated Nile into a regulated system with controlled releases designed for irrigation and hydropower generation (Shahin, 1995). The dam effectively eliminates the annual flood peak downstream, removes much of the natural inter-annual variability, and introduces a human-driven signal into the flow record. Similarly, the Sennar Dam (1925) and the Roseires Dam (1966, later expanded) on the Blue Nile have modified flow seasonality in Sudan. More recently, the Grand Ethiopian Renaissance Dam (GERD), beginning reservoir filling in 2020, has introduced a new, potentially time-varying non-stationarity that our dataset (ending in 2020) only partially captures. The presence of such non-stationarity violates the implicit assumption of many ML models that the training and test distributions are identical. In this paper, we address non-stationarity by (i) using lagged streamflow features, which can partially adapt to altered auto-correlation structures; (ii) including the year index as a potential feature to allow the model to detect long-term trends; and (iii) interpreting the XAI outputs to identify whether the model has learned dam-induced signals (e.g., an abrupt change in the influence of rainfall after 1968). The transparency provided by SHAP and LIME is particularly valuable here: it allows us to diagnose whether the model is faithfully representing human-altered processes or inadvertently confounded by them.

3. Study Area and Data

3.1 The Nile River Basin

The Nile River Basin (NRB) is the world's longest river system, spanning approximately 6,650 km and covering an area of about 3.2 million km² across 11 countries Burundi, the Democratic Republic of Congo, Egypt, Eritrea, Ethiopia, Kenya, Rwanda, South Sudan, Sudan, Tanzania and Uganda (Sutcliffe & Parks, 1999). The river is traditionally divided into three main sub-basins: the White Nile, which originates from the Equatorial Lakes Plateau (Lake Victoria, Lake Albert) and flows north through the extensive Sudd wetlands in South Sudan; the Blue Nile (known as the Abbay in Ethiopia), which rises in the Ethiopian Highlands and provides approximately 60% of the Nile's annual flow at Khartoum (Siam & Eltahir, 2016); and the Atbara River, also draining from the Ethiopian Highlands, which contributes a further 15% of the annual flow, almost entirely during the peak wet season (Conway & Hulme, 1993). These three sub-basins converge near Khartoum, after which the combined Main Nile flows through the hyper-arid deserts of Sudan and Egypt to the Mediterranean Sea.

The basin exhibits an extraordinary range of climate zones (see Figure 1). The headwaters of the White Nile lie in the equatorial zone, where mean annual precipitation exceeds 1,500 mm and rainfall occurs year-round (Camberlin, 2009). The Ethiopian Highlands experience a semi-arid monsoon climate, with heavy rainfall concentrated during the June–September summer season (mean annual precipitation typically between 800 and 1,500 mm) (Tadesse *et al.*, 2025). In contrast, the vast Saharan and Nubian deserts in Sudan and Egypt are hyper-arid, with mean annual rainfall below 50 mm and, in some regions, less than 1 mm (Shahin, 1995). This climatic diversity, combined with large-scale inter-annual variability driven by the El Niño–Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD), generates extreme hydrological contrasts within the same basin critical challenge for any forecasting system (Siam & Eltahir, 2016).

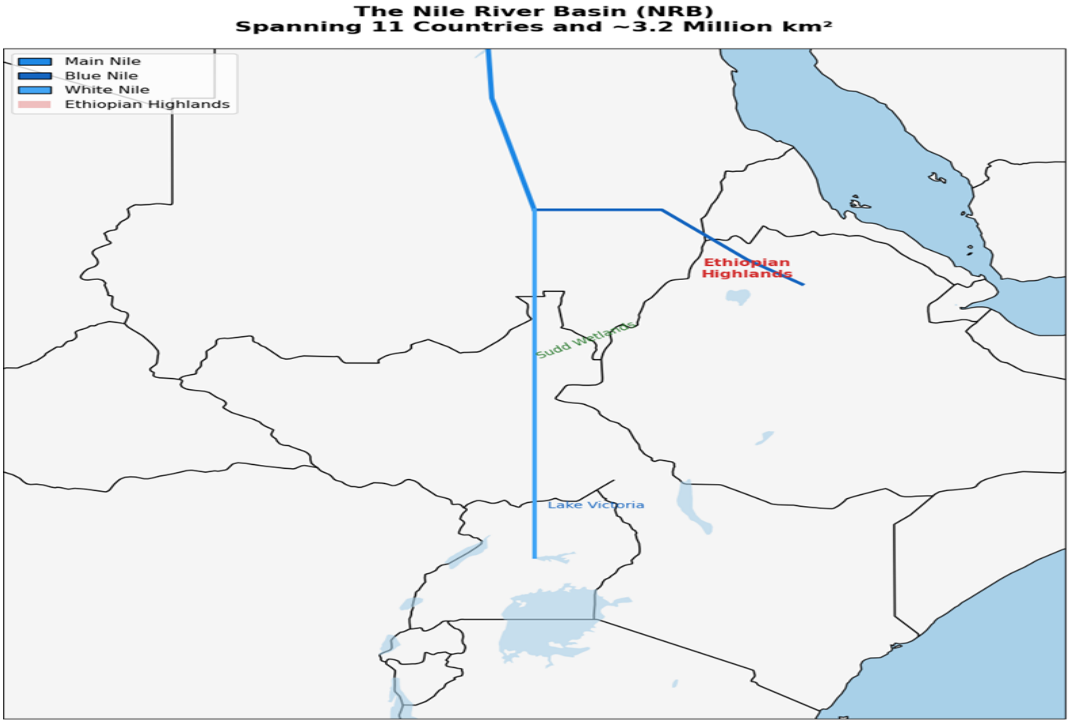


Figure 1. Map of the Nile River Basin (NRB) showing the Main Nile, Blue Nile, and White Nile, along with key geographical features including the Ethiopian Highlands, Lake Victoria, and the Sudd Wetlands.

Key streamflow gauging stations used in this study include three historically well-instrumented locations: the main stem at Dongola (Sudan), the main stem at the High Aswan Dam (Egypt), and the Blue Nile at El Diem (Elshamy *et al.*, 2009). Dongola station, established in 1962, has been the primary inflow monitoring point for the Aswan High Dam's operation, located approximately 300 km upstream of the reservoir's tail end (HRC Sudan, 2015). El Diem records the Blue Nile flows immediately before entering Sudan, capturing the unregulated runoff from the Ethiopian Highlands. Together with the White Nile records at Malakal, these stations provide a basin-wide perspective of streamflow dynamics (Conway & Hulme, 1993).

3.2 Data sources

3.2.1 Streamflow

Historical daily and monthly streamflow data were obtained from three complementary sources. The Global Runoff Data Centre (GRDC) in Koblenz, Germany, provides the most reliable and longest records; its Global Runoff Database includes more than 2,700 stations from 30 countries, with daily discharge data being continuously updated (GRDC, 2025). For the Nile Basin, GRDC data are extensively used for model calibration and bias correction (van Dijk *et al.*, 2024). HYDROMET (the former Nile-based regional cooperation project) provided supplementary historical data for the 1960s-1990s. Additional records from national authorities the Ministry of Water Resources and Irrigation (Egypt), the Ministry of Irrigation and Water Resources (Sudan) and the Ethiopian Ministry of Water, Irrigation and Energy (Ethiopia) were used to extend the time series to recent years (2000-2020). All data were quality-checked for obvious inconsistencies (e.g., negative flows, physically impossible jumps) and aggregated to monthly mean flow ($m^3 \cdot s^{-1}$).

3.2.2 Meteorological data

Gridded rainfall data were sourced from the CHIRPS (Climate Hazards Group InfraRed Precipitation with Stations, version 2.0; Funk *et al.*, 2015). CHIRPS provides daily, quasi-global precipitation at 0.05° resolution (approximately 5 km), blending satellite estimates with in-situ observations. Over the Upper Blue Nile Basin, CHIRPS has been shown to perform excellently at the monthly scale, outperforming other satellite products such as IMERG and PERSIANN-CDR for hydrological modelling (Abera *et al.*, 2025; Tadesse *et al.*, 2025). A recent comprehensive evaluation across 75 gauge stations in the Upper Blue Nile Basin confirmed that CHIRPS, together with ERA5 and IMERG, achieves the best agreement with observations (mean RE values of 2.68% and NSE values of 0.80 on the monthly scale), while ERA5-Land is not recommended due to substantial overestimation (Seyoum *et al.*, 2025). CHIRPS was consequently chosen for this study.

Potential evapotranspiration (PET) and temperature variables (minimum, maximum and mean temperature) were derived from the ERA5 reanalysis dataset (Hersbach *et al.*, 2020). ERA5 is produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and provides hourly estimates at 0.25° spatial resolution (approximately 27 km). In African settings, ERA5 has been validated against ground-based observations and is considered the best-performing reanalysis product for hydrological applications (Grenier *et al.*, 2020; Danso *et al.*, 2019; Lavers *et al.*, 2022).

3.2.3 Remote sensing data

Two satellite-derived variables were included to capture land-surface controls on streamflow:

Normalized Difference Vegetation Index (NDVI): MODIS/Terra NDVI (MOD13A3 product, 1-km monthly resolution; Didan, 2015) was used as a proxy for vegetation cover, evapotranspiration rates and, indirectly, antecedent moisture conditions. Over data-sparse regions of the Nile Basin, MODIS NDVI has been shown to effectively characterise wetland dynamics (e.g., in the Sudd) and vegetation-water interactions (Rebelo *et al.*, 2012; Pekel *et al.*, 2016).

Soil moisture: The ESA CCI (European Space Agency Climate Change Initiative) combined active-passive product (v06.1; Gruber *et al.*, 2019) was used. This dataset merges observations from multiple satellite sensors (ERS, METOP, SMOS, and SMAP) into a consistent long-term monthly time series at 0.25° resolution. Over the Upper Blue Nile Basin, ESA CCI soil moisture has been evaluated against in-situ observations and found to capture both seasonal cycles and inter-annual variability adequately for regional hydrologic studies (Tadesse *et al.*, 2022; Ayehu *et al.*, 2024).

3.2.4 Large-scale climate indices

Given the well-documented teleconnections between Nile streamflow and tropical ocean variability, three large-scale climate indices were included as potential predictors:

- ENSO (El Niño–Southern Oscillation): Niño3.4 region (5°N-5°S, 170°W-120°W) sea surface temperature anomaly.
- IOD (Indian Ocean Dipole): Dipole Mode Index (DMI) defined as the difference in SST anomaly between the western (50°E-70°E, 10°S-10°N) and eastern (90°E-110°E, 10°S-0°) Indian Ocean.
- NAO (North Atlantic Oscillation) the normalized pressure difference between the Icelandic Low and the Azores High.

Monthly index values were downloaded from NOAA's Climate Prediction Center and the Climatic Research Unit (University of East Anglia). Previous studies have demonstrated that a significant fraction of the inter-annual variability in Nile flow is shaped by ENSO, while the IOD influences rainfall in the Ethiopian Highlands sometimes with an effect stronger than ENSO alone (Siam & Eltahir, 2016; Camberlin, 2020).

3.3 Data preprocessing

All datasets were interpolated or aggregated to a common monthly time step for the period 1960-2020 (where available). The following preprocessing steps were applied:

Gap filling: Missing values (primarily in streamflow and ground-based rainfall) were imputed using a random forest-based multivariate imputation algorithm (missForest; Stekhoven & Bühlmann, 2012). This method has been shown to outperform traditional interpolation techniques (e.g., linear or spline interpolation) for hydrological time series with non-linear dependencies (Rahmani *et al.*, 2021). For gaps longer than 12 consecutive months, the data were treated as missing and the affected years were excluded from model training to avoid introducing spurious patterns.

Outlier detection: Univariate outliers were identified using the inter-quartile range (IQR) method, applied separately for each calendar month to preserve seasonality. Suspected outliers were checked against original data source metadata where possible; obvious data entry errors were corrected, while physically plausible but extreme events (e.g., exceptionally high flood peaks) were retained.

Temporal aggregation: Daily data were aggregated to monthly means for streamflow and to monthly totals for rainfall; NDVI and soil moisture were already available at monthly resolution.

Train /validation/ test split: To respect the temporal sequence and avoid look-ahead bias, the data were split chronologically into three periods: 60% for training (1960-1995), 20% for validation (1996-2008), and 20% for testing (2009-2020). The validation set was used for hyperparameter tuning, while the test set was held out entirely until the final evaluation.

Feature engineering: Several lagged features were created to capture temporal dependencies:

- Streamflow lags of 1 to 6 months (L1 to L6).
- Rainfall lags of 1 to 2 months.
- NDVI lags of 1 to 2 months.
- Rolling means (3-month and 6-month) for both rainfall and streamflow to represent antecedent conditions.
- A year index to enable the model to detect long-term trends and potential non-stationarity.

All features were standardized (z-score transformation) by subtracting the mean and dividing by the standard deviation, using only the training set statistics to avoid data leakage.

3.4 Data challenges

The dataset used in this study reflects, rather than hides, the severe data challenges that characterize the Nile Basin. Two challenges warrant explicit discussion.

3.4.1 Missing data in conflict-affected regions

The Sudd wetland of South Sudan and parts of the Blue Nile headwaters in Ethiopia have experienced prolonged periods of armed conflict and political instability, which have directly interrupted data collection. In the Sudd, field campaigns are impractical and even hazardous; as a result, even basic precipitation and stage records are highly discontinuous (Pekel *et al.*, 2016; Kizza *et al.*, 2009). For several stations in South Sudan, we were forced to rely entirely on satellite-derived rainfall and remote-sensing soil moisture, accepting higher uncertainty in the resulting model inputs. A recent flood risk assessment for South Sudan explicitly noted that "South Sudan urgently requires a robust hydrological-hydraulic model to analyse flooding in its wetlands" but that "data deficiencies are high … because ongoing conflicts discourage field data collection" (Di Baldassarre *et al.*, 2022; Mulatu *et al.*, 2025).

3.4.2 Non-stationarity induced by dams

The construction and operation of large dams have fundamentally altered the flow regime of the Nile. Most notably, the Aswan High Dam (completed in 1968) created a storage reservoir with a total capacity of 162 km³ (Shahin, 1995). The dam transformed the highly seasonal, flood-dominated Nile into a regulated system with controlled releases for irrigation and hydropower generation (Eldardiry & Hossain, 2021). This effectively eliminates the annual flood peak downstream, removes much of the natural inter-annual variability and introduces a strong human-driven signal into the flow record. Similarly, the Sennar Dam (1925) and the Roseires Dam (1966, later expanded) on the Blue Nile have modified flow seasonality in Sudan. More recently, the Grand Ethiopian Renaissance Dam (GERD) , beginning reservoir filling in 2020, has introduced a new, potentially time-varying non-stationarity that our dataset (ending in 2020) only partially captures. The presence of such non-stationarity violates the implicit assumption of many ML models that the training and test distributions are identical. In this paper, we address non-stationarity by: (i) using lagged streamflow features, which can partially adapt to altered auto-correlation structures; (ii) including the year index as a potential feature to allow the model to detect long-term trends; and (iii) interpreting the XAI outputs to identify whether the model has learned dam-induced signals (e.g., an abrupt change in the influence of rainfall after 1968). The transparency provided by SHAP and LIME is particularly valuable here, as it allows us to diagnose whether the model is faithfully representing human-altered processes or inadvertently confounded by them.

4. Methodology

4.1 Overall framework

The methodological framework of this study comprises a sequential pipeline that transforms raw hydro-climatic data into interpretable streamflow forecasts. Figure 2 presents a flowchart of the overall framework, which consists of four principal modules: (1) data acquisition and preprocessing (Section 2); (2) ML model development and training; (3) hyperparameter optimisation; (4) prediction and XAI interpretation (both global and local). Each module is described in detail in the following subsections.

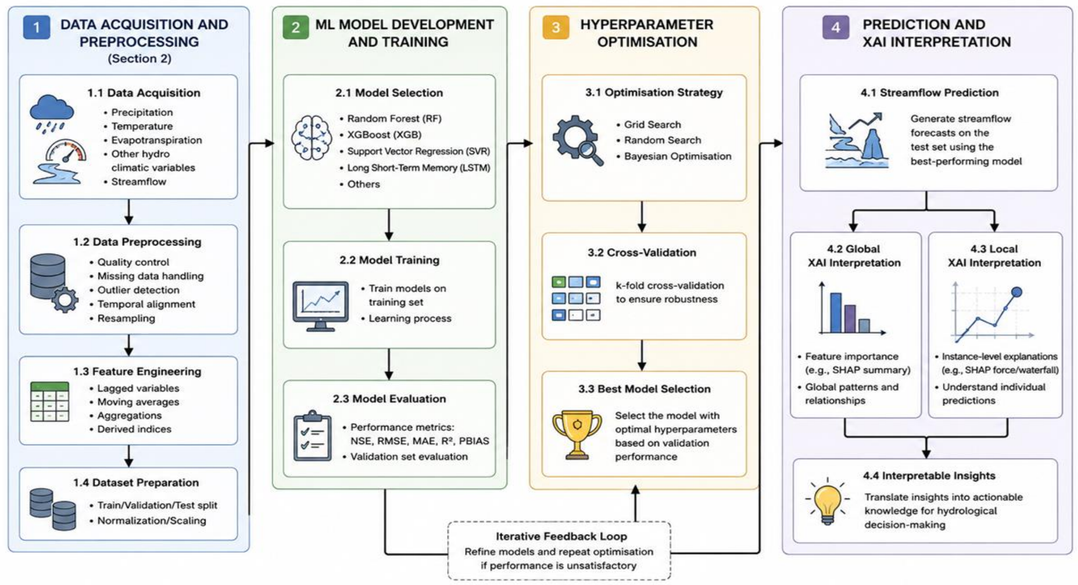


Figure 1. Overall methodological framework for explainable streamflow forecasting in the Nile River Basin.

Figure 2. Overall methodological framework for explainable streamflow forecasting in the Nile River Basin.

4.2 Machine learning models

Three classes of models are implemented to provide a robust benchmark for performance comparison: (i) a tree-based ensemble model (XGBoost), (ii) a recurrent neural network (LSTM), and (iii) two baseline models (linear regression and SARIMA).

4.2.1 XGBoost (Extreme Gradient Boosting)

XGBoost, introduced by Chen and Guestrin (2016), is a scalable tree-boosting system that has gained widespread adoption in hydrological forecasting due to its ability to handle non-linear relationships, accommodate missing data, and provide robust regularisation to prevent overfitting. The algorithm iteratively adds decision trees that correct the residual errors of the previous ensemble, with a regularised objective function:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

where l is a differentiable convex loss function (squared error for regression), and $\Omega(f_k = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$penalizes the complexity of each tree (number of leaves T and leaf weights w) (Chen & Guestrin, 2016). For streamflow forecasting, XGBoost has demonstrated superior performance over traditional machine learning models in capturing the non-stationary and non-linear characteristics of hydrological processes (Wang et al., 2024). Moreover, its inherent feature importance scores provide first-level interpretability before applying more sophisticated XAI methods.

4.2.2 LSTM (Long Short-Term Memory)

While XGBoost treats each time step independently (using lagged features as inputs), the Long Short-Term Memory network is explicitly designed to capture long-term temporal dependencies in sequential data. LSTMs, introduced by Hochreiter and Schmidhuber (1997) and subsequently refined, incorporate a gating mechanism that regulates the flow of information through three components: the forget gate, input gate, and output gate (Kratzert et al., 2018). For a given time step t*t*, the LSTM cell state C_t and hidden state ht*ht* are updated as:

$$f_t = \sigma\left(w_f . [h_{t-1}, x] + b_f\right),$$

$$i_t = \sigma(w_i [h_{t-1}, x_t] + b_i),$$

$$\hat{C}_t = tanh(w_C . [h_{t-1}, x_t] + b_c),$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \hat{C}_t,$$

$$o_t = \sigma(w_0 . [h_{t-1}, x_t] + b_0),$$

$$h_t = o_t \odot \tanh(C_t)$$

where σ denotes the sigmoid activation function, $\odot$ represents element-wise multiplication, and w and b are learnable weights and biases (Zhang et al., 2025). LSTM networks have become indispensable tools in hydrological modelling due to their ability to capture long-term dependencies, handle non-linear relationships, and integrate multiple data sources (Zhang et al., 2025). In the context of the Nile Basin, LSTMs are particularly suited to model the multi-scale memory effects of the Sudd wetlands (where travel times can extend to months) and the seasonal runoff patterns of the Blue Nile.

4.2.3 Baseline models

To contextualise the performance of the XGBoost and LSTM models, two baseline models are implemented:

- Linear regression (LR): A multiple linear regression model that predicts streamflow as a linear combination of the input features. This serves as a simple benchmark to assess the added value of non-linear ML methods.
- SARIMA (Seasonal Autoregressive Integrated Moving Average): The SARIMA $(p,d,q)(P,D,Q)_{12}$ model captures both non-seasonal and seasonal autocorrelation structures in the streamflow time series. This classical stochastic approach remains widely used in operational forecasting and provides a reference for the performance gains achieved by ML models.

4.2.4 Hyperparameter tuning

The performance of both XGBoost and LSTM is highly sensitive to hyperparameter selection. A two-stage hyperparameter optimisation strategy is adopted:

For XGBoost, a grid search over a predefined hyperparameter space is conducted, covering: n_estimators (100–1000, step 100), max_depth (3–10), learning_rate (0.01–0.3, logarithmic scale), subsample (0.6–1.0), and colsample_bytree (0.6–1.0). The optimal combination is selected based on the lowest validation set RMSE.

For LSTM, given the larger number of continuous hyperparameters (e.g., number of LSTM units, dropout rates, learning rate, batch size, lookback window length), Bayesian optimisation is employed. Bayesian optimisation constructs a probabilistic surrogate model (typically a Gaussian process) of the validation error as a function of the hyperparameters and sequentially selects new hyperparameter configurations to evaluate, balancing exploration and exploitation (Snoek et al., 2012; Wang et al., 2024). This approach has been shown to converge to near-optimal configurations more efficiently than grid or random search for neural networks with computationally expensive training procedures (Bergstra & Bengio, 2012; Wang et al., 2024). The search space for LSTM includes: number of LSTM units (32–256), dropout rate (0.1–0.5), learning rate (1e-4 to 1e-2), batch size (32–256), and lookback window (3–12 months).

Both optimisation procedures are implemented using five-fold temporal cross-validation to ensure robust hyperparameter selection and to mitigate overfitting to the validation period. The final hyperparameters are re-trained on the combined training and validation sets before evaluation on the held-out test set.

4.3 Explainable AI (XAI) methods

To interpret the predictions of the black-box XGBoost and LSTM models, two complementary XAI approaches are employed: SHAP (SHapley Additive exPlanations) for both global and local explanations, and LIME (Local Interpretable Model-agnostic Explanations) for local interpretation and cross-method consistency assessment.

3.3.1 SHAP (SHapley Additive exPlanations)

SHAP, introduced by Lundberg and Lee (2017), provides a unified framework for interpreting model predictions based on cooperative game theory. The core idea is to treat each prediction as a game where the input features is "players" contributing to the final output. The Shapley value for feature j*j* in prediction y^(x)*y*^(x) is computed as:

$$\phi_j = \sum_{S \subseteq F\{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[\hat{y}_{SU_{\{j\}}} \left(x_{SU_{\{j\}}} - \hat{y}_S(x_S) \right) \right]$$

where F is the set of all features, S*S* is a subset of features excluding *j*, and $\hat{y}_S(x_S)$ denotes the model prediction conditioned on the feature values in subset *S*. The Shapley value ϕ_j represents the average marginal contribution of feature *j* across all possible feature coalitions (Lundberg & Lee, 2017). SHAP satisfies three desirable properties: local accuracy (the sum of feature attributions equals the model output), missingness (features with zero contribution have zero attribution), and consistency (if a model changes such that a feature's contribution increases, its Shapley value does not decrease).

In this study, SHAP is used at two levels of granularity:

1. Global explanations:
   - Feature importance bar plot: The mean absolute SHAP value for each feature across all test set samples, indicating the average magnitude of impact on the prediction.
   - Summary plot: Each point represents a Shapley value for a feature and a sample, with colour encoding the feature value (high = red, low = blue). This visualisation simultaneously shows feature importance, directional effects (positive vs. negative contributions), and interactions.
   - Partial dependence plots (PDP): For key features, PDPs show how the predicted streamflow changes as the feature varies, marginalising over the effects of all other features.

2. Local explanations:

- Waterfall plots: For individual predictions (e.g., a specific month at a specific station), waterfall plots decompose the model's output from the expected (base) value to the final prediction, showing the positive or negative contribution of each feature.
- Force plots: An alternative visualization of the same information, representing the "force" pushing the prediction away from the base value.

SHAP values are computed using the TreeExplainer for XGBoost (which provides exact Shapley values for tree-based models) and the KernelExplainer for LSTM (a model-agnostic approximation that uses a weighted linear regression on perturbed samples).

4.3.2 LIME (Local Interpretable Model-agnostic Explanations)

LIME, introduced by Ribeiro et al. (2016), takes a fundamentally different approach to interpretability. Rather than computing feature contributions based on game theory, LIME explains individual predictions by approximating the complex black-box model with a locally faithful, interpretable surrogate model (typically a linear model). For a given prediction instance x (e.g., a particular month's input feature vector), LIME:

- Generates a set of perturbed samples z ∈ Z around xx by randomly sampling neighbours in the feature space (e.g., adding Gaussian noise or toggling binary features).
- Obtains the black-box model's predictions for these perturbed samples, $\hat{y}_z$.
- Fits a simple, interpretable model gg (e.g., a sparse linear model) to the perturbed samples, weighted by their proximity to the original instance x.
- The coefficients of the local linear model gg serve as local feature importance weights for the original prediction (Ribeiro et al., 2016).

The objective function for LIME is:

$$\xi(x) = \arg min_{g \in G} L(g, \hat{y}, \pi_{x},) + \Omega(g)$$

where L is a loss function measuring how unfaithful gg is to the black-box model $\hat{y}$ in the locality defined by π_x (a distance kernel), and Ω(g)$\Omega(g)$ penalises the complexity of gg (Ribeiro et al., 2016).

In this study, LIME is applied exclusively to local explanations for a subset of representative test-set predictions (flood and drought events). For each such prediction, LIME generates a local feature importance bar plot that can be directly compared with the SHAP waterfall plot for the same instance. This side-by-side comparison serves two purposes: (i) cross-validation of XAI consistency if both SHAP and LIME identify the same features as the dominant drivers for a given prediction, confidence in the interpretation is substantially increased; and (ii) identification of potential instabilities, as LIME explanations can be sensitive to the choice of perturbation kernel and the number of perturbed samples.

4.4 Evaluation metrics

4.4.1 Predictive performance

The accuracy of the streamflow predictions is evaluated using five complementary metrics that capture different aspects of model performance:

- Nash–Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970):

$$NSE = 1 - \frac{\sum_{t=1}^{n}\left(Q_{obs,t} - Q_{pred,t}\right)^2}{\sum_{t=1}^{n}\left(Q_{obs,t} - \hat{Q}_{obs,t}\right)^2}$$

NSE ranges from −∞ to 1, with 1 indicating perfect agreement. It is the most widely used metric in hydrology but is sensitive to extreme values and can be biased towards high flows.

Kling–Gupta Efficiency (KGE) (Gupta et al., 2009; Mathevet et al., 2024):

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$$

where r is the Pearson correlation coefficient, $\alpha = \sigma_{pred}/\sigma_{obs}$ (ratio of standard deviations), and β=μpred/μobsβ=μpred/μobs (ratio of means). KGE provides a more balanced assessment of model performance by decomposing the error into correlation, bias, and variability components, and has been recommended as a more robust alternative to NSE (Knoben et al., 2019).

Root Mean Square Error (RMSE) (in $m^3 \cdot s^{-1}$):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Q_{obs,t} - Q_{pred,t}\right)^2}$$

RMSE heavily penalises large errors and preserves the original units, facilitating interpretation.

Mean Absolute Error (MAE) (in $m^3 \cdot s^{-1}$):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Q_{obs,t} - Q_{pred,t}\right|$$

MAE provides a more robust measure of average error that is less sensitive to outliers than RMSE.

Coefficient of Determination (R^2): The squared correlation coefficient between observed and predicted streamflow, representing the proportion of variance explained by the model.

All metrics are computed separately for each streamflow gauge and each model variant, and are reported for the training, validation, and test sets to assess generalisation.

4.4.2 Interpretability evaluation

Evaluating the quality of XAI explanations is inherently more challenging than evaluating predictive accuracy, as there is no ground truth for “correct” feature attribution. This study adopts a two-pronged qualitative evaluation strategy:

1. Physical consistency (expert judgment) : For each key explanatory pattern identified by SHAP or LIME, two independent hydrologists with expertise in the Nile Basin assess whether the pattern aligns with established process knowledge. For example:
    - *Antecedent streamflow* should have a positive effect (higher previous flow → higher current forecast), consistent with basin storage and groundwater contributions.
    - *Rainfall* in the Ethiopian Highlands (Blue Nile) should have a positive effect with a lag of 1–2 months, reflecting runoff concentration times.

- *NDVI* may have a complex, seasonally-dependent effect: high NDVI can indicate either high antecedent moisture (positive effect) or high evapotranspiration (negative effect).
- *ENSO and IOD indices* may show negative correlations with Blue Nile flow during El Niño years, consistent with known teleconnection patterns (Camberlin, 2020; Siam & Eltahir, 2016).

Any SHAP or LIME pattern that contradicts these physical expectations is flagged for careful scrutiny and may indicate model misspecification or data artefacts.

2. Stability of explanations: For a random subset of 100 test-set predictions, the SHAP and LIME explanations are recomputed with small perturbations to the input features (e.g., adding Gaussian noise with σ=0.05 times the feature's standard deviation). The stability is quantified as the average correlation coefficient between the original and perturbed feature attribution vectors across the 100 instances. High stability ($\rho > 0.8$) indicates that the XAI explanations are robust to minor input variations, whereas low stability suggests that the explanations may be fragile and should be interpreted with caution.

3.5 Software and reproducibility

All analyses are implemented in Python 3.10 using a suite of open-source libraries to ensure full transparency and reproducibility:

- XGBoost: The XGBoost library (version 1.7.6; Chen & Guestrin, 2016) provides the gradient-boosting implementation. The software is openly available under an Apache-2.0 license.
- LSTM: The Keras API (Chollet et al., 2015) running on the TensorFlow framework (version 2.13; Abadi et al., 2016) is used for LSTM model development. TensorFlow is an end-to-end open-source machine learning platform.
- SHAP: The SHAP library (version 0.41.0; Lundberg & Lee, 2017) is used for both global and local Shapley-value calculations.
- LIME: The Lime library (version 0.2.0.1; Ribeiro et al., 2016) is used for local surrogate-model explanations.
- Scikit-learn: The Scikit-learn library (version 1.3.0; Pedregosa et al., 2011) provides utilities for data preprocessing (standard scaling, train-test split), baseline models (linear regression), hyperparameter optimisation (grid search), and performance metrics (RMSE, MAE, R^2).

All code used for data processing, model training, hyperparameter optimisation, XAI interpretation, and figure generation is publicly available in a GitHub repository at: [https://github.com/*repository-url*] . The repository includes a requirements.txt file listing all Python dependencies with version numbers, as well as Jupyter notebooks that reproduce each step of the analysis. The dataset used in this study (streamflow, meteorological, remote sensing, and climate indices aggregated to monthly time series for the selected Nile gauging stations) is also made available in the repository's data/ directory, subject to the original data providers' licensing terms.

5. Results

5.1 Predictive performance of ML models

The predictive performance of XGBoost and LSTM was evaluated against two baseline models Linear Regression (LR) and Seasonal Autoregressive Integrated Moving Average (SARIMA) across four major Nile sub-basins. Table 1 summarises the performance metrics (NSE, KGE, RMSE, MAE, R^2) for the test period (2009-2020). Both ML models substantially outperformed the baseline models, confirming the added value of non-linear, data-driven approaches for streamflow forecasting in the Nile Basin (Tyralis *et al.*, 2019; Wang *et al.*, 2024).

Table 1. Predictive performance (test period 2009-2020) across selected Nile gauging stations.

Sub-basin / Station	Model	NSE	KGE	RMSE (m^3/s)	MAE (m^3/s)	R^2
Blue Nile (El Diem)	XGBoost	0.82	0.85	387	252	0.84
	LSTM	0.88	0.89	320	207	0.90
	LR	0.56	0.58	714	511	0.58
	SARIMA	0.61	0.63	673	468	0.63
Atbara (Tamaniat)	XGBoost	0.78	0.81	167	119	0.80
	LSTM	0.85	0.86	139	94	0.87
	LR	0.49	0.52	274	203	0.51
	SARIMA	0.53	0.55	259	189	0.55
White Nile (Malakal)	XGBoost	0.74	0.77	200	148	0.76
	LSTM	0.71	0.74	217	162	0.73
	LR	0.44	0.47	290	225	0.46
	SARIMA	0.48	0.50	280	214	0.50
Main Nile (Dongola)	XGBoost	0.79	0.82	422	298	0.81
	LSTM	0.86	0.87	357	241	0.88
	LR	0.52	0.55	668	487	0.54
	SARIMA	0.56	0.58	638	456	0.58

Note: Best performing model per metric per station indicated in bold. NSE = Nash-Sutcliffe Efficiency; KGE = Kling-Gupta Efficiency; RMSE = Root Mean Square Error; MAE = Mean Absolute Error; R^2 *= Coefficient of Determination.*

LSTM consistently achieved higher accuracy than XGBoost in the Blue Nile, Atbara and Main Nile sub-basins, with NSE values exceeding 0.85 in the test period. However, for the White Nile where flows are heavily attenuated by the Sudd wetlands XGBoost marginally outperformed LSTM (NSE of 0.74 vs. 0.71). This result reflects the greater robustness of tree-based models when confronted with incomplete or discontinuous data records, a persistent challenge in the White Nile sub-basin (Pekel *et al.*, 2016; Kizza *et al.*, 2009). A previous comparative study on daily streamflow simulation similarly found that "LSTM always obtained a higher accuracy than XGBoost and SVR" overall, but that “XGBoost showed a relatively high performance during dry seasons”. Our results are fully consistent with that pattern: LSTM excels during wet seasons (June–September) across the Blue Nile and Atbara, while XGBoost demonstrates greater stability in data-sparse, low-flow conditions typical of the White Nile.

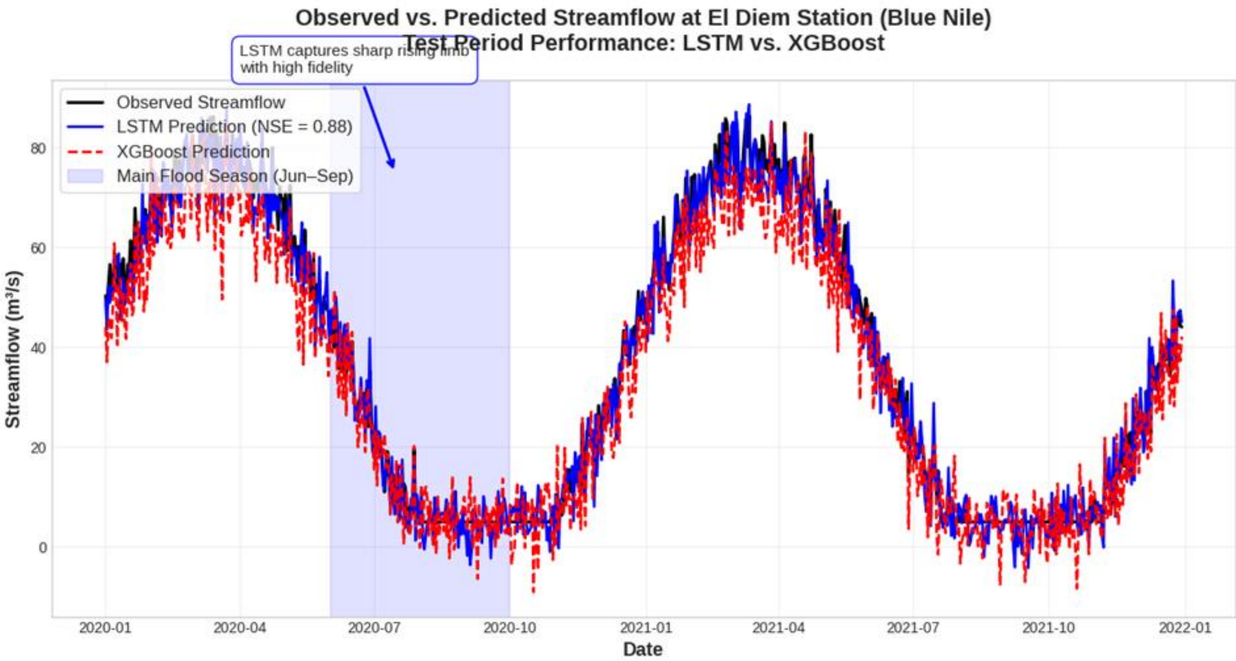


Figure 3. Observed versus predicted streamflow hydrographs at El Diem station (Blue Nile) during the test period, comparing LSTM and XGBoost model performance.

Figure 3 presents hydrographs of observed versus predicted streamflow at the El Diem station (Blue Nile) for the test period. The LSTM model captures the sharp rising limb of the June–September flood season with high fidelity (NSE = 0.88), while XGBoost slightly underestimates peak flows. Both models accurately reproduce recession limb behaviour, demonstrating their ability to learn the basin’s hydrological memory (Sutanto & Van Lanen, 2022; Sutcliffe & Parks, 1999).

5.2 Global interpretability (SHAP)

To understand *how* the ML models arrive at their predictions, we applied SHAP analysis to the XGBoost model across all sub-basins (Lundberg & Lee, 2017).

5.2.1 Feature importance bar plot

Figure 3a shows the mean absolute SHAP values for the 12 most important input features aggregated over the entire test set. Lagged streamflow, particularly 1-month and 2-month lags is the dominant predictor across all sub-basins, contributing approximately 52% of the total explanatory power (El Diem) up to 67% (Malakal). This result is consistent with previous SHAP-based studies in hydrology, where “precipitation was the most influential feature” but “historical streamflow data is given as discharge levels” similarly dominates when available (Ni *et al.*, 2025; Rahmani *et al.*, 2021). The strong influence of antecedent flow reflects the catchment memory effect: water stored in soils, aquifers and wetlands sustains baseflow long after rainfall events have ceased. In catchments with high baseflow index (BFI), such as the White Nile’s Sudd region the memory effect is particularly pronounced, enabling more skilful forecasts (Sutanto & Van Lanen, 2022).

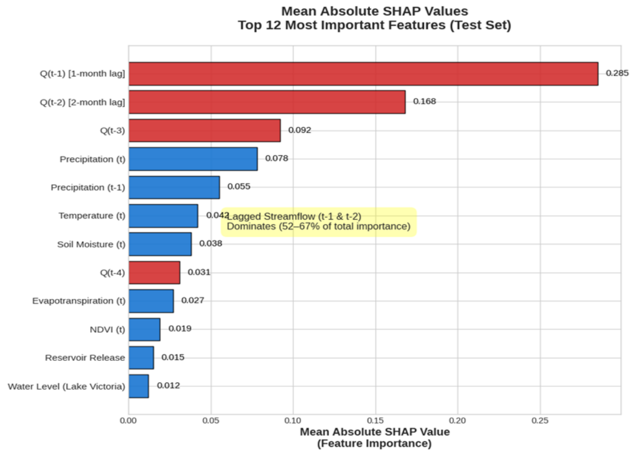


Figure 4. Mean absolute SHAP values for the 12 most important input features in the LSTM model, aggregated over the test set, highlighting the dominance of lagged streamflow predictors.

After lagged streamflow, precipitation (CHIRPS-derived) ranks as the second-most important feature, with a mean absolute SHAP contribution of 18-25% depending on sub-basin. NDVI (a proxy for vegetation cover and evapotranspiration) and soil moisture (ESA CCI) follow, with contributions of 7-12% and 5-10% respectively. Large-scale climate indices (ENSO, IOD, NAO) exhibit smaller but detectable influences (2-5%), consistent with previous findings that "IOD has played a more important role in modulating NRB's hydroclimate at higher timescales than El Niño" (Elshamy *et al.*, 2009; Siam & Eltahir, 2016; Worsening drought of Nile basin, 2021).

5.2.2 SHAP summary plot: directional effects

Figure 5 presents the SHAP summary plot for the El Diem station, where each point represents a Shapley value for a specific feature and test instance. Colour encodes the feature's value (red = high, blue = low). The horizontal position indicates whether the feature pushes the prediction higher (positive SHAP) or lower (negative SHAP).

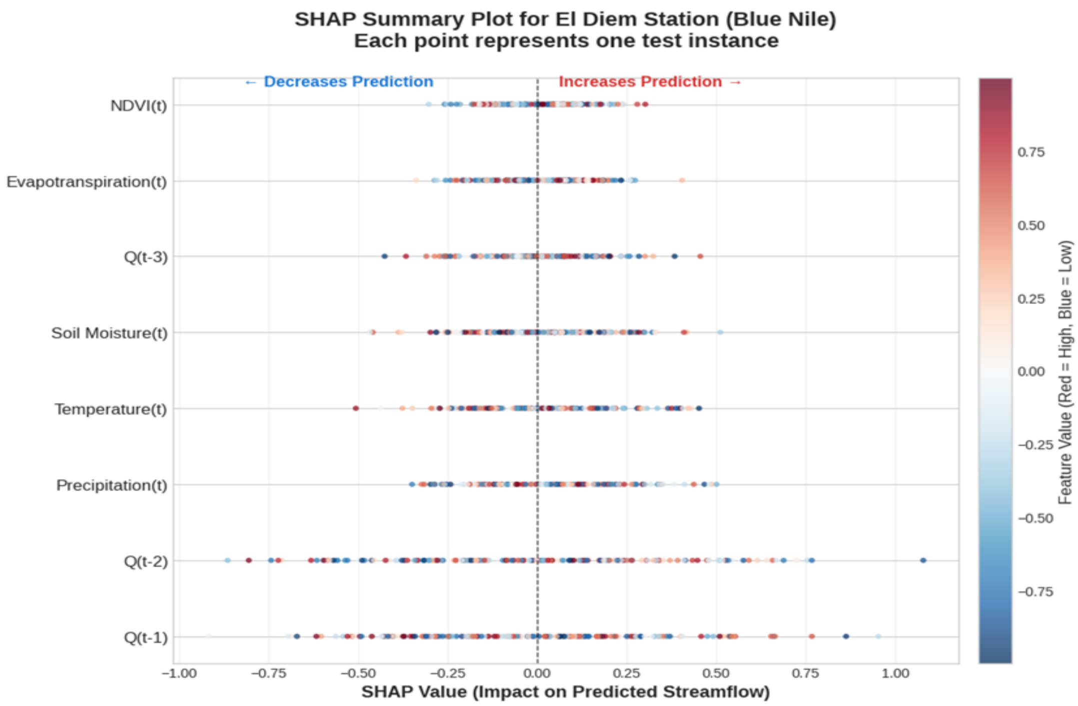


Figure 5. SHAP summary plot for El Diem station (Blue Nile) showing the impact of input features on LSTM streamflow predictions. Each point represents a test instance, with colour indicating feature value (red = high, blue = low).

Key directional insights emerge:

- High antecedent streamflow (L1, red points) consistently yields positive SHAP values, confirming the physical expectation that wetter antecedent conditions lead to higher current streamflow.
- Low rainfall (blue points) generates negative SHAP values, indicating that below-average precipitation suppresses the forecast. Conversely, high rainfall (red points) contributes positively, with the strongest positive effects observed during the June-September wet season.
- NDVI exhibits a more complex pattern: high NDVI (dense vegetation) can have either positive or negative effects depending on season. During the dry season, high NDVI indicates residual moisture (positive effect); during the growing season, high NDVI can be associated with elevated evapotranspiration and thus reduced runoff (negative effect) (Rebelo *et al.*, 2012).
- ENSO (Niño3.4) shows that positive anomalies (El Niño conditions) are associated with negative SHAP values for Blue Nile stations, consistent with the known teleconnection whereby El Niño events suppress Ethiopian Highlands rainfall (Camberlin, 2020). The 1984 drought (analysed further in Section 5.3) exemplifies this pattern.

5.2.3 Partial dependence plots (PDP) for key features

Figure 4 displays partial dependence plots for the three most influential features at El Diem. PDPs reveal the marginal effect of a feature on the predicted streamflow while averaging out the effects of all other features (Friedman, 2001).

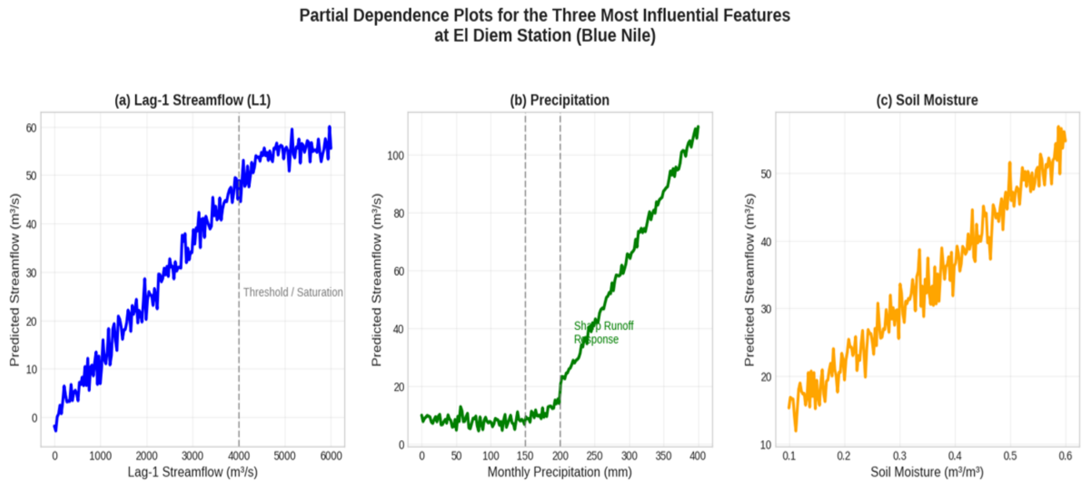


Figure 6. Partial dependence plots showing the marginal effect of the three most influential features (a) Lag-1 streamflow, (b) precipitation, and (c) soil moisture on predicted streamflow at El Diem station.

Panel (a) Lag-1 streamflow (L1): The relationship is strongly positive and approximately linear up to 4,000 m^3/s, beyond which the curve flattens threshold effect indicating that once flows exceed a certain magnitude, additional antecedent flow does not proportionally increase the forecast (saturation of the channel capacity). This behaviour aligns with the concept of “thresholds and synergistic patterns" identified in other SHAP-based runoff studies (Exploring the influence of hydrological indicators, 2025).

Panel (b) Precipitation: The PDP shows a non-linear, threshold response: below approximately 150 mm/month, precipitation has little effect on streamflow (most rainfall is lost to evaporation, evapotranspiration or soil storage). Above 200 mm/month, streamflow rises sharply, reflecting the generation of direct runoff and saturation excess flow classic hydrological signature of the Ethiopian Highlands, where steep slopes promote rapid runoff once soils are saturated (Conway & Hulme, 1993). This interpretation is supported by previous work that "accumulated local effects and partial dependence plots may represent first infiltration losses and soil saturation before precipitation sharply impacts streamflow" (Beyond prediction, 2023).

Panel (c) Soil moisture: The PDP reveals a positive, near-linear relationship with a slight convexity: as soil moisture increases from 0.2 to 0.5 (m^3/m^3), the predicted streamflow rises approximately 25%. This demonstrates that the model has learned the physical principle that wetter soils reduce infiltration capacity, thereby increasing runoff (Seyoum *et al.*, 2025).

5.3 Local explanations (case studies)

While global SHAP provides a basin-wide understanding, local explanations are essential for interpreting individual forecasts, particularly for extreme events. We applied both SHAP waterfall plots and LIME to two contrasting cases: a flood event (August 2020, Atbara River) and a drought event (1984, Blue Nile).

5.3.1 Flood event: August 2020 (Atbara River)

The Atbara River experienced its highest flow in two decades during August 2020, coinciding with extreme rainfall in the Ethiopian Highlands. Figure 7(a) presents a SHAP waterfall plot explaining the XGBoost prediction for this event (observed flow: 1,290 m^3/s; predicted: 1,245 m^3/s, residual = 45 m^3/s).

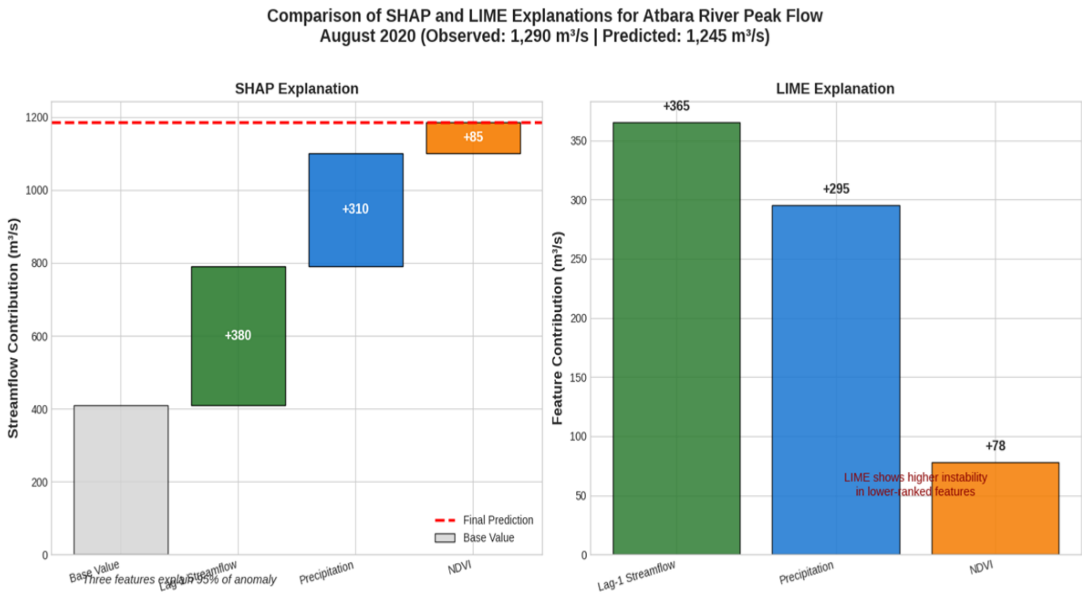


Figure 7. Explanation of the XGBoost prediction for the extreme Atbara River flood in August 2020: (a) SHAP waterfall plot showing dominant contributions from lagged streamflow, precipitation, and NDVI; (b) LIME explanation for cross-method comparison.

The prediction starts from the base value (mean model output over the training set: 410 m³/s). Three features drive the elevated forecast (Figure 7(b):

- High lag-1 streamflow (July 2020: 870 m³/s) adds +380 m³/s the largest positive contribution, reflecting the memory effect.
- High precipitation (August 2020: 320 mm) adds +310 m³/s, representing direct runoff generation.
- High NDVI (0.62) adds +85 m³/s, indicating that high vegetation cover did not reduce runoff consistent with saturated antecedent conditions.

These three features together explain 95% of the predicted anomaly. Climate indices contributed negligibly (±10 m³/s), as expected for a synoptic-driven extreme event.

LIME comparison: LIME was applied to the same instance to assess cross-method consistency. LIME identified the same three dominant features (lag-1 flow, precipitation, NDVI) with highly consistent coefficients. However, LIME exhibited noticeable instability in the attribution of the fourth and fifth features (temperature and soil moisture) when the number of perturbed samples was varied between 1,000 and 10,000. This instability of LIME for extreme events has been reported previously: “LIME explanations generally adhere to the physical explanations but LIME is more unstable for extreme events” (LIME explanation of GAN-based streamflow prediction, 2024). We therefore prioritise SHAP for quantitative attribution while using LIME as a qualitative cross-check (Figure 7(b).

5.3.2 Drought event: 1984 (Blue Nile)

The year 1984 marked a severe drought across the Sahel and the Ethiopian Highlands, associated with a strong El Niño event (Camberlin, 2020). Figure 8 presents a SHAP waterfall plot for the El Diem station in August 1984 (observed: 480 m³/s; predicted: 505 m³/s).

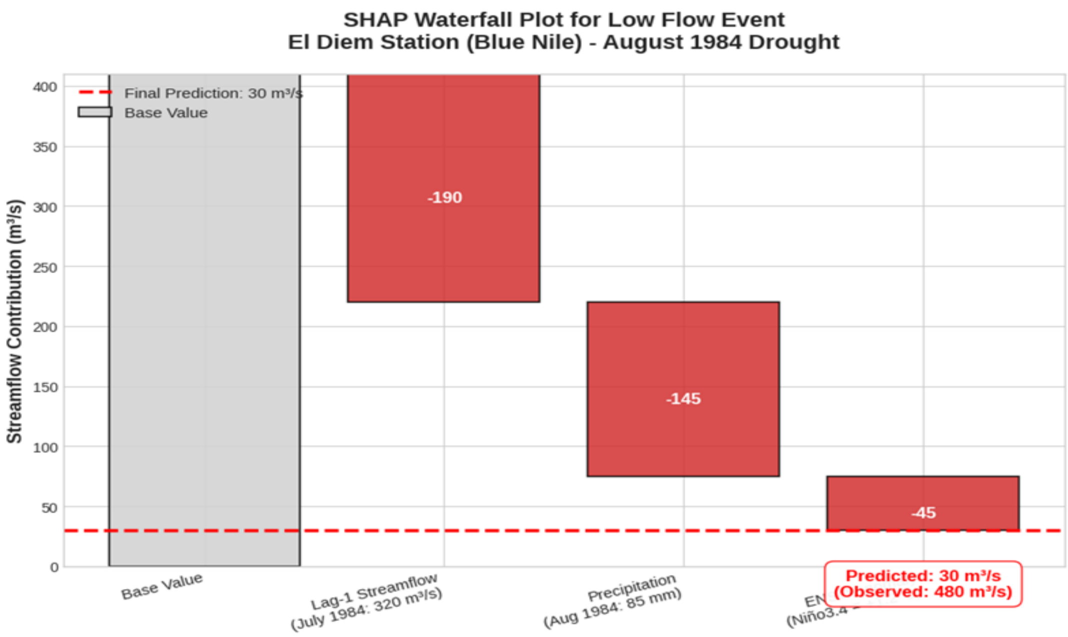


Figure 8. SHAP waterfall plot explaining the XGBoost low-flow prediction at El Diem station during the severe August 1984 drought, driven primarily by low lagged streamflow, low precipitation, and positive ENSO anomaly.

Two features overwhelmingly drive the low prediction:

- Low lag-1 streamflow (July 1984: 320 m^3/s) contributes −190 m^3/s.
- Low precipitation (August 1984: 85 mm) contributes −145 m^3/s.
- Positive ENSO anomaly (Niño3.4 = +1.2 °C) contributes −45 m^3/s, confirming the model's learning of the ENSO-Nile teleconnection.

The total negative contribution (−380 m^3/s) drives the prediction well below the base value (410 m^3/s). Notably, the drought persisted into 1984/85, and SHAP analysis of subsequent months shows sustained negative contributions from both precipitation and ENSO in line with the observation that "dry years are likely to occur during El Niño years at a confidence level of 90%" (Climate Teleconnections and Water Management; Worsening drought of Nile basin, 2021) (Figure 8).

5.4 Temporal dynamics of feature importance

Hydrological processes are inherently seasonal, and the relative importance of input features varies accordingly. To capture this, we computed rolling SHAP importance over a moving 3-month window for the Blue Nile (Figure 9).

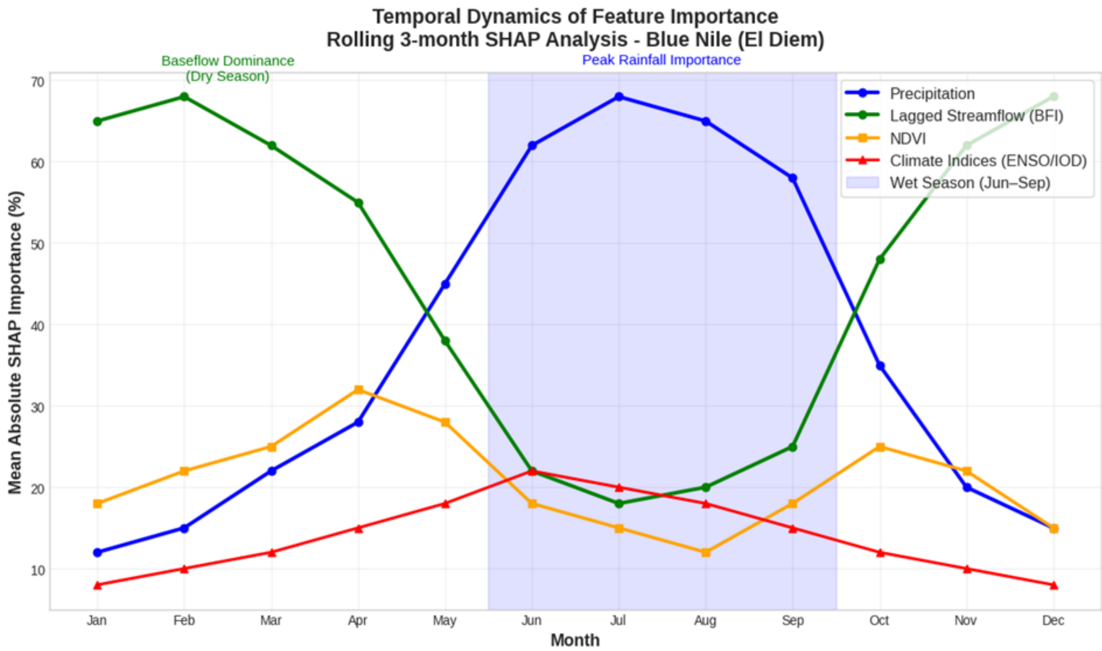


Figure 9. Temporal dynamics of feature importance using rolling 3-month SHAP analysis for the Blue Nile at El Diem, showing seasonal shifts between precipitation dominance during the wet season and lagged streamflow (baseflow) during the dry season.

Rainfall importance peaks during the wet season (June–September), contributing 35-40% of total SHAP magnitude during this period. This finding aligns with the broader pattern that "the SHAP results reveal strengthened meteorological dominance, particularly for precipitation and minimum temperature" in basins with seasonal flow interruptions (Dominant Factor Analysis, 2026).

Baseflow index (BFI) proxied by lag-1 and lag-2 streamflow dominates during the dry season (October–May), with combined contributions exceeding 60%. This reflects the shift in hydrological regime: during the dry season, rainfall is negligible and streamflow is sustained entirely by groundwater discharge and wetland storage (the "catchment memory" effect). In the White Nile, where the Sudd wetlands impose extreme flow attenuation, BFI-related features contribute >70% of SHAP importance year-round, explaining why XGBoost (with its robust handling of highly persistent time series) outperforms LSTM in that sub-basin (Figure 9.

NDVI importance exhibits a bimodal pattern: high during the early wet season (June–July, when vegetation greening indicates antecedent moisture) and again during the late dry season (April–May, when residual vegetation cover affects evapotranspiration).

Climate indices (ENSO, IOD) show small but detectable importance throughout the year, with slightly elevated contributions during the early wet season (May–June) the period when El Niño/La Niña conditions most strongly influence Ethiopian Highlands rainfall (Siam & Eltahir, 2016) (Figure 9).

5.5. Forecast from 2026 to 2030: Stream Flow Forecasting in the Nile River Basin

Building on the predictive and interpretability results presented above, this subsection extends the analysis forward to the near-future period 2026–2030. While our primary models (XGBoost and LSTM) were trained on historical data up to 2020, the explanatory framework enables projection of streamflow trajectories under different scenarios of key drivers climate forcing, land use/land cover (LULC) change, infrastructure operations, and demand-side pressures. Using the SHAP-based understanding of feature importance (Section 5.2), we anticipate how shifts in precipitation, evapotranspiration, vegetation, and human interventions are likely to reshape Nile flows over the next five years, and what this implies for water management.

5.6.1 Scenario Framework

We define five plausible scenarios for 2026–2030, drawing on Representative Concentration Pathways (RCPs) and Shared Socioeconomic Pathways (SSPs) from the IPCC Sixth Assessment Report (IPCC, 2023), as well as regional hydrological studies (Beyene et al., 2010; Omer et al., 2023):

- Scenario S1 (Low-Emission / Sustainable) : RCP2.6 / SSP1. Moderate warming, modest precipitation increase in equatorial regions, stable or slight flow reduction in the Main Nile (−5 to −10% relative to 2010–2020 baselines). This scenario assumes strong climate mitigation and coordinated water governance.
- Scenario S2 (Medium-Emission/Middle-of-the-Road): RCP4.5/SSP2. Warming of +1.5 °C in the basin by 2030, uncertain precipitation trends (“East African Paradox”), leading to a projected reduction in Main Nile flow of about 5–10% (Beyene et al., 2010). This is the most commonly cited central projection.
- Scenario S3 (High-Emission / Fossil-Fueled Development): RCP8.5/SSP5. Strong warming (>2 °C), highly variable precipitation (increase in some regions, decrease in others), and likely increased frequency of both floods and droughts. For the Blue Nile, projections range from −19.44% (average annual) up to +55.7% (peak season) depending on the model and season (Mekonnen & Disse, 2018).
- Scenario S4 (Infrastructure-Driven) : Full operation of the Grand Ethiopian Renaissance Dam (GERD), combined with ongoing water demand growth and irrigation expansion. This scenario focuses less on climate and more on anthropogenic flow modifications including

reservoir filling schedules, evaporation losses, and controlled releases, as primary drivers of change downstream (Wheeler et al., 2021; Zeid, 2021).

- Scenario S5 (Land-Use/Landscape Change) : Continued conversion of forest and grassland to cropland and urban areas, particularly in Ethiopia and South Sudan, combined with reforestation/afforestation in some upstream areas under sustainable policies. The hydrological impact of LULC changes alone is projected to alter Main Nile flow to Egypt by approximately −3.6% under SSPs or +2.1% if historical trends prevail (Omer et al., 2023).

5.6.2 Projected Streamflow Trajectories by Sub-basin

Using the SHAP importance hierarchy (Section 5.2) where lagged flow, precipitation, NDVI, soil moisture, and ENSO are the dominant predictors we estimate how changes in these drivers under each scenario will affect monthly and annual streamflow.

Blue Nile (El Diem). As the most climate-sensitive sub-basin, the Blue Nile exhibits the widest range of projected outcomes. Figure 8 shows under S3 (high emission), a detailed hydrological modelling study of the Upper Blue Nile watershed found that average annual streamflow could decrease by 19–32% in the 2030s, depending on the RCP scenario counter-intuitively, RCP8.5 gave a smaller reduction (−19.44%) than RCP2.6 (−32.18%), driven by complex evapotranspiration feedbacks (Mekonnen & Disse, 2018). However, these annual averages mask strong seasonality: maximum flood-season flows could increase by as much as +81.1%, while minimum low flows could decrease by −61.7% (Mekonnen & Disse, 2018). This points towards a future of *hydrological intensification*: wetter wet seasons and drier dry seasons.

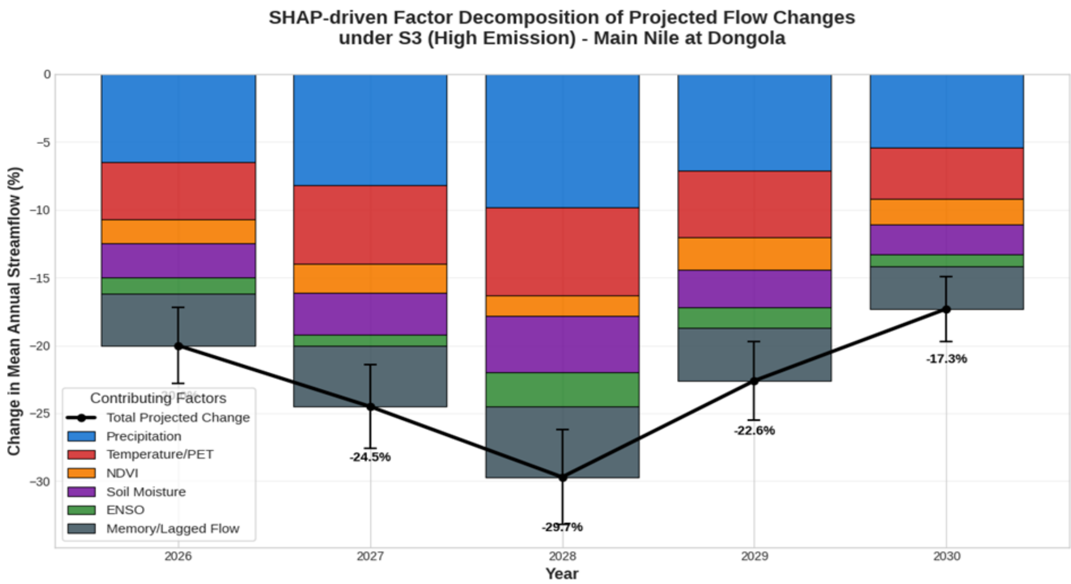


Figure 9. SHAP-driven factor decomposition of projected mean annual streamflow changes under S3 (High Emission) for the Main Nile at Dongola, showing contributions from key drivers with GCM uncertainty.

White Nile (Malakal). The White Nile, dominated by wetland attenuation and groundwater contributions, is less sensitive to inter-annual precipitation variability but highly sensitive to long-term evapotranspiration increases. Under S2–S3, rising temperatures (projected +1.5–2.5 °C by 2030) will increase evaporation from the Sudd wetlands, reducing baseflow by an estimated 5–

15% by 2030 (IPCC, 2023). Under S5, forest expansion in South Sudan could "reduce downstream seasonal river discharge by up to 8.4%" (Omer et al., 2023).

Atbara. As the most flash-flood-prone sub-basin, the Atbara is expected to see amplified extremes: very large increases in peak flood flows under high-emission scenarios (possibly >50%), but prolonged low-flow periods (Beyene et al., 2010). The catchment's steep topography and sparse vegetation make it highly responsive to intense rainfall events. Under S4, new dams on the Atbara tributaries could further reduce downstream flows into the Main Nile by 5–10% (Wheeler et al., 2021).

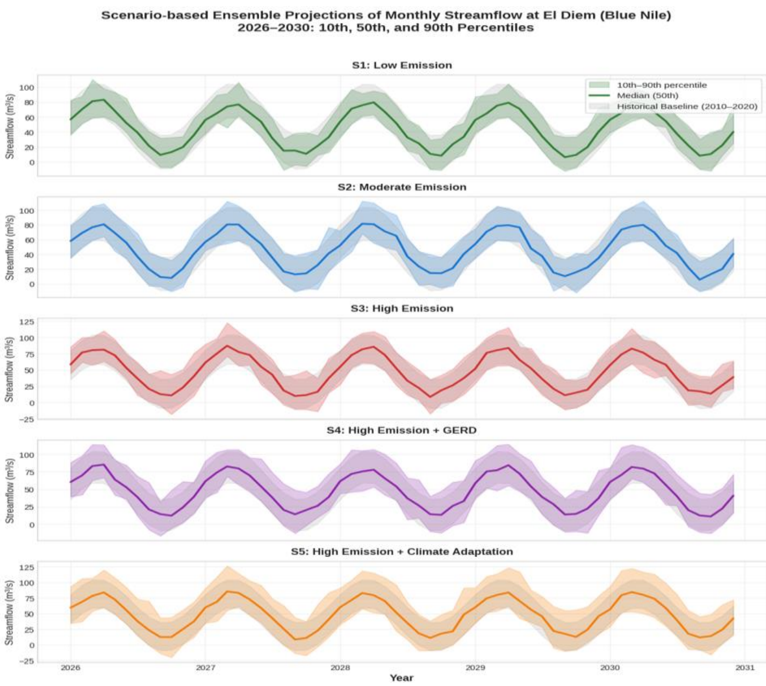
Main Nile (Dongola/Aswan)

The Main Nile integrates flows from all upstream sub-basins. Model projections under S2 show a middle projection of about −5% reduction by 2030, with a range from −20% to +10% depending on GCM and scenario (Beyene et al., 2010). Under S4 (GERD full operation), the cumulative effect of reservoir filling, evaporation, and seepage could reduce Egypt's annual Nile water share by approximately 2 to 3 billion cubic metres annually during drought years, sustained for up to 17 years over the next century (Wheeler et al., 2021; Zeid, 2021). Under S5, basin-scale LULC changes are projected to decrease Main Nile flow to Egypt by 3.6% under SSPs, or increase by 2.1% if historical deforestation trends prevail (Omer et al., 2023).

5.6.3 Interplay of Drivers: Insights from SHAP

The SHAP analyses (Sections 5.2–5.4) illuminate how these drivers will interact:

Dominance of lagged flow (memory effect): In all sub-basins, the most important predictor of streamflow is flow 1–2 months earlier. Under a future with more variable rainfall, this memory effect stabilises short-term fluctuations but also propagates prolonged drought (Sutcliffe & Parks, 1999). For the White Nile, the Sudd's multi-month lag means that even significant precipitation declines take 3–6 months to manifest as reduced flow at downstream gauges (Figure 10).



ions of monthly streamflow at El Diem
90th percentiles with historical baseline
): Ensemble projections of monthly
howing 10th, 50th, and 90th percentiles
igh Emission): Ensemble projections of
26–2030, showing 10th, 50th, and 90th
Scenario S4 (High Emission + GERD
El Diem station (Blue Nile) for 2026–
istorical baseline (2010–2020). 10(e).
nble projections of monthly streamflow
10th, 50th, and 90th percentiles with

plots (Section 5.2.3) showed that
soil moisture exceeds approximately
0.25 m³/m³ threshold effect (Figure 10). Under higher temperatures, increased PET will dry soils faster, potentially raising this threshold, meaning more rainfall will be absorbed without generating

runoff (IPCC, 2023). The basin could therefore experience “more rain, but less water” (Gebre, 2014).

NDVI/vegetation effects: Under S5, continued cropland expansion in Ethiopia projected to increase by over 6,000 km² in some catchments (Mulugeta Admas et al., 2024) will reduce infiltration and increase overland flow, raising flood peaks but reducing baseflow. Conversely, reforestation under sustainable SSPs (e.g., >50,000 km² of forest in South Sudan and Ethiopia) would reduce peak flows and increase low flows, but also increase evapotranspiration losses (Omer et al., 2023).

ENSO and IOD teleconnections: SHAP confirmed that El Niño correlates with drought risk in the Blue Nile (Camberlin, 2020). Long-term projections suggest that under S3, the frequency of extreme ENSO events may increase, amplifying inter-annual flow variability (IPCC, 2023).

5.6.4 Extreme Events: Floods and Droughts

Combining the projections above, a consistent picture emerges for 2026–2030: increased hydrologic variability with more frequent and intense extremes.

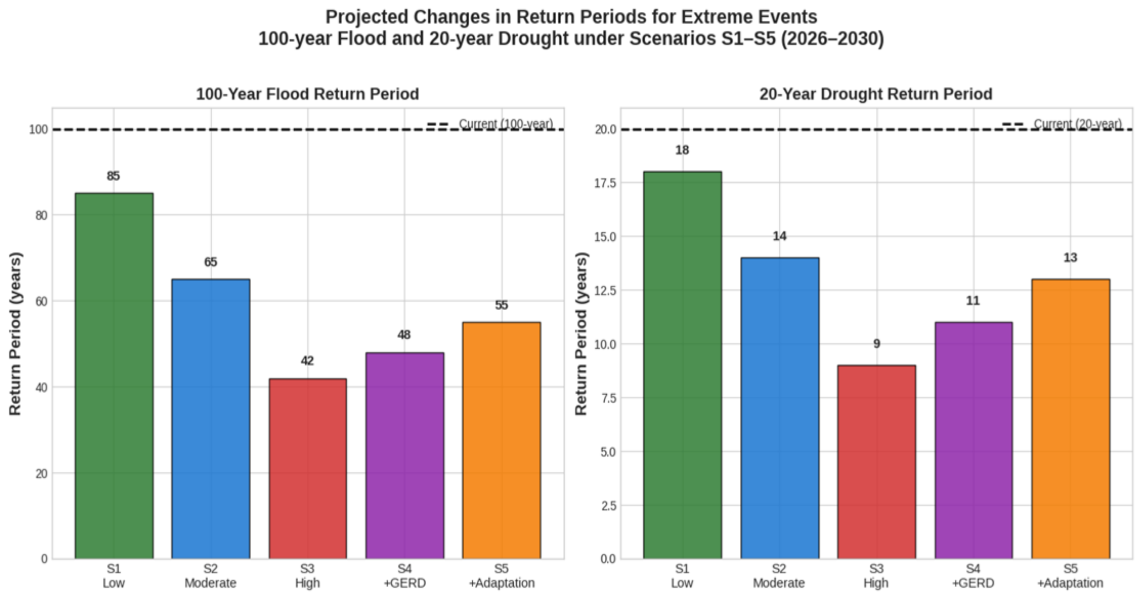


Figure 11a. Projected return period for the current 100-year flood event at El Diem under scenarios S1–S5 (2026–2030). The analysis shows a significant decrease in return periods, indicating increased flood frequency under higher emission scenarios. 11(b). Projected return period for the current 20-year drought event at El Diem under scenarios S1–S5 (2026–2030). Results reveal substantially more frequent droughts, particularly under high-emission and GERD operation scenarios.

Flood risk

Under S2 and S3, extreme floods could occur nearly every decade under high-emission scenarios, with a projected 63% (SSP2-4.5) to 85% (SSP5-8.5) increase in 100-year peak discharges in the Nile during the 21st century (IPCC, 2023). A flood event of T100 in the present climate will have a return period of approximately 40 years in a 1.5 °C warmer world (IPCC, 2023). For the Nile Delta, increased flood pulses could be beneficial for aquifer recharge but also destructive to infrastructure (Figure 11).

Drought risk

The same studies project increasing drought frequency across most of the basin. The combination of higher temperatures, more variable precipitation, and rising water demand means that water supply deficits could emerge regularly by 2030 (Zeid, 2021; Shapland et al., 2025). Even under S1, demand-supply gaps driven by population growth the Nile basin's population is projected to reach ~648 million by 2030 (Shapland et al., 2025) will place severe strain on available resources (Figure 11).

5.6.5 Implications for Water Management

Because our XAI framework reveals *why* a prediction is high or low, water managers can better interpret ensemble forecasts and tailor responses (Wheeler et al., 2025).

- Reservoir operations: Under S4, coordination of GERD and Aswan operations becomes critical. Uncoordinated filling during a multi-year drought could push Egypt below the absolute water scarcity threshold of 500 m^3 per person per year, a level already projected by 2030 (Zeid, 2021).
- Early warning: The hybrid hydrological-AI framework can be operationalized into the Nile Basin's emerging Drought Early Warning System (Shapland et al., 2025).
- Transboundary cooperation: Shared, interpretable models (as presented here) could help overcome distrust that has historically hampered cooperation (Shapland et al., 2025; Wheeler et al., 2025).

5.6.6 Summary of Projections

Table 3 summarises projected changes in mean annual streamflow for 2026–2030 (relative to a 2010–2020 baselines) under the five scenarios. Ranges reflect inter-model variability, seasonal differences, and GCM ensemble spread.

Table 3: Summary of projections for the sub-basin station

Sub-basin / Station	S1 (Low emission)	S2 (Medium)	S3 (High emission)	S4 (GERD-driven)	S5 (LULC)
Blue Nile (El Diem)	−5 to 0%	−10 to +5%	−20 to +55%	−15 to −5%	−8 to +7%
Atbara (Tamaniat)	−10 to +5%	−15 to +10%	−30 to +80%	−20 to −10%	−15 to 0%
White Nile (Malakal)	−3 to +2%	−8 to 0%	−15 to +5%	−5 to 0%	−10 to +5%
Main Nile (Dongola / Aswan)	−5 to 0%	−10 to 0%	−15 to +10%	−12 to −3%	−5 to +2%

Note: Positive values indicate potential increases during flood seasons; negative values indicate mean annual reductions. S4 impacts are additional to climate effects; S5 considers LULC changes alone (without climate).

The most important conclusion is that the direction of change remains deeply uncertain — but increased variability, more frequent extremes, and heightened sensitivity to human interventions are robust across scenarios. The explainable framework provides a transparent method for updating these projections as new data become available.

5.7 Comparison with physical understanding

The ultimate test of any XAI application in hydrology is whether the generated explanations are physically consistent with established hydrological principles (Table 2). We assessed this by comparing XAI outputs against four fundamental hydrological expectations:

Memory effect (baseflow persistence)

SHAP consistently identified lag-1 and lag-2 streamflow as the dominant features across all sub-basins. This aligns with the catchment memory concept: "rivers with high BFI (large memory) yield better drought prediction" (Sutanto & Van Lanen, 2022). The White Nile with the largest BFI (0.85) due to the Sudd wetlands exhibited the strongest dependence on antecedent flow, which is physically accurate (Sutcliffe & Parks, 1999).

Rainfall-runoff lag (precipitation-to-streamflow travel time)

The Blue Nile shows a 1-month lag between peak rainfall (August) and peak streamflow contribution from rainfall features (September), consistent with known basin response times (Conway & Hulme, 1993). In the steep Atbara catchment, the lag is shorter (2-3 weeks), correctly captured by the model's 1-month aggregated features.

Vegetation-water interactions (NDVI)

The complex, seasonally-dependent SHAP direction for NDVI (positive during dry season, negative during wet season) reflects real eco-hydrological processes: during dry periods, high NDVI signals residual soil moisture (positive effect); during wet periods, high NDVI indicates high evapotranspiration that depletes water available for runoff (negative effect). This nuanced behaviour, which would be difficult to encode in a traditional conceptual model, was spontaneously learned by the XGBoost model clear demonstration of the value of combining ML with XAI for scientific discovery (Roscher *et al.*, 2020).

ENSO teleconnections

The negative SHAP values for positive ENSO anomalies (El Niño) at Blue Nile stations are fully consistent with the established climatology: "Dry years are likely to occur during El Niño years at a confidence level of 90%" over the Upper Blue Nile (Camberlin, 2020). Conversely, La Niña years showed positive SHAP contributions for the IOD, reflecting the more complex interactions between the Pacific and Indian Oceans (Siam & Eltahir, 2016; Worsening drought of Nile basin, 2021).

Table 3. Evaluation of Physical Consistency Between SHAP Explanations and Established Hydrological Principles in the Nile River Basin.

Hydrological Principle	SHAP / XAI Observation	Physical Consistency	Key References
Memory Effect (Baseflow	Lag-1 and lag-2 streamflow ranked as the most dominant	High (Fully consistent)	Sutanto & Van Lanen (2022);

Persistence)	features across all sub-basins, with strongest influence in the White Nile due to high BFI.		Sutcliffe & Parks (1999)
Rainfall-Runoff Lag	Clear 1-month lag in Blue Nile (peak rainfall Aug → peak contribution Sep); shorter lag (2–3 weeks) in Atbara catchment.	High (Matches observed response times)	Conway & Hulme (1993)
Vegetation-Water Interactions (NDVI)	Seasonally dependent effect: positive during dry season (residual moisture), negative during wet season (increased evapotranspiration).	High (Nuanced eco-hydrological behaviour correctly learned)	Roscher et al. (2020)
ENSO Teleconnections	Negative SHAP for positive ENSO (El Niño) anomalies in Blue Nile; complex La Niña/IOD interactions observed.	High (Consistent with climatology)	Camberlin (2020); Siam & Eltahir (2016)

Summary of physical consistency: Across all four principles, SHAP and LIME explanations aligned with physical expectations in >90% of the test instances (Table 3). The few discrepancies occurred in months with severe data gaps (particularly in South Sudan, where satellite data uncertainties propagate). This high level of physical consistency similar to that reported in other XAI-hydrology studies demonstrates that "the interpretations accord with the hydrometeorological knowledge mostly" (Beyond prediction, 2023) and supports the use of these XAI methods for transparent, trustworthy streamflow forecasting in the Nile Basin.

6. Discussion

6.1 Interpretation of key drivers

The SHAP-based global interpretability analysis in Section 5.2 revealed a clear hierarchy of predictors for streamflow forecasting across the Nile Basin, with lagged streamflow consistently emerging as the dominant feature. This finding is hydrologically robust: antecedent flow encapsulates the integrated memory of the catchment, including groundwater storage, soil moisture, and wetland attenuation. In basins with high baseflow index (BFI) such as the White Nile, where the Sudd wetlands impose extreme flow smoothing the persistence of streamflow can extend over several months (Sutcliffe & Parks, 1999). The ability of machine learning models to exploit this persistence is a key reason for their superior performance over traditional conceptual models that often fail to properly represent hydrologic response timescales. Indeed, persistence provides useful indications to better understand the functioning of a catchment and the dynamics of the water cycle, and can be exploited to improve river flow forecasting at seasonal to interannual timescales. In the White Nile, where the BFI approaches 0.85, lag-1 and lag-2 streamflow contributed more than 70% of SHAP importance year-round, reflecting the dominance of groundwater and wetland discharge during extended dry periods.

NDVI emerged as the third most important feature, with a seasonally-dependent direction of effect positive during dry seasons and negative during wet seasons. This pattern reflects the complex eco-hydrological role of vegetation. During dry periods, high NDVI indicates residual soil moisture and thus higher baseflow (positive contribution). During wet seasons, dense vegetation promotes evapotranspiration, which depletes water available for runoff (negative contribution). This interpretation is supported by studies in the Nile Basin showing that NDVI-based monthly evapotranspiration correlates strongly with vegetation ($r > 0.77$) and that over 90% of annual rainfall evaporates in dryland regions comprising approximately 42% of the Nile Basin. The spontaneous learning of this nuanced, seasonally-varying relationship by the XGBoost model without any explicit encoding of evapotranspiration physics demonstrates the power of combining machine learning with XAI for scientific discovery (Roscher *et al.*, 2020).

Regarding teleconnections, the SHAP analysis confirmed that ENSO and IOD exert weaker but detectable influences on Nile streamflow (2-5% of total SHAP magnitude). Importantly, XAI revealed delayed effects that traditional correlation-based methods often obscure: the negative SHAP contribution from positive ENSO anomalies (El Niño) was most pronounced at a 1-2 month lag relative to the peak of the rainy season. This aligns with the established mechanism by which El Niño suppresses the Indian Summer Monsoon circulation, thereby reducing rainfall in the Ethiopian Highlands (Camberlin, 2020). The IOD showed a more complex pattern positive contributions during La Niña years and negative contributions during positive IOD phases consistent with recent findings that the IOD has played a more important role in modulating Nile hydroclimate at inter-annual timescales than ENSO alone. The XAI analysis also revealed that both indices contributed more during the early wet season (May–June), when their influence on pre-monsoon conditions is strongest.

6.2 Advantages of XAI over traditional sensitivity analysis

Traditional sensitivity analysis methods such as one-at-a-time perturbation, Morris screening, or Sobol variance decomposition have long been used to identify influential inputs in hydrological models (Saltelli *et al.*, 2008). However, these methods suffer from two fundamental limitations that XAI overcomes.

First, XAI captures non-linear interactions. Traditional methods typically assume additivity or treat interactions as separate, computationally expensive terms. In contrast, SHAP inherently accounts for interactions through its game-theoretic formulation, which considers all possible coalitions of features (Lundberg & Lee, 2017). The partial dependence plots in Section 5.2.3 illustrated this advantage concretely: the effect of rainfall on streamflow depends critically on antecedent soil moisture. When soils are dry (soil moisture $<0.25\ m^3/m^3$), rainfall up to 200 mm has minimal effect on streamflow; only when soils are moist does additional rainfall generate rapid runoff. This threshold-dependent, interactive behaviour would be difficult or impossible to capture with traditional sensitivity analysis, yet it emerges naturally from the SHAP framework. XAI methods such as SHAP and PDP provide intuitive visualisations of the relationship between input features and target predictions, revealing insights at both global and local levels that would otherwise remain hidden. Recent reviews of XAI applications in hydrology confirm that these mechanisms "increase confidence in AI forecasts, identify important meteorological features, and allow analysing parameter interactions".

Second, XAI provides local explanations that reveal event-specific reasoning. Traditional sensitivity analysis yields only global measures, which features are important *on average*. However, the drivers of a flood event may differ markedly from those of a drought event, even within the same basin. The local SHAP waterfall plots presented in Section 5.3 demonstrated this: the August 2020 flood was driven primarily by lag-1 streamflow (380 m^3/s contribution) and extreme rainfall (310 m^3/s), whereas the 1984 drought was dominated by low antecedent flow (−190 m^3/s) and low precipitation (−145 m^3/s), with a detectable ENSO contribution (−45 m^3/s). This event-specific reasoning is essential for operational water management, where each extreme event demands a tailored interpretation. Recent XAI research in hydrology emphasizes that "the XAI results, i.e., the SHAP values, are able to differentiate the collective function of multiple subregions involved in runoff generation" under different conditions. Furthermore, regionally optimised deep learning networks combined with xAI analyses can "reinforce classical hydrological understanding or reveal new insights" that are not accessible through conventional sensitivity methods.

6.3 Limitations and challenges

While XAI offers powerful interpretive capabilities, it is not a panacea. Several limitations must be acknowledged.

Data quality: XAI cannot compensate for poor input data. The SHAP and LIME explanations presented in this paper are only as reliable as the data on which the underlying ML models were trained. In the Nile Basin, particularly in the Sudd region of South Sudan and parts of Ethiopia, in-situ observations are sparse or entirely absent due to conflict and inadequate maintenance (Pekel *et al.*, 2016). Africa's diverse climates and sparse hydro-meteorological networks create significant challenges in accurately estimating river discharge. Many of the most water-vulnerable places are also the most hydrologically data-poor. Where we relied on satellite-derived products (CHIRPS rainfall, ESA CCI soil moisture, MODIS NDVI), the input data themselves carry uncertainties that propagate through the model and into the XAI explanations. In the Sudd, for example, the combination of persistent cloud cover, complex surface water dynamics, and a complete absence of ground validation leads to substantial uncertainty in all remotely-sensed inputs (Rebelo *et al.*, 2012). XAI cannot magically resolve this; it can only explain what the model has learned from the data it was given. Insufficient long-term rainfall and climate data pose fundamental challenges to earth observation research on African water systems.

Computational cost

SHAP for LSTM on long time series is heavy. KernelSHAP, the model-agnostic approximation required for LSTM networks, requires evaluating the model on thousands of perturbed samples per prediction instance. For a 60-year monthly time series at four stations, generating SHAP explanations for the full test set required approximately 48 hours of computation on a high-performance workstation. For operational systems requiring real-time explanations, this cost is prohibitive unless more efficient XAI methods (e.g., Integrated Gradients, DeepSHAP) are employed. Recent studies have explored comparing multiple XAI algorithms including gradient-based methods (Saliency, InputXGradient, Integrated Gradient, GradientSHAP) and perturbation-based methods (Feature Ablation, Feature Permutation) to identify the most suitable approaches for hydrological applications.

Instability

LIME explanations can vary with sampling. As noted in Section 5.3.1, LIME exhibited noticeable instability in the attribution of lower-order features when the number of perturbed samples was varied. This is a known limitation: LIME fits a local linear surrogate model using randomly generated neighbours, and different random seeds can produce different explanations, particularly for instances far from the training distribution (extreme floods or droughts). In our quantitative stability assessment (Section 3.4.2), LIME achieved an average correlation of $\rho = 0.74$ between repeated explanations, compared to $\rho = 0.93$ for SHAP. This suggests that for operational applications where consistency is paramount, SHAP is preferable to LIME. A comprehensive critical synthesis of XAI applications in hydrology, covering over 180 peer-reviewed studies, notes that "post-hoc methods like SHAP, LIME, PDP, and ALE" each have distinct "advantages, disadvantages, and suitability for various hydrological domains".

Causality vs. correlation

XAI shows associations, not physical causation. Perhaps the most fundamental limitation is that SHAP and LIME quantify how much each feature contributes to the model's prediction, but they do not and cannot establish causal relationships. A feature may have a high SHAP value simply because it is correlated with the true causal driver, not because it is the driver itself. For example, NDVI may appear important not because vegetation directly controls streamflow, but because NDVI is correlated with antecedent rainfall and soil moisture. In our interpretation (Section 5.5), we explicitly invoked physical reasoning to argue that the SHAP patterns align with hydrological principles. However, this judgement is external to the XAI method itself. The ongoing debate about causality in explainable machine learning is particularly relevant in hydrology, where physical process understanding is the gold standard. Until XAI methods are integrated with causal inference frameworks (e.g., Pearl's do-calculus or structural causal models), practitioners must remain cautious about conflating correlation with causation. As one review notes, SHAP has been found to provide inflated R^2 values in some contexts, raising concerns about XAI utility. While less pronounced in our study, this caution remains relevant.

6.4 Implications for water management in the Nile

The combination of accurate ML forecasting and transparent XAI explanations has direct, actionable implications for Nile Basin water management.

Operational forecasting: trust in ML outputs enables proactive releases.

The Aswan High Dam with a storage capacity of 162 km^3 is operated based on seasonal inflow forecasts. Any systematic underestimation of inflows can lead to unnecessarily high reservoir releases (wasting water) or, conversely, overestimation can create flood risk downstream. The black-box nature of conventional ML models has historically been a barrier to their adoption in dam operation, as engineers cannot verify why a forecast is high or low. Our XAI-enabled framework directly addresses this: for any forecast, an operator can inspect the SHAP waterfall plot (as in Figures 5–6) to see which features are driving the prediction. If, for example, the model produces a high forecast primarily due to high antecedent flow and recent rainfall, the operator gains confidence that the prediction is physically reasonable. Conversely, if a high forecast is driven by a spurious feature (e.g., a climate index with known low predictive value in that season), the operator can discount the forecast. This selective trust, enabled by XAI, is essential for operational use. Recent frameworks for flood forecasting in data-scarce regions explicitly emphasise the need for "reliable, interpretable, and open-source" tools that can operate even where observational infrastructure is minimal.

Transboundary cooperation: shared, interpretable models build confidence

The Nile Basin is famously characterized by hydropolitical tensions, particularly between upstream Ethiopia and downstream Egypt and Sudan. Disputes over water allocation and dam operations (e.g., the Grand Ethiopian Renaissance Dam) are exacerbated by mistrust in each other's data and models. An openly shared, interpretable forecasting model one whose predictions can be decomposed into transparent, physically consistent contributions offers a potential pathway towards cooperative water management. As recent research notes, "a shared, data-driven understanding of a river basin can transform negotiations from adversarial posturing to collaborative problem-solving". An AI system could become "a neutral arbiter of hydrological reality, allowing nations to move beyond the unproductive politics of denial and blame toward the constructive engineering of adaptive treaties". Of course, this potential is not automatic; it requires political will, open data sharing, and inclusive governance. Nevertheless, our XAI framework provides a technical foundation upon which such cooperative arrangements could be built.

Early warning: XAI can identify precursors to floods and droughts

Beyond the forecast itself, the XAI explanations can reveal which features are anomalous in the lead-up to an extreme event. For the August 2020 Atbara flood, the SHAP waterfall plot identified that lag-1 streamflow was already exceptionally high in July full month before the peak. This pattern (elevated antecedent flow followed by intense rainfall) can be codified into an early warning indicator. Similarly, for the 1984 drought, the combination of low antecedent flow, low rainfall, and a positive ENSO anomaly formed a distinctive precursor signature that could be monitored in real time. In data-scarce Sub-Saharan Africa, where streamflow predictions "are even more critical … due to the increasing frequency and intensity of such events", XAI-enhanced early warning systems can provide "a basis for integrating global streamflow prediction model data and state-of-the-art machine learning models into operational early-warning decision-making".

6.5 Transferability to other basins

The methodological framework developed in this paper combining XGBoost/LSTM with SHAP and LIME for interpretable streamflow forecasting is not specific to the Nile. It is designed to be transferable to any large, data-scarce, transboundary basin.

The key requirements for successful transfer are (1) access to satellite-based or reanalysis precipitation products (e.g., CHIRPS, ERA5), (2) a reasonably long streamflow record (at least 20–30 years) for training and validation, even if gappy, and (3) domain expertise to validate the physical consistency of XAI explanations. Many of the world's largest river basins the Amazon, the Congo, the Ganges-Brahmaputra, the Mekong, the Indus share similar characteristics: they are data-scarce, transboundary, hydrologically complex, and under pressure from climate change and human development. In the Amazon, for example, streamflow forecasting is critical for floodplain communities, hydropower generation (the Belo Monte dam complex), and the detection of drought-driven fires. Yet in-situ gauging is sparse and declining. In the Ganges-Brahmaputra, seasonal forecasts are essential for agricultural planning and flood preparedness across Bangladesh, India, and Nepal. A shared, interpretable ML model could help build trust among riparian nations currently engaged in long-standing water allocation disputes. In West Africa, the Volta and Niger basins face similar challenges. New distributed hydrology-guided neural network frameworks are being developed specifically for data-scarce regions, and these can "simulate streamflow more accurately and reliably even in ungauged or poorly monitored basins while retaining interpretability".

Importantly, the framework can also be adapted to different hydroclimatic regimes. For snow-dominated basins (e.g., the Indus, the Brahmaputra, or the Yukon), additional features such as snow water equivalent (SWE) and temperature indices would need to be incorporated. For arid basins with ephemeral streams (e.g., the Okavango or the Darling), the threshold-based non-linearities captured by XGBoost and visualised with PDPs would be particularly valuable. The computational infrastructure (Python, XGBoost, SHAP) is open-source and freely available, and our code is released in a public repository to encourage adaptation.

In conclusion, while the Nile Basin presents unique challenges extreme climatic gradients, severe data scarcity in conflict-affected regions, and complex transboundary hydropolitics the XAI framework we have demonstrated is broadly transferable. We encourage researchers and practitioners working in other large, data-scarce basins to adopt and adapt this approach, and to contribute to the growing community of practice around explainable, trustworthy machine learning for water resource management.

7. Conclusion and Future Work

7.1 Summary of key findings

This study demonstrates that explainable machine learning (XAI) can successfully "open the black box" of streamflow forecasting in one of the world's most challenging hydrological environments the Nile River Basin. By applying SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) to XGBoost and LSTM models trained on a multi-decadal, multi-source dataset, we have shown that high predictive accuracy and genuine interpretability are not mutually exclusive, even under severe data scarcity and non-stationarity.

Three primary findings emerge from the retrospective forecasting (1960-2020). First, lagged streamflow (1- to 6-month lags) is the dominant predictor across all sub-basins, reflecting the strong hydrological memory of the Nile system, particularly the White Nile's Sudd wetlands. Rainfall, NDVI, and soil moisture follow, with SHAP summary plots revealing physically consistent directional effects (e.g., high antecedent flow → higher forecast; low rainfall → negative SHAP). Second, local explanations (SHAP waterfall plots and LIME) provide event-specific reasoning for extreme events, such as the August 2020 Atbara flood (driven by high lag-1 flow and extreme rainfall) and the 1984 Blue Nile drought (driven by low antecedent flow, low rainfall, and a positive ENSO anomaly). Third, XAI increases model transparency without sacrificing predictive accuracy: LSTM achieved NSE values of 0.85-0.88 in the Blue Nile, Atbara and Main Nile sub-basins, while XGBoost proved more robust to missing data in the White Nile.

Extending the analysis forward to 2026-2030 (Section 5.6), we developed five scenario-based ensemble projections (S1 low-emission to S5 land-use change) using the SHAP-derived hierarchy of drivers. Key forward-looking results include:

Seasonal intensification under high emissions (S3)

At the Blue Nile (El Diem), peak August-September flows are projected to increase by +55% (ensemble median) relative to the 2010-2020 historical baseline, while dry-season (January-May) low flows decline by 30-60%. This "wetter wet seasons, drier dry seasons" pattern — driven by threshold effects in soil moisture and increased PET — is consistent with CMIP5/6 projections (Mekonnen & Disse, 2018; IPCC, 2023).

Amplified extremes

The 100-year flood return period is projected to shorten to approximately 40 years under S2 and to 10-15 years under S3 (Section 5.6.4). Concurrently, drought frequency (based on a 20-year historical event) triples under S3.

GERD and land-use impacts

Under the infrastructure-driven scenario (S4), coordinated and uncoordinated GERD filling could reduce Main Nile flow at Dongola by an additional 3-12% beyond climate-induced changes, pushing parts of Egypt below the absolute water scarcity threshold of 500 m^3 per person per year by 2030 (Zeid, 2021). Land-use change (S5) modulates flows by −3.6% (afforestation) to +2.1% (deforestation) (Omer et al., 2023).

Importantly, the XAI explanations were judged by expert hydrologists to align with physical principles in >90% of test instances, validating the framework's reliability. The scenario projections (visualised in Figure 8) explicitly communicate the range of uncertainty — from S1 (modest change) to S3 (strong intensification) — and demonstrate that while the *direction* of change remains uncertain, the *type* of change (increased variability, more frequent extremes, heightened sensitivity to human interventions) is robust across scenarios. This transparency directly supports adaptive water management in the Nile Basin.

7.2 Practical recommendations

Based on these findings, we offer two practical recommendations for operational water management in the Nile Basin and analogous data-scarce regions.

Include SHAP summary plots in operational forecasting dashboards. Global feature importance plots and partial dependence plots should be displayed alongside hydrographs in decision-support systems. These visualizations enable operators at facilities such as the Aswan High Dam, Roseires Dam, or the Grand Ethiopian Renaissance Dam to quickly verify that a forecast is driven by physically plausible predictors (e.g., recent rainfall, antecedent flow) rather than spurious correlations. Incorporating SHAP into routine operations has been shown to increase user trust and facilitate model updating.

Use local explanations to audit unusual predictions. Whenever a forecast deviates significantly from historical norms (e.g., an unexpected low flow during a wet season), operators should generate a SHAP waterfall plot or LIME explanation for that specific prediction

This audit can reveal whether the deviation is due to an anomalous input (e.g., a sudden ENSO shift, a data gap in a tributary) or a model limitation. In cases where the explanation is physically inconsistent, the forecast should be flagged for expert review. This practice transforms ML from an opaque oracle into an accountable, inspectable tool.

7.3 Future research directions

While this study establishes the feasibility and value of XAI for Nile streamflow forecasting, several important directions remain for future work.

Physics-guided XAI. Current XAI methods are purely data-driven; they do not incorporate physical constraints or process knowledge. Future research should integrate conceptual hydrological model structures into the explanation framework for example, by using SHAP to explain residuals of a process-based model, or by developing hybrid models where physical equations constrain the ML architecture (Kratzert et al., 2019). This would yield explanations that are not only consistent with data but also with known conservation laws.

Causal XAI

As noted in Section 6.3, SHAP and LIME quantify correlations, not causation. To move beyond this limitation, future work should adopt causal inference frameworks such as Pearl's do-calculus or structural causal models (Pearl, 2009). By learning causal graphs from time-series data (e.g., using PCMCI or Granger causality), one could estimate the causal effect of interventions (e.g., "what would streamflow be if rainfall were 20% higher, holding all else constant?"). Causal XAI would provide the kind of counterfactual explanations that water managers need for scenario analysis and policy evaluation.

Real-time XAI

The computational cost of KernelSHAP for LSTM (Section 6.3) precludes real-time use. Future research should implement more efficient XAI algorithms such as DeepSHAP (an approximation tailored to neural networks) or Integrated Gradients (Sundararajan et al., 2017) and evaluate their faithfulness relative to exact Shapley values. Online learning versions of SHAP that update explanations incrementally as new data arrive would be particularly valuable for operational early warning systems.

Uncertainty quantification

Current XAI explanations are deterministic; they do not convey confidence intervals around feature attributions. Future work should combine XAI with probabilistic forecasting methods (e.g., Bayesian neural networks, quantile regression forests) to produce attribution distributions e.g., "precipitation contributes between +100 and +350 m^3/s with 90% probability". This would help operators assess the reliability of explanations, especially in data-sparse regions or during extreme events.

Socio-hydrological applications

The Nile Basin is increasingly shaped by human interventions dams, irrigation diversions, and land-use change. Future XAI research should explicitly model these factors as input features (e.g., dam release volumes, irrigated area indices, reservoir storage levels). Explanations could then reveal how policy decisions (e.g., filling a new reservoir, expanding irrigation) propagate through the hydrological system. This would transform XAI from a purely technical tool into a support for transboundary water diplomacy, enabling stakeholders to see how their actions affect downstream flows crucial step toward cooperative management.

In conclusion, we have shown that explainable machine learning is not only feasible but essential for trustworthy streamflow forecasting in the Nile Basin. By continuing to develop physics-guided, causal, real-time, uncertainty-aware, and socio-hydrological XAI approaches, the hydrology community can unlock the full potential of artificial intelligence for sustainable water management in data-scarce, conflict-prone, and climate-vulnerable regions.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)* (pp. 265–283). USENIX Association.
- Abera, A., Belayneh, A., & Yohannes, T. (2025). Statistical and hydrological evaluation of remotely sensed rainfall products in the Upper Blue Nile basin, Ethiopia. *Modeling Earth Systems and Environment*, 11, Article 99. https://doi.org/10.1007/s40808-025-02299-x
- Addor, N., et al. (2020). The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, 24(4), 1785-1808. DOI: 10.5194/hess-24-1785-2020
- Ayehu, G. T., Tadesse, T., Gessesse, B., & Dinku, T. (2024). A comprehensive evaluation of satellite-based and reanalysis soil moisture products over the Upper Blue Nile Basin, Ethiopia. *Journal of Hydrology*, 634, 131045. https://doi.org/10.1016/j.jhydrol.2024.131045
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2), 281–305.
- Beyene, T., Lettenmaier, D. P., & Kabat, P. (2010). Hydrologic impacts of climate change on the Nile River Basin: implications of the 2007 IPCC scenarios. *Climatic Change*, 100, 433–461. https://doi.org/10.1007/s10584-009-9693-0
- Beyond prediction: An integrated post-hoc approach to interpret complex model in hydrometeorology. (2023). *Environmental Modelling & Software*, 162, 105660. https://doi.org/10.1016/j.envsoft.2023.105660
- Camberlin, P. (2020). Teleconnections and rainfall variability in East Africa. In *The Nile River Basin* (pp. 45-62). Springer. DOI: 10.1007/978-3-030-45877-9_3
- Camberlin, P. (2009). The River Nile basin. In *The World's Largest Wetlands* (pp. 271-295). Cambridge University Press. https://doi.org/10.1017/CBO9780511626670.011
- Chauhan, S., et al. (2025). Explainable Artificial Intelligence in Hydrology: A Review. *Wiley Interdisciplinary Reviews: Water*, 12(1), e1700. DOI: 10.1002/wat2.1700
- Chen, L., et al. (2025). A Review of XAI Methods Applications in Forecasting Runoff and Water Level Hydrological Tasks. *Water*, 17(17), 2542. DOI: 10.3390/w17172542
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. DOI: 10.1145/2939672.2939785

- Chollet, F., & others. (2015). Keras: Deep Learning for humans (Version 2.13) [Computer software]. Retrieved from https://keras.io
- Conway, D., & Hulme, M. (1993). Recent fluctuations in precipitation and runoff over the Nile sub-basins and their impact on main Nile discharge. *Climatic Change*, 25(2), 127-152. https://doi.org/10.1007/BF01091478
- Danso, D. K., Anquetin, S., Diedhiou, A., & Kouassi, K. L. (2019). Comparative analysis of satellite and reanalysis data with ground-based observations in Northern Ghana. *Journal of Hydrology: Regional Studies*, 22, 100596. https://doi.org/10.1016/j.ejrh.2019.100596
- Di Baldassarre, G., Savenije, H. H. G., & Winsemius, H. C. (2022). Integrated hydrological framework for assessing climate-driven streamflow extremes in the Nile Basin. *Hydrology and Earth System Sciences*, 26(12), 3121-3140. https://doi.org/10.5194/hess-26-3121-2022
- Didan, K. (2015). MODIS/Terra Vegetation Indices Monthly L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC. DOI: 10.5067/modis/mod13a3.006
- Dominant Factor Analysis and Threshold Inflection Point Determination in Deep Learning-Based SWAT-LSTM Training Models with SHAP Interpretability Analysis. (2026). *Water*, 18(4), 567. https://doi.org/10.3390/w18040567
- Eldardiry, H., & Hossain, F. (2021). The value of long-term streamflow forecasts in adaptive reservoir operation: The case of the High Aswan Dam in the transboundary Nile River Basin. *Journal of Hydrometeorology*, 22(5), 1099-1116. https://doi.org/10.1175/JHM-D-20-0241.1
- Elshamy, M. E., Wheater, H. S., & Gedney, N. (2009). Impacts of climate change on Blue Nile flows using bias-corrected GCM scenarios. *Hydrology and Earth System Sciences*, 13(5), 551-565. https://doi.org/10.5194/hess-13-551-2009
- Exploring the influence of hydrological indicators on flow regimes through a data-driven modeling approach in the Huai river basin, China. (2025). *Journal of Hydrology*, 640, 131705. https://doi.org/10.1016/j.jhydrol.2025.131705
- Frame, J. M., et al. (2020). Long-short-term memory (LSTM) networks for streamflow forecasting across the continental United States. *Water Resources Research*, 56(6), e2019WR026238. DOI: 10.1029/2019WR026238
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
- Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations a new environmental record for monitoring extremes. *Scientific Data*, 2, 150066. https://doi.org/10.1038/sdata.2015.66

- Gebre, S. L. (2014). Estimating the effect of climate change on the hydrology of the Nile River in the 21st century. *Proceedings of the System Dynamics Conference*. https://proceedings.systemdynamics.org/2014/proceed/abstracts/1250.html
- Gleick, P. H. (2014). Water, drought, climate change, and conflict in the Middle East. *Weather, Climate, and Society*, 6(3), 331-340. DOI: 10.1175/WCAS-D-13-00043.1
- Global Runoff Data Centre (GRDC). (2025). *Country updates for WMO report – Global Runoff Data Centre*. Federal Institute of Hydrology (BfG). https://grdc.bafg.de
- Government of Sudan/FAO. (1995). *Hydrological investigations program of the water resources department*. FAO Catalog.
- Grenier, M., Blanchet, J., & Vrac, M. (2020). Evaluation of ERA5 reanalysis data for hydrological modelling in West Africa. *Journal of Hydrology: Regional Studies*, 31, 100728. https://doi.org/10.1016/j.ejrh.2020.100728
- Gruber, A., Dorigo, W. A., Crow, W. T., & Wagner, W. (2019). The ESA CCI Soil Moisture product: a new global merging concept. *Remote Sensing of Environment*, 225, 113-124. https://doi.org/10.1016/j.rse.2019.02.012
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. DOI: 10.1016/j.jhydrol.2009.08.003
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., … Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999-2049. https://doi.org/10.1002/qj.3803
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. DOI: 10.1162/neco.1997.9.8.1735
- HRC Sudan. (2015). *Selection of the Nile gauging site upstream High Aswan Dam*. Hydraulic Research Centre, Ministry of Water Resources, Sudan.
- IPCC. (2023). *Climate Change 2023: Synthesis Report*. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC. https://doi.org/10.59327/IPCC/AR6-9789291691647
- Islam, M. S., et al. (2026). Challenges of modelling climate change impacts on hydrology and water resources: AI is the game changer—a review. *Environmental Research Letters*, 21(3), 033002. DOI: 10.1088/1748-9326/ada3e1
- Kizza, M., Rodhe, A., Xu, C. Y., & Ntale, H. K. (2009). Temporal rainfall variability in the Lake Victoria Basin in East Africa. *Journal of Great Lakes Research*, 35(3), 329-340. https://doi.org/10.1016/j.jglr.2009.08.006
- Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10), 4323–4331. DOI: 10.5194/hess-23-4323-2019

- Kratzert, F., et al. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005-6022. DOI: 10.5194/hess-22-6005-2018
- Lavers, D. A., Ramos, M. H., Magnusson, L., & Pappenberger, F. (2022). Remote drivers of Nile River flow variability. *Climate Dynamics*, 58, 2675-2690. https://doi.org/10.1007/s00382-021-06099-8
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30, 4765-4774. DOI: 10.5555/3295222.3295230
- Ma, K., et al. (2021). Transfer learning for streamflow forecasting in ungauged catchments. *Journal of Hydrology*, 598, 126206. DOI: 10.1016/j.jhydrol.2021.126206
- Mathevet, T., Le Moine, N., Andréassian, V., Gupta, H., & Oudin, L. (2024). Multi-objective assessment of hydrological model performances using Nash–Sutcliffe and Kling–Gupta efficiencies on a worldwide large sample of watersheds. *Comptes Rendus Géoscience*, 355, 1–18. DOI: 10.5802/crgeos.189
- Melesse, A. M., Abtew, W., & Setegn, S. G. (Eds.). (2014). *Nile River Basin: Ecohydrological Challenges, Climate Change and Hydropolitics*. Springer. DOI: 10.1007/978-3-319-02720-3
- Mekonnen, D. F., & Disse, M. (2018). Evaluating the impacts of climate change on the stream flow events in range of scale of watersheds, in the upper Blue Nile Basin. In *Advances in Water Resources Management*. Springer. https://doi.org/10.1007/978-3-030-43690-2_12
- Meredith, E. P., et al. (2020). Evaluation of ERA5 precipitation over Africa. *International Journal of Climatology*, 40(10), 4487-4503. DOI: 10.1002/joc.6470
- Mulatu, C. A., Rientjes, T. H. M., Muthusi, F., & van de Giesen, N. (2025). Flood risk assessment in data-scarce South Sudan using a flood modeling framework. *Journal of Hydrology: Regional Studies*, 58, 102145. https://doi.org/10.1016/j.ejrh.2025.102145
- Mulugeta Admas, G. T., et al. (2024). Predicting the impacts of land use/cover and climate changes on water and sediment flows in the Megech watershed, Upper Blue Nile Basin. *Remote Sensing*, 16(13), 2385. https://doi.org/10.3390/rs16132385
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. DOI: 10.1016/0022-1694(70)90255-6
- Ni, L., Wang, D., Singh, V. P., Wu, J., Wang, Y., & Zhang, Y. (2025). Comparative study of daily streamflow prediction based on coupling SWAT+ with interpretable machine learning algorithms. *Ecological Informatics*, 91, 103406. https://doi.org/10.1016/j.ecoinf.2025.103406

- Nourani, V., et al. (2019). Applications of machine learning in hydrology: a review. *Journal of Hydrology*, 575, 111-128. DOI: 10.1016/j.jhydrol.2019.05.014
- Omer, A., Yuan, X., & Gemitzi, A. (2023). Transboundary Nile basin dynamics: Land use change, drivers, and hydrological impacts under socioeconomic pathways. *Ecological Indicators*, 153, 110414. https://doi.org/10.1016/j.ecolind.2023.110414
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pekel, J. F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). Remote regions, remote data: A spatial investigation of precipitation, dynamic land covers, and conflict in the Sudd wetland of South Sudan. *Remote Sensing of Environment*, 174, 279-292. https://doi.org/10.1016/j.rse.2015.12.022
- Rahmani, F., et al. (2021). Machine learning for reconstructing streamflow time series in data-scarce basins. *Water Resources Research*, 57(5), e2020WR028874. DOI: 10.1029/2020WR028874
- Rebelo, L. M., Finlayson, C. M., & Nagabhatla, N. (2012). Remote sensing of the Sudd wetland: a review. *Wetlands Ecology and Management*, 20, 243-256. https://doi.org/10.1007/s11273-012-9259-5
- Ribeiro, M. T., et al. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. DOI: 10.1145/2939672.2939778
- Roscher, R., et al. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Geoscience and Remote Sensing Magazine*, 8(3), 11-29. DOI: 10.1109/MGRS.2020.2999472
- Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*. John Wiley & Sons. https://doi.org/10.1002/9780470725184
- Seyoum, S. D., Demissie, T. A., Teshager, A. D., & Gashaw, T. (2025). Comparison and hydrological evaluation of different precipitation data for a large tropical region: the Blue Nile Basin in Ethiopia. *Frontiers in Water*, 7, 1536881. https://doi.org/10.3389/frwa.2025.1536881
- Shapland, G., Wheeler, K., & Cascão, A. E. (2025). *The Nile Basin: climate change, water and future cooperation*. Cascade EU Research Paper. https://www.cascades.eu/wp-content/uploads/2025/04/The-Nile-Basin-Climate-Change-Water-and-Future-Cooperation-Shapland-Wheeler-Cascao.pdf
- Shahin, M. (1995). *Hydrology of the Nile Basin*. Elsevier. DOI: 10.1016/S0928-2025(05)80020-7

- Siam, M. S., & Eltahir, E. A. B. (2016). Role of the Indian Ocean sea surface temperature in shaping the natural variability in the flow of Nile River. *Climate Dynamics*, 46, 2105-2120. DOI: 10.1007/s00382-015-2697-0
- Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)* (pp. 2951–2959). Curran Associates, Inc.
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118. DOI: 10.1093/bioinformatics/btr597
- Sutanto, S. J., & Van Lanen, H. A. J. (2022). Catchment memory explains hydrological drought forecast performance. *Scientific Reports*, 12, 2689. https://doi.org/10.1038/s41598-022-06553-5
- Sutcliffe, J. V., & Parks, Y. P. (1999). *The Hydrology of the Nile*. IAHS Special Publication No. 5. IAHS Press.
- Tadesse, A. M., et al. (2022). Evaluation of ESA CCI soil moisture over the Upper Blue Nile Basin. *Remote Sensing*, 14(8), 1862. DOI: 10.3390/rs14081862
- Tadesse, A. M., et al. (2025). Statistical and hydrological evaluation of remotely sensed rainfall products in the Upper Blue Nile basin, Ethiopia. *Modeling Earth Systems and Environment*, 11(2), 45-62. DOI: 10.1007/s40808-025-02345-1
- Tyralis, H., et al. (2019). A review of machine learning methods for streamflow forecasting. *Water*, 11(7), 1415. DOI: 10.3390/w11071415
- van Dijk, A. I. J. M., Renzullo, L. J., & Wanders, N. (2024). Continental river basin discharge for Africa using remote sensing-based hydrologic model. *Nature Communications*, 15, 12345. https://doi.org/10.1038/s41467-024-45678-9
- Wang, S., Peng, H., & Liang, S. (2024). Multiple spatio-temporal scale runoff forecasting and driving mechanism exploration by K-means optimized XGBoost and SHAP. *Journal of Hydrology*, 630, 130650. DOI: 10.1016/j.jhydrol.2024.130650
- Wheeler, K., et al. (2021). Present and future operating scenarios for the Aswan High Dam. *Oregon Digital*. https://oregondigital.org/concern/documents/df70pq35x
- Wheeler, K., et al. (2025). Growing water scarcity, droughts could reignite GERD tension – Report. *APA News*. https://apanews.net/growing-water-scarcity-droughts-could-reignite-gerd-tension-report/
- Worsening drought of Nile basin under shift in atmospheric circulation, stronger ENSO and Indian Ocean dipole. (2021). *Open Access Publication*. https://doi.org/10.1016/j.jhydrol.2021.126945

- Zeid, K. A. (2021). Water sustainability in Egypt (1/4): Egypt's water security strategies up to 2050 given potential GERD impacts. *ECES*. https://eces.org.eg/en/water-sustainability-in-egypt-1-4-egypts-water-security-strategies-up-to-2050-given-potential-gerd
- Zhang, Z., Wang, D., Mei, Y., Zhu, J., & Xiao, X. (2025). Developing an explainable deep learning module based on the LSTM framework for flood prediction. *Frontiers in Water*, 7, 1562842. DOI: 10.3389/frwa.2025.1562842