https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

From Forecast to Alert: Designing an AI-Driven Flood Early Warning System for the White Volta Basin Using Open Satellite Data

Joseph Junior Obeng[1 *] and **Hongyong YUAN**[1,]

[1] School of Safety Science, Tsinghua University, Beijing 100084, China

*Corresponding author

Contact: Joseph Obeng Junior (Yue-sf25@mail.tsinghua.edu.cn)

Abstract

Flood early warning in the White Volta Basin of northern Ghana is complicated by unmonitored dam releases from Burkina Faso's Bagre Reservoir, which existing globally calibrated systems do not account for. We present an end-to-end AI-driven flood early warning system built entirely from open satellite data. An ensemble of Random Forest, XGBoost, and LSTM models trained on GRDC discharge, CHIRPS rainfall, ERA5-Land reanalysis, and a novel JRC-derived Bagre storage proxy achieved Kling-Gupta Efficiency scores of 0.984, 0.974, and 0.957 at 1-, 3-, and 5-day lead times on an independent test period, exceeding the GloFAS v2.1 African median benchmark of approximately 0.35, though direct comparison against GloFAS v4 at Nawuni was not undertaken. A four-tier alert system calibrated to 30-year flood return periods achieved a cross-validated Red-tier probability of detection of 0.902 (false alarm ratio 0.134) at one-day lead, declining to 0.762 at five days; higher-tier skill rests on leave-one-year-out cross-validation rather than held-out evidence, as the test period contains no Orange or Red events. Sentinel-1 SAR mapping confirmed that threshold exceedances correspond to observed inundation extents of 50 to 149 km². The system integrates into Ghana's existing myDEWETRA-VOLTALARM platform without requiring new institutional infrastructure.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

Keywords: flood early warning systems; machine learning; White Volta Basin; Ghana; Sentinel-1; CHIRPS; Bagre Dam; alert thresholds; leave-one-year-out cross-validation

1. Introduction

Flooding kills more people in West Africa than any other natural hazard, and the trend is getting worse. Across the region, a statistically significant upward trajectory in flood frequency has been documented over recent decades, with event rates exceeding 0.50 additional floods per year in parts of the Volta Basin (Koubodana et al. 2025). Ghana sits near the centre of this pattern. The White Volta Basin in the country's north experiences annual inundation that regularly destroys crops, displaces communities, and kills livestock, yet the timing and severity of the worst events are often determined not by local rainfall but by operational decisions made across an international border, at Burkina Faso's Bagre Dam.

The Bagre Reservoir was constructed primarily for irrigation and hydroelectric generation, but its releases during high-storage periods can send flood pulses downstream into Ghana within two to three days. Communities in the Central Gonja and Savannah districts have developed a deep familiarity with this dynamic over generations, but formal early warning systems have yet to integrate it. The myDEWETRA-VOLTALARM platform, deployed in the White Volta Basin through a partnership between Ghanaian agencies and the Italian Civil Protection Department, issues five-day probabilistic flood forecasts using GloFAS global model output, a system not calibrated to local conditions and without explicit representation of Bagre reservoir state

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

(Katsekpor et al. 2025a). As a consequence, warnings can arrive too late, cover the wrong locations, or fail to differentiate between a locally driven rainfall event and a Bagre-release flood that demands a categorically different response.

The literature on machine learning for hydrological forecasting has expanded rapidly over the past decade. Long Short-Term Memory networks have demonstrated strong skill for streamflow prediction in both gauged and ungauged basins, in some settings outperforming process-based models when trained on sufficiently large datasets (Kratzert et al. 2018). Ensemble combinations of LSTM with gradient-boosted trees have further improved skill for extreme event detection, which is the operationally critical capability for a flood warning system. In the White Volta Basin itself, Katsekpor et al. (2025b) demonstrated that LSTM and Random Forest models trained on CHIRPS rainfall and ERA5 reanalysis data can predict discharge at 1-, 5-, and 10-day intervals with high skill, establishing the methodological precedent for the present study. At global scale, Nearing et al. (2024) showed that AI-based systems can match the accuracy of next-day GloFAS nowcasts at five-day lead time in ungauged watersheds, a finding that suggests locally calibrated models with access to gauge data should perform substantially better still, provided the evaluation is conducted on a genuinely independent period.

What has not been demonstrated for the White Volta Basin, or, to our knowledge, for any comparable West African river system, is a complete end-to-end early warning system design that moves from satellite data ingestion through AI forecasting, calibrated threshold-based alert generation, satellite-derived inundation validation, and a concrete operational architecture compatible with Ghana's institutional infrastructure. The gap between a well-performing discharge prediction model and an operational flood warning system is substantial. It involves translating continuous discharge forecasts into discrete, actionable alert tiers; calibrating those tiers to return

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

periods that are meaningful to emergency managers and communities; validating that threshold exceedances correspond to real inundation; and specifying how the alert output integrates with the existing institutional chain from national forecasting centre to at-risk household. Each of these steps requires deliberate design choices that go beyond model performance optimisation, and each is addressed in the present study.

This paper makes four specific contributions. First, it incorporates Bagre reservoir upstream storage as an explicit model input, a variable that influences downstream flood risk profoundly but has been absent from all previous formal modelling work on the White Volta. Second, it designs a four-tier alert classification system calibrated to 30-year flood frequency analysis at Nawuni, producing discharge thresholds grounded in the statistical properties of the observed record. Third, it validates alert thresholds against Sentinel-1 synthetic aperture radar inundation maps for three major flood events, providing spatial evidence that the threshold-to-inundation relationship is empirically defensible. Fourth, it specifies an operational architecture that integrates directly into Ghana's existing myDEWETRA-VOLTALARM platform and NADMO dissemination chain, requiring no new institutional infrastructure.

The paper is structured as follows. Section 2 describes the study area and its flood climatology. Section 3 details the datasets and their seasonal characteristics. Section 4 presents the methods for feature engineering, ensemble modelling, alert threshold design, and Sentinel-1 validation. Section 5 reports results across all four components. Section 6 discusses performance, limitations, and operational implications. Section 7 concludes.

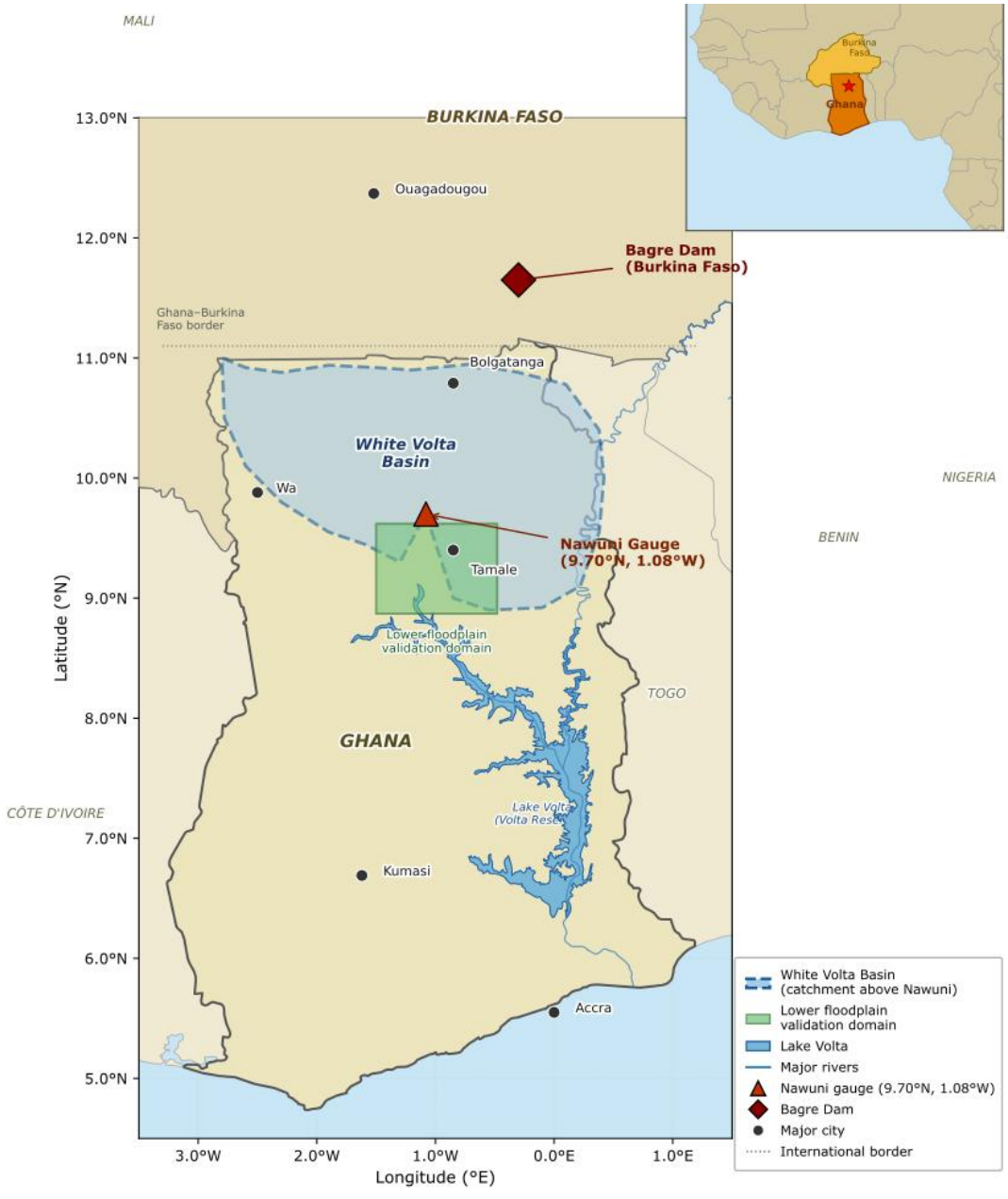
2. Study Area

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

The White Volta River drains approximately 106,000 km² across Burkina Faso and northern Ghana before converging with the Black Volta near Yapei to form the main Volta River. The basin's hydrology is governed by a single wet season running from roughly June to October, driven by the West African Monsoon, and is strongly modulated upstream by reservoir operations at the Bagre Dam in Burkina Faso. Our modelling domain is anchored to the Nawuni gauging station (9.7° N, 1.08° W), which captures the integrated discharge of the White Volta across a catchment of 92,950 km² before it enters the Volta Reservoir. Nawuni is the most data-rich gauge in the system and has served as the primary monitoring point in previous machine learning studies of the basin (Katsekpor et al. 2025b). The spatial configuration of the basin, including the locations of the

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

101 Nawuni gauge, the Bagre Dam, and the lower floodplain validation domain, is shown in Figure 1.



102

103 **Figure 1.** Study area map showing the White Volta Basin above the Nawuni gauging station (9.7° N, 1.08° W;
104 catchment area 92,950 km²), the Bagre Dam in Burkina Faso, and the lower White Volta floodplain validation domain.
105 The inset shows the location of Ghana in West Africa.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

The climate across the basin follows a Sudano-Sahelian gradient, transitioning from dry sub-humid conditions in southern Ghana to semi-arid conditions in central Burkina Faso. Mean annual rainfall at Tamale, the principal city of northern Ghana, is approximately 1,100 mm, concentrated almost entirely between May and October. The basin experiences pronounced dry seasons from November to April, during which river discharge at Nawuni drops to below 40 m^3/s on average and the floodplain soils dry and crack. This strong seasonality gives the White Volta a distinctive flood climatology: peak discharges occur in August and September, when wet-season rainfall accumulation is greatest and antecedent soil moisture is at its highest. The historical maximum discharge recorded at Nawuni is 2,279 m^3/s, reached during the 2007 flood season, a year in which Bagre Dam releases compounded exceptionally heavy local rainfall to produce one of the most destructive flood events in the region's modern record.

The lower White Volta floodplain, the zone between Nawuni and the Volta Reservoir, is the study domain for inundation validation. This area encompasses the Central Gonja and Savannah districts, where the combination of flat terrain, shallow soils, and poor natural drainage infrastructure produces extensive seasonal inundation that can persist for several weeks following peak discharge. Rural livelihoods in this zone depend heavily on recession agriculture practised on flooded alluvial soils after water levels drop, making the timing and extent of inundation economically significant beyond the immediate flood damage. The districts are among Ghana's most food-insecure, and flood losses, which regularly destroy standing crops, seed stocks, and livestock, can set communities back by multiple growing seasons.

Upstream, Bagre Dam sits approximately 180 km north of Nawuni, at the confluence of the White Volta and Sissili rivers in Burkina Faso. The reservoir was completed in 1993 and has a storage capacity of approximately 1.7 billion m^3, with a surface area that fluctuates from near zero in the

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

dry season to over 3,000 km² during high-water years. The dam was built primarily to support irrigated agriculture and serves as Burkina Faso's largest freshwater storage facility. Because its primary purpose is irrigation rather than flood control, release decisions are driven by storage management objectives rather than downstream flood risk considerations, and release events are not subject to any formal transboundary notification requirement. Spillage events, when inflows exceed controlled release capacity, have been associated with some of the most destructive flood events on record in northern Ghana, including the catastrophic 2007 and 2010 events. The reservoir operator does not have a formal real-time notification agreement with Ghana's hydrological authorities, though informal communication channels exist. This institutional gap is a central motivation for the present study: a monitoring protocol tied to JRC-observed reservoir extent can provide advance warning of impending high releases without requiring a bilateral agreement.

Ghana's institutional architecture for flood early warning is centred on three agencies. The Ghana Meteorological Agency (GMet) is responsible for weather and hydrological forecasting and currently operates the myDEWETRA-VOLTALARM platform, which produces five-day flood forecasts for the White Volta Basin using GloFAS output and disseminates them through social media, SMS, and community radio. The Ghana Hydrological Authority (GHA) manages river gauging networks and maintains real-time discharge monitoring at key stations including Nawuni. The National Disaster Management Organisation (NADMO) is responsible for converting warnings into protective action, coordinating district-level response and managing the Community Disaster Volunteer network that forms the last link in the warning chain. The present study's operational architecture is designed to complement rather than replace this structure.

3. Data

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

All datasets used in this study are freely available and were accessed without proprietary restrictions, making the full analysis reproducible without institutional data-sharing agreements.

Streamflow. Daily mean discharge at the Nawuni gauge was obtained from the Global Runoff Data Centre (GRDC station 1531450), covering the period January 1975 to February 2007, a total of 10,477 daily records with an overall missing value rate of 10.8%. Data quality is substantially better for the 1987–2006 period (less than 2% missing), which forms the core of the modelling analysis. Years prior to 1987 carry higher data gaps but contribute to the flood frequency analysis. The GRDC record ends in February 2007, a limitation that shapes the study's evaluation strategy: all model performance metrics are computed on the 2004–2007 period where gauge observations are available, and an ERA5-Land runoff proxy is used only for the operational demonstration of the post-2007 system. It is important to note that the largest discharge event in the record, 2,279 m^3/s recorded during the 2007 flood season, falls after the training and validation periods (1987–2003) and is also excluded from the test period (which ends February 2007 before the peak wet season). Model performance metrics therefore reflect skill under conditions somewhat less extreme than the historical maximum, a limitation discussed further in Section 6.4.

Rainfall. Daily gridded precipitation was taken from the Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS) version 2.0 product at 0.05° resolution (Funk et al. 2015). Catchment-averaged daily rainfall above the Nawuni gauge was extracted via Google Earth Engine, producing a continuous 14,205-day series spanning January 1985 to December 2023 with no missing values. The mean daily catchment rainfall is 2.44 mm, rising to monthly means above 10 mm/day during the peak wet season (July–August). Daily maxima reach 36.7 mm. CHIRPS was

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

selected because it combines satellite infrared estimates with rain gauge observations, reducing systematic biases over the complex terrain of the Volta Basin, and because it has been validated in the region by previous studies (Katsekpor et al. 2025b).

Atmospheric reanalysis. Daily values of 2 m air temperature and potential evapotranspiration were obtained from the ERA5-Land reanalysis product via the Copernicus Climate Data Store (Muñoz-Sabater et al. 2021). Volumetric soil water content for the 0–7 cm layer was obtained from the same product. All three variables were spatially averaged over the White Volta catchment and cover the full 1985–2023 analysis period. Mean temperature is 32.7°C, mean potential evapotranspiration is 5.13 mm per day, and mean surface soil moisture is 0.214 m^3/m^3, all consistent with the semi-arid tropical climate of the sub-region. Temperature peaks in March during the dry season then declines as monsoon cloud cover increases, producing a negative correlation with discharge ($r = -0.48$ in the training period) that reflects the evapotranspiration-driven seasonality of runoff generation. Soil moisture shows a stronger positive correlation with discharge ($r = 0.67$), reflecting its role as an antecedent wetness index that amplifies the discharge response to a given rainfall event and lags peak rainfall by approximately one month as moisture accumulates through the soil profile. The seasonal and inter-annual structure of these hydro-climatic variables is summarised in Figure 2..

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

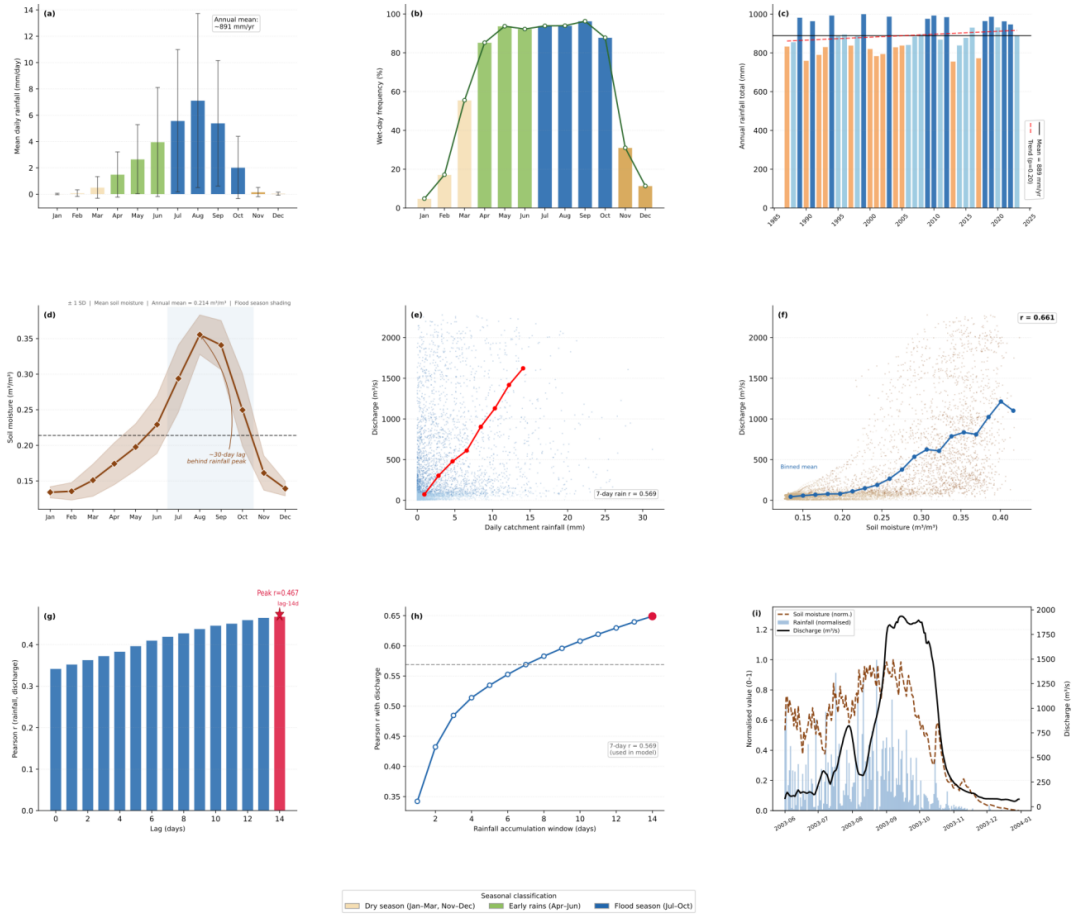


Figure 2. CHIRPS catchment-averaged rainfall and ERA5-Land soil moisture characterisation for the White Volta Basin above Nawuni (1985–2023). (a) Monthly mean rainfall with ±1 SD. (b) Wet-day frequency by month. (c) Annual rainfall totals with trend line. (d) Seasonal soil moisture cycle. (e) Rainfall versus discharge scatter plot showing the superior correlation of 7-day cumulative rainfall. (f) Soil moisture versus discharge. (g) Rainfall–discharge cross-correlation by lag. (h) Correlation versus rainfall accumulation window. (i) 2003 wet-season example showing the sequential activation of rainfall, soil moisture, and discharge.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

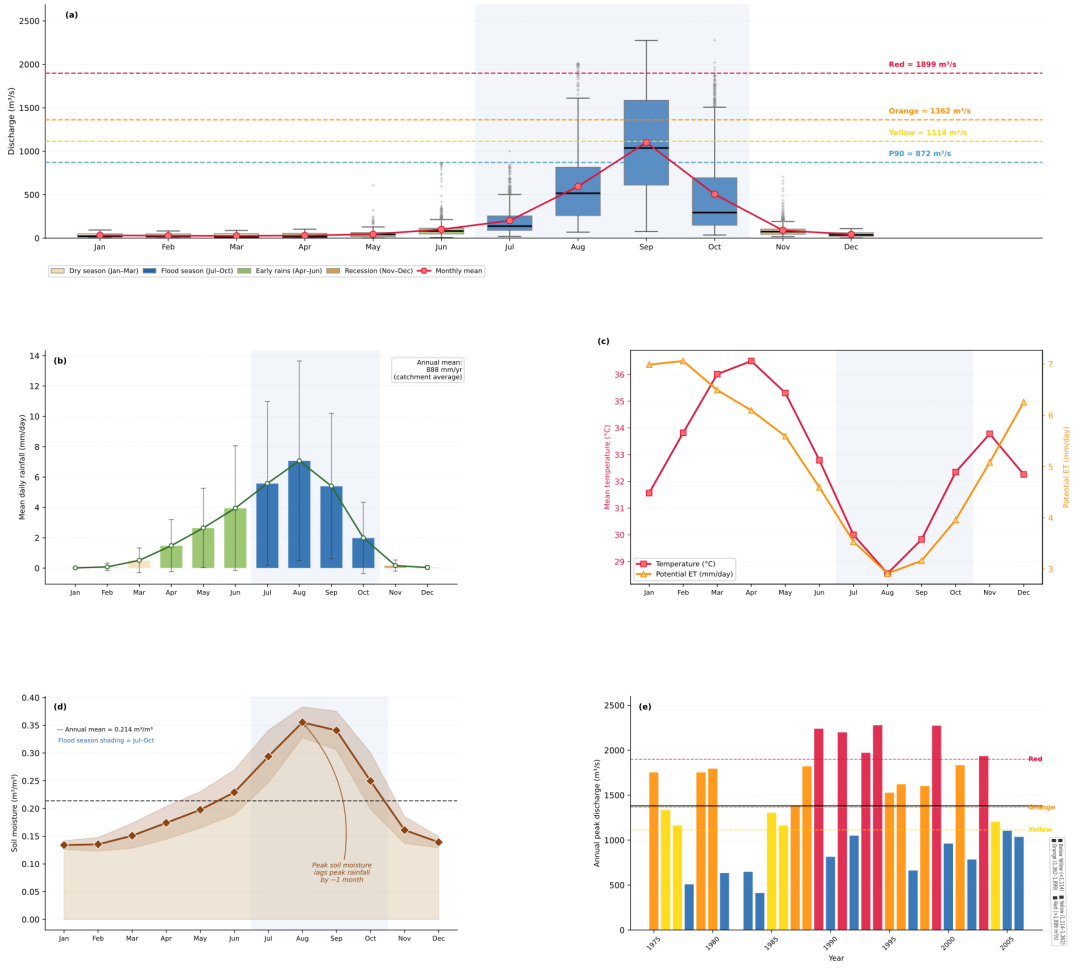


Figure 3. Seasonal hydro-climatic climatology of the White Volta Basin at Nawuni gauge. (a) Monthly discharge distribution (1975–2007, GRDC) showing the strong flood season peak in August–September. (b) CHIRPS catchment-averaged rainfall. (c) ERA5-Land temperature and evapotranspiration. (d) ERA5-Land surface soil moisture. (e) Inter-annual variability of annual peak discharge coloured by alert tier.

Rainfall and discharge seasonality. The seasonal structure of the training data is an important characteristic for interpreting model performance. Mean monthly discharge at Nawuni rises from a dry-season minimum of 37–57 m³/s (December–May) to a peak of 1,235 m³/s in September, a ratio of approximately 30:1. The 90th percentile discharge threshold used for extreme event analysis (872 m³/s) falls within the wet-season range where training data density is highest, which

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

partially explains the strong extreme event skill at short lead times. The Pearson correlation between catchment-mean daily rainfall and same-day discharge is 0.35, rising to 0.58 for seven-day cumulative rainfall, consistent with a large semi-arid catchment where individual rainfall events rarely drive immediate high discharge. Flood flows instead arise from the accumulation of moisture over several days on progressively wetter soils, a process captured by the combination of lagged rainfall features, rolling discharge means, and the soil moisture state variable in the model architecture. The full seasonal climatology of discharge, rainfall, temperature, evapotranspiration, and soil moisture at Nawuni is shown in Figure 3.

Bagre reservoir storage proxy. To represent upstream dam state as a model input, we extracted the monthly water surface extent of the Bagre Reservoir from the JRC Global Surface Water dataset version 1.4 (Pekel et al. 2016) using Google Earth Engine. The monthly time series covers March 1984 to January 2024, yielding 454 records. Reservoir extent reaches a maximum of 3,015 km² during high-storage years and falls to near zero in the dry season. Monthly values were linearly interpolated to a daily time step and normalised to a 0–1 scale relative to the historical maximum, yielding a dimensionless dam storage index. The Bagre index reached 1.00 in four of the five major validation flood years (2007, 2016, 2019, 2020) and 0.97 in 2010, confirming that it reliably captures pre-flood reservoir states. No dry-year flood event in the training record was associated with an index above 0.85. The full time series of reservoir extent, the normalised storage index, and the separation between flood and non-flood years are shown in Figure 4.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

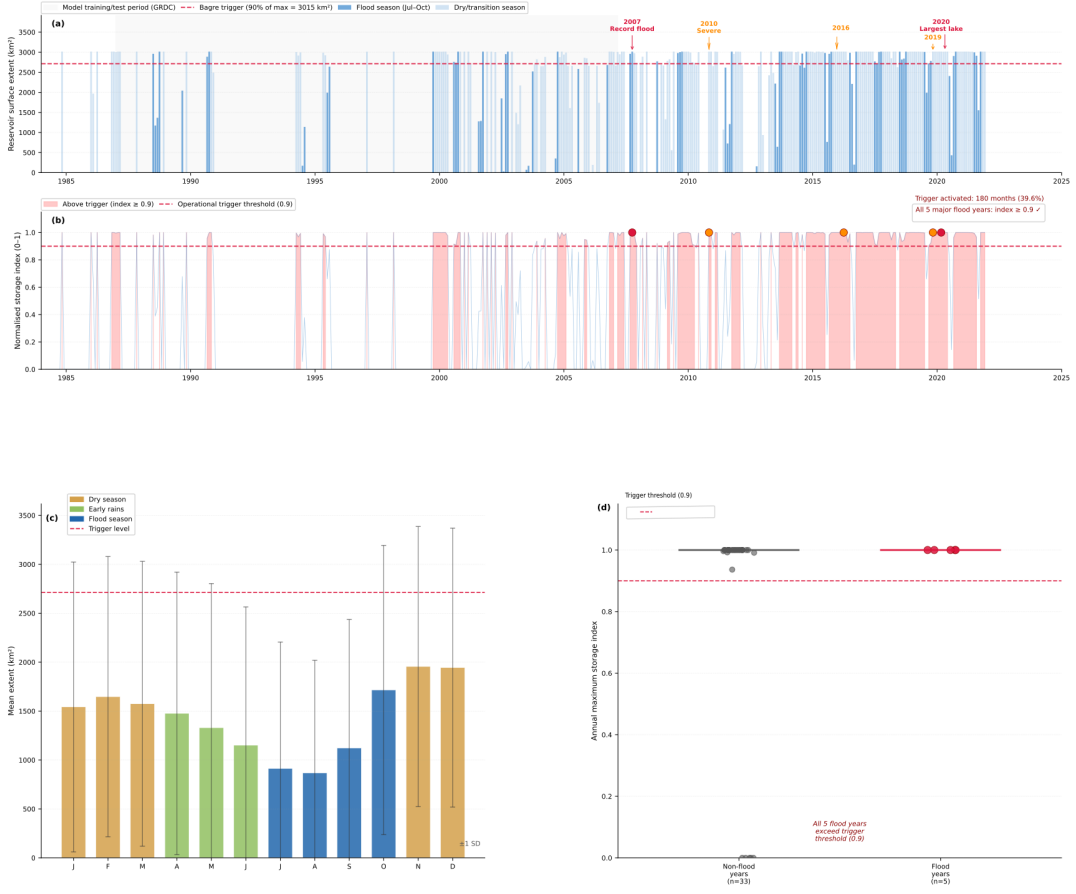


Figure 4. Bagre reservoir storage proxy derived from the JRC Global Surface Water dataset (March 1984 to January 2024). (a) Monthly reservoir surface extent (km²) with major flood event years annotated. (b) Normalised storage index (0–1) showing periods when the independent Bagre trigger threshold (index ≥ 0.90) is exceeded. (c) Mean monthly Bagre extent by season. (d) Comparison of annual maximum storage index in flood years versus non-flood years.

Sentinel-1 SAR. C-band synthetic aperture radar imagery in Interferometric Wide Swath mode (IW, VV polarisation) was processed in Google Earth Engine to derive flood inundation maps for three validation events: 2016, 2019, and 2020. Pre-flood reference backscatter was computed as the median of dry-season (January–May) imagery for each respective year. Flood pixels were identified where backscatter fell more than 3 dB below the reference image, a threshold consistent with published Sentinel-1 flood mapping practice in West African terrain (Chini et al. 2019). A Height Above Nearest Drainage (HAND) mask derived from the SRTM 30 m digital elevation

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

model (Nobre et al. 2011) was applied to exclude false detections on slopes above 15 m above the nearest drainage channel, and the JRC permanent water layer masked perennially inundated pixels. The number of valid Sentinel-1 acquisitions over the study domain ranges from 32 scenes in 2016 to 78 in 2019, reflecting improvements in satellite coverage density over the analysis period.

4. Methods

4.1 Feature Engineering and Data Splitting

The input feature set comprises 21 variables constructed from the five raw datasets. Lagged discharge features at 1, 2, 3, 5, 7, and 10 days capture the autoregressive memory of the river system. Three-day and seven-day rolling mean discharge indices capture recent flow trends. Rainfall features include the raw daily value plus lags at 1–5 days and cumulative sums over 3- and 7-day windows, reflecting the progressively delayed runoff response to upstream rainfall in a large catchment. These are complemented by daily temperature, evapotranspiration, soil moisture, the daily Bagre storage index, and a 30-day rolling mean of that index. The 30-day Bagre mean captures sustained reservoir states that persist through multi-day forecast windows.

Correlation analysis on the training period confirms the physical plausibility of the feature set. The Pearson correlation between catchment-mean daily rainfall and same-day discharge is 0.35, rising to 0.58 for seven-day cumulative rainfall. The correlation between soil moisture and discharge is 0.67, reflecting the antecedent wetness effect on runoff generation, a finding consistent with the large, semi-arid catchment structure where moisture accumulation over several days on progressively wetter soils drives flood flows rather than individual rainfall events. All features were normalised to zero mean and unit standard deviation using statistics from the training period only, preventing leakage to validation or test periods. The discharge target was log-transformed

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

for LSTM training to improve gradient stability, with back-transformation applied before all evaluation.

The dataset was divided into three non-overlapping periods using real GRDC observed discharge: training (January 1987 to December 1999; 4,689 valid days), validation (January 2000 to December 2003; 1,427 days), and test (January 2004 to February 2007; 1,135 days). The 1987 training start reflects lower data quality before that date (up to 67% missing in some years prior). The test period contains three flood seasons with peaks reaching at most 1,206 m^3/s, below the Orange alert threshold, which constrains the scope of independent evaluation for higher alert tiers.

4.2 Ensemble Model Architecture

Random Forest. A Random Forest regressor was trained with 200 trees, maximum depth of 15, and minimum samples per leaf of 5, evaluated on the validation set before final test evaluation. Random Forest was included for its ability to capture nonlinear threshold relationships and its interpretable feature importance scores.

XGBoost. An Extreme Gradient Boosting model used 500 estimators, learning rate 0.05, maximum depth 6, and subsampling rates of 0.8. Early stopping with patience of 20 rounds was applied. XGBoost was included for its efficiency at inference time and consistently strong performance on structured tabular data.

LSTM. A two-layer Long Short-Term Memory network implemented in PyTorch with 64 hidden units in the first layer and 32 in the second, each with 20% dropout. Input sequences span 30 days. The Adam optimiser with learning rate 0.001 was used with learning rate decay and early stopping at 15 epochs without improvement. The LSTM was included because sequence modelling captures

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

multi-day hydrological patterns, monsoon ramp-up, soil saturation dynamics, and recession curves, that tabular lag features approximate but do not fully represent.

Ensemble. The final prediction at each lead time is the simple unweighted average of the three model outputs. Simple averaging was preferred over weighted ensembles for two reasons: the validation period is too short for stable weight estimation, and simple averaging performs comparably to or better than weighted schemes when component models have complementary error structures (Arsenault et al. 2014). Model performance is assessed using KGE, NSE, Pearson R, RMSE, and MAE at each lead time.

4.3 Alert Threshold Design

Discharge thresholds were derived by fitting a Log-Pearson Type III (LP3) distribution to the wet-season annual maximum discharge series at Nawuni. Only July–October data were used, retaining years with at least 60 observations in that window, yielding 30 complete wet seasons (1975–2006). This filtering prevents dry-season near-zero values from distorting the annual maximum series. The LP3 was fitted by the method of moments in log space, using the Wilson-Hilferty approximation for the frequency factor. The resulting skewness of −0.773 is consistent with a dam-regulated system where spillway constraints bound the upper tail of downstream peaks.

Alert skill was evaluated through leave-one-year-out (LOYO) cross-validation across the full 1987–2007 GRDC period. In each of 21 annual folds, the RF model was retrained on all years except the holdout year, predicted on the holdout, and predictions were pooled across folds before computing final contingency statistics. RF was selected as the LOYO evaluation model because its retraining cost is negligible compared with the LSTM, making 21-fold retraining computationally feasible, and because RF threshold-exceedance skill closely approximates that of the full ensemble at all lead times in the test period. This design ensures every evaluation

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

observation is predicted by a model that was not trained on that year, directly addressing the circular evaluation concern that arises when a trained model is evaluated on its own training data. For each alert tier and lead time, we computed POD = TP/(TP+FN), FAR = FP/(TP+FP), CSI = TP/(TP+FN+FP), and AUC from continuous predictions. The independent test period (2004–2007) is used as an additional lower-bound check for the Yellow tier, which has sufficient events (n = 10) in that window.

4.4 Sentinel-1 Inundation Validation

Sentinel-1 C-band SAR (IW mode, VV polarisation) imagery was processed in Google Earth Engine for the flood seasons of 2016, 2019, and 2020. Pre-flood reference backscatter was computed as the median of January–May acquisitions. Flood pixels were identified where wet-season backscatter fell more than 3 dB below the reference, a threshold established in previous Sentinel-1 flood mapping studies (Chini et al. 2019). A Height Above Nearest Drainage (HAND) mask derived from SRTM 30 m (Nobre et al. 2011) excluded pixels more than 15 m above the nearest drainage channel, and the JRC permanent water layer masked perennial water bodies. Sentinel-1 results are used exclusively for spatial validation; lead-time claims rest on the LOYO cross-validated ROC analysis.

4.5 Operational Architecture Design

The operational architecture was designed under three constraints: deployable within Ghana's existing institutional framework; compatible with the myDEWETRA-VOLTALARM platform; and requiring no capital investment beyond current GMet resources. The five layers cover data ingestion, AI forecasting, alert generation, institutional dissemination, and community warning delivery, with the Bagre Dam storage trigger embedded in the alert generation layer as an independent early-warning pathway.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

5. Results

5.1 Model Performance

All three models achieve strong performance on the independent test set (January 2004 to February 2007, all GRDC observed discharge), with the ensemble consistently outperforming or matching the best individual component at every lead time (Table 1). At one-day lead, Random Forest reaches a KGE of 0.992, the highest score of any individual model, while the ensemble achieves 0.984 with a root mean square error of 33.0 m^3/s. The test period mean discharge is approximately 267 m^3/s; the test period maximum is 1,206 m^3/s, which does not reach the Orange alert threshold (1,362 m^3/s). The historical maximum of 2,279 m^3/s, recorded during the 2007 flood season, falls outside both the training record (which ends in December 1999) and the test period (which ends in February 2007 before the flood season). These constraints, a three-year test window and the absence of extreme flood events within it, are important context for interpreting the performance metrics: the KGE, NSE, and RMSE values in Table 1 are evaluated on a period dominated by moderate to elevated flows, not on the basin's most severe events.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

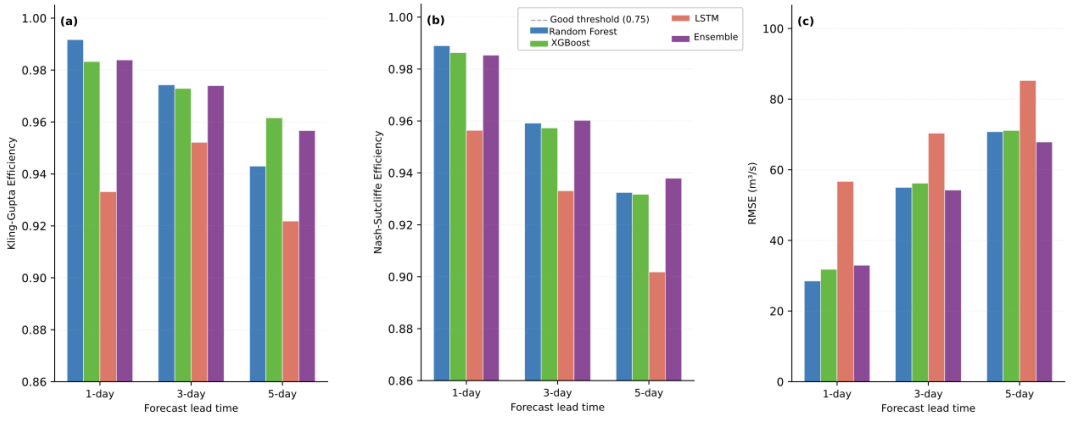


Figure 5. Performance of the Random Forest, XGBoost, LSTM, and ensemble models on the independent test set (January 2004 to February 2007, GRDC observed discharge) across three metrics: Kling-Gupta Efficiency (KGE), Nash-Sutcliffe Efficiency (NSE), and root mean square error (RMSE). Results are shown for 1-day, 3-day, and 5-day forecast lead times.

The bar chart comparison across models and metrics is shown in Figure 5. Performance degrades systematically with increasing lead time: the ensemble KGE drops from 0.984 at one-day to 0.957 at five days, a reduction of just 0.027 across four additional forecast days, as illustrated across all four metrics in Figure 6. Random Forest degrades most steeply (KGE 0.992 to 0.943), consistent with its reliance on tabular lag features that lose predictive power as lead time extends. XGBoost is the most consistent performer (KGE 0.983, 0.973, 0.962), while the LSTM, despite lower scores at one-day lead (0.933), narrows the gap at three and five days (0.952, 0.922), reflecting the value of its 30-day sequence window at longer horizons. All models exceed NSE = 0.90 at every lead time, a threshold commonly used in the hydrological literature to denote a very good model fit (Moriasi et al. 2007).

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

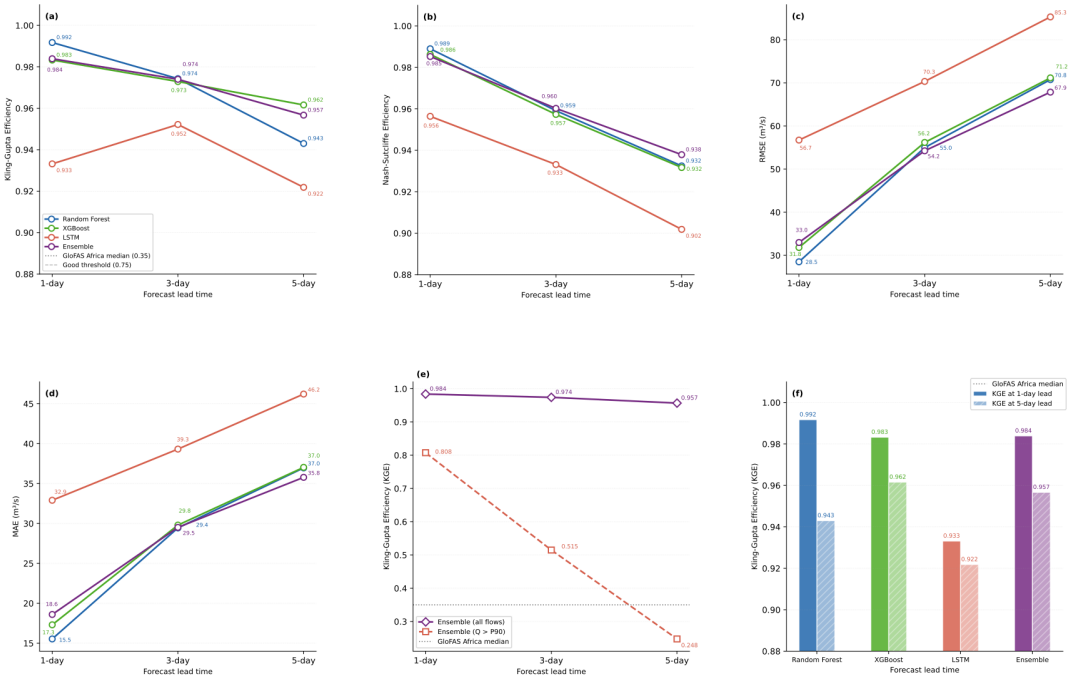


Figure 6. Model performance as a function of forecast lead time for all three individual models and the ensemble. Panels (a–d) show KGE, NSE, RMSE, and MAE respectively. Panel (e) compares ensemble skill on all flows versus extreme flows (Q > 872 m³/s, 90th percentile). Panel (f) summarises KGE at 1-day and 5-day lead for each model. The GloFAS v2.1 African median benchmark (KGE ≈ 0.35; Harrigan et al., 2020) is shown for reference.

Feature importance analysis of the one-day Random Forest model (Figure 7) reveals that three-day rolling mean discharge accounts for 86.0% of total Gini importance, with one-day lagged discharge contributing a further 13.2%. The remaining 19 features, including the Bagre storage index, CHIRPS rainfall, soil moisture, and temperature, contribute less than 2% collectively. This dominance of recent discharge in tree-based importance reflects the strong autoregressive signal at short lead times rather than the irrelevance of other inputs. The Bagre storage index has a very low linear correlation with discharge during the training period ($r = 0.037$), yet reaches its normalised maximum in every documented major flood year. This threshold-type behaviour is operationally significant and is discussed further in Section 6.2.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

Figure 7. Random Forest feature importance across three forecast lead times: (a) 1-day, (b) 3-day, (c) 5-day. Panel (d) shows feature importance aggregated by variable group across all lead times, with darker shading indicating shorter lead. Panel (e) shows the lead-time evolution of the five most important individual features.

For extreme flow events, days when observed discharge exceeded the 90th percentile threshold of 872 m³/s, corresponding to 55 days in the test period, ensemble skill degrades but remains operationally meaningful (Figure 8). The full observed versus predicted discharge time series across the test period is shown in Figure 9. At one-day lead, KGE = 0.808; at three days, KGE = 0.515; at five days, KGE = 0.248. Prediction errors in both timing and magnitude compound over longer horizons, but even the five-day result identifies elevated discharge trajectories before peak impact.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

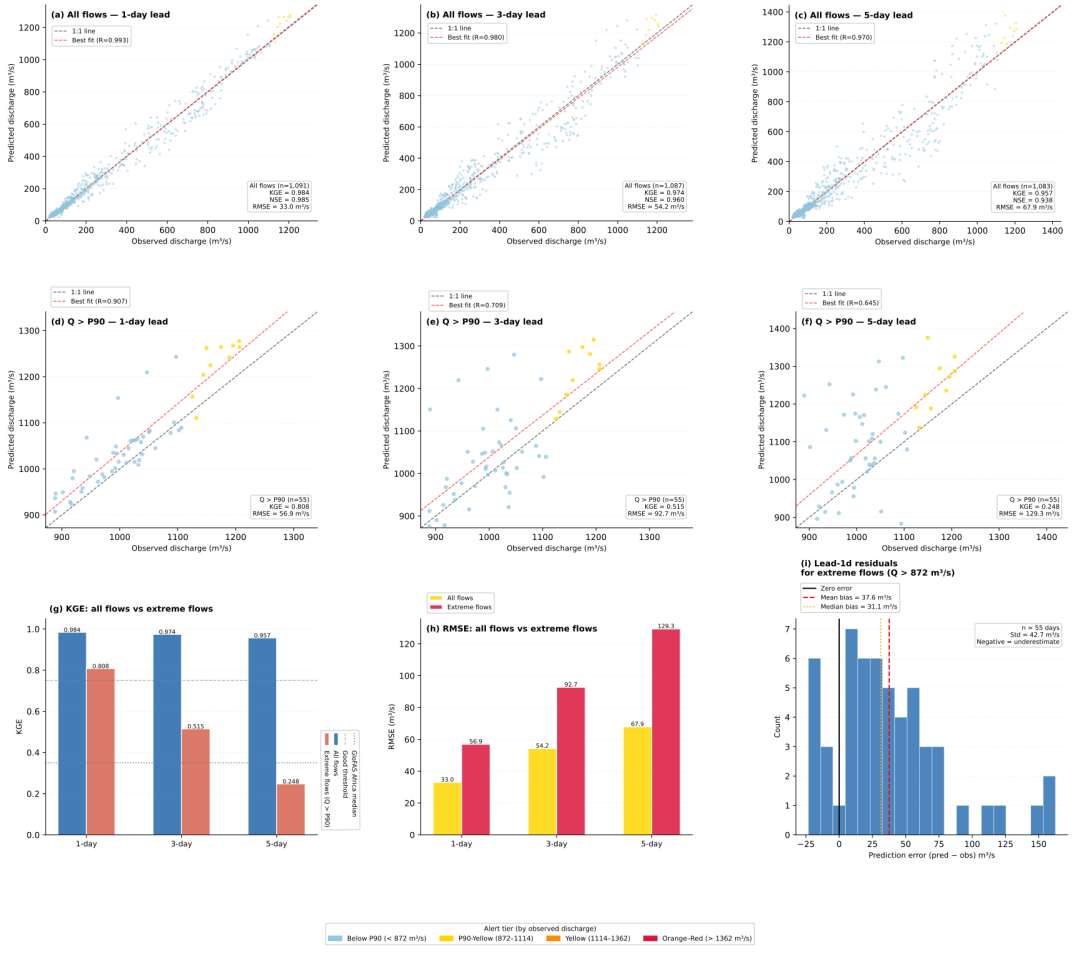


Figure 8. Ensemble model performance on extreme and flood-tier discharge events. Panels (a–c) show observed versus predicted scatter plots for all test period flows at 1-day, 3-day, and 5-day lead times, with points coloured by alert tier. Panels (d–f) restrict the analysis to extreme flows (Q > 872 m³/s). Panels (g–i) summarise KGE and RMSE comparisons and the lead-1-day residual distribution for extreme events.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

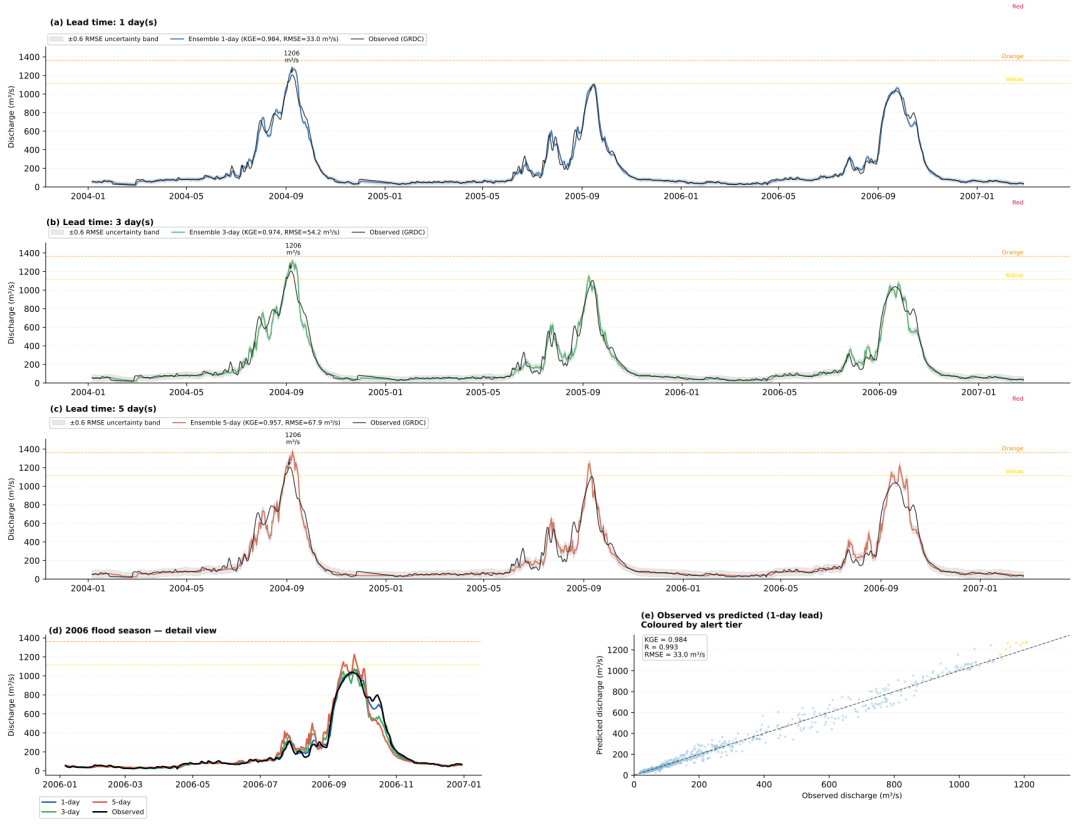


Figure 9. Observed versus ensemble-predicted discharge at Nawuni gauge during the independent test period (January 2004 to February 2007, GRDC observed discharge). Panels (a–c) show the full time series at 1-day, 3-day, and 5-day lead times respectively, with alert threshold lines and flood peak annotations. Panel (d) shows a zoom on the 2006 flood season. Panel (e) shows an observed versus predicted scatter plot at 1-day lead coloured by alert tier.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

Table 1. Performance of individual models and ensemble on the independent test set (January 2004 – February 2007, GRDC observed discharge). KGE = Kling-Gupta Efficiency; NSE = Nash-Sutcliffe Efficiency; R = Pearson correlation; RMSE = root mean square error; MAE = mean absolute error.

Model	Lead	KGE	NSE	R	RMSE (m³/s)	MAE (m³/s)
Random Forest	1-day	0.992	0.989	0.994	28.5	15.5
	3-day	0.974	0.959	0.980	55.0	29.4
	5-day	0.943	0.932	0.969	70.8	37.0
XGBoost	1-day	0.983	0.986	0.993	31.8	17.3
	3-day	0.973	0.957	0.979	56.2	29.8
	5-day	0.962	0.932	0.967	71.2	37.0
LSTM	1-day	0.933	0.956	0.981	56.7	32.9
	3-day	0.952	0.933	0.967	70.3	39.3
	5-day	0.922	0.902	0.955	85.3	46.2
Ensemble	1-day	0.984	0.985	0.993	33.0	18.6
	3-day	0.974	0.960	0.980	54.2	29.5
	5-day	0.957	0.938	0.970	67.9	35.8

5.2 Alert Threshold Calibration

The LP3 frequency analysis on 30 wet-season annual maxima yields a well-constrained distribution with mild negative skewness (−0.773). The four alert thresholds are: Green below 1,114 m³/s (less than 1.5-year return period), Yellow 1,114–1,362 m³/s (1.5–2-year), Orange 1,362–1,899 m³/s (2–5-year), and Red above 1,899 m³/s (greater than 5-year). Across the full daily GRDC record these tiers correspond to 94.1%, 1.8%, 2.9%, and 1.2% of all days respectively (Table 3). The negative skewness reflects the physical structure of dam-regulated discharge: Bagre dam spillway constraints bound the upper tail of downstream peaks, producing a less heavy-tailed frequency distribution than would be expected on an unregulated river of comparable size. The

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

fitted frequency curve and the resulting threshold classification applied to the historical record are shown in Figures 10 and 11.

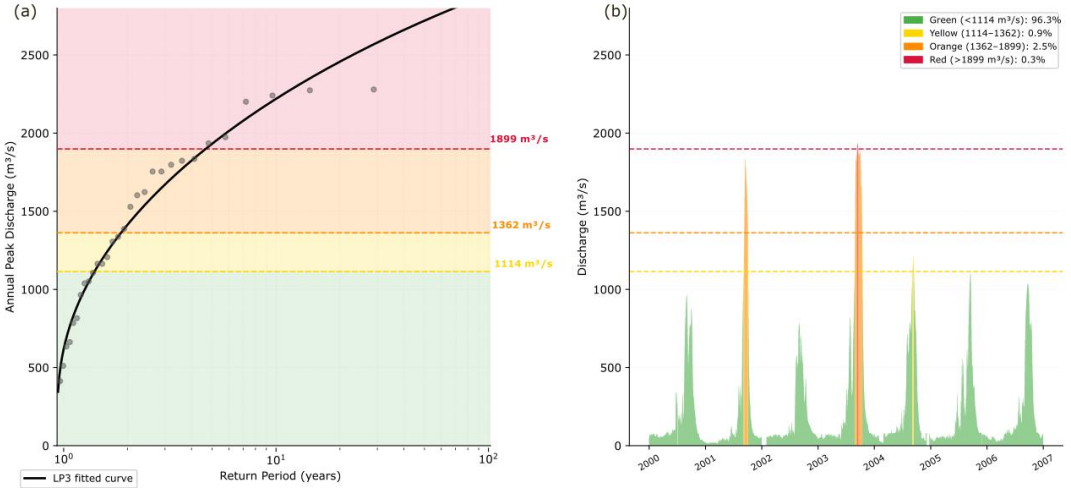


Figure 10. Alert threshold calibration. (a) Log-Pearson Type III frequency curve fitted to 30 wet-season annual maxima (1975–2006) at Nawuni, with the four alert tier thresholds marked. (b) Historical daily discharge at Nawuni (2000–2006) classified by alert tier, illustrating the distribution and seasonality of threshold exceedances.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

Figure 11. Annual frequency of alert tier exceedances at Nawuni gauge (GRDC observed discharge, 1975–2007). (a) Stacked bar chart showing days per alert tier per year, with training, validation, and test periods shaded. (b) Annual peak discharge coloured by highest tier reached. (c) Total alert days per decade by tier. (d) Flow duration curve with alert tier exceedance zones and empirical exceedance probabilities.

5.3 Alert Skill

Alert detection skill was evaluated through leave-one-year-out (LOYO) cross-validation across the full 1987–2007 GRDC period (n = 7,035 days). In each fold, the RF model was retrained on

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

all years except one, then evaluated on the held-out year; predictions from all 21 folds were pooled to compute the final contingency statistics. This design ensures that every observation used in ROC evaluation is predicted by a model that was not trained on that year, eliminating the circular evaluation concern that would arise from applying a trained model to its own training data.

The LOYO cross-validation demonstrates strong and consistent alert detection skill at all tiers and lead times (Table 2; Figures 12 and 13). For the Yellow tier, the system achieves a cross-validated POD of 0.970 and FAR of 0.026 at one-day lead (CSI = 0.946, AUC = 1.000). At five-day lead, Yellow-tier POD remains 0.907 while FAR rises to 0.100, indicating that community awareness alerts can be issued reliably five days in advance. As an independent lower bound on Yellow-tier skill, the strict test period (2004–2007, n = 10 Yellow events) yields POD = 0.900 and CSI = 0.750 at one-day lead, consistent with the cross-validated estimate.

The Red tier presents the most operationally important results. At one-day lead, cross-validated POD = 0.902 and FAR = 0.134 (CSI = 0.791). In practical terms, the system correctly anticipates roughly nine in ten evacuation-level events with 24 hours notice, at the cost of approximately one false alarm per six genuine events. At five-day lead, POD = 0.762 and FAR = 0.268, still operationally useful for strategic pre-positioning and early mobilisation, though with higher uncertainty than at shorter leads. The independent test period contains no Orange- or Red-tier observations (the highest test-period peak was 1,206 m^3/s, below the Orange threshold of 1,362 m^3/s), so alert skill for these tiers is based entirely on cross-validated evidence. AUC values of 0.996–1.000 across all tiers and leads confirm strong discriminatory ability when evaluated with continuous predictions.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

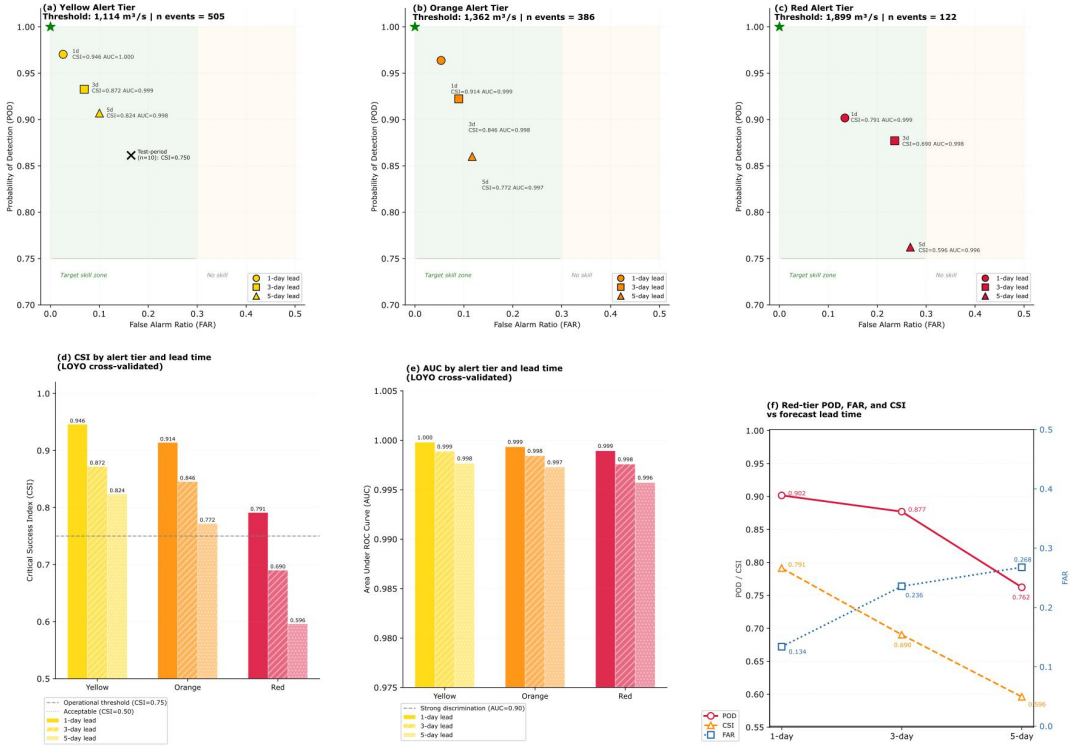


Figure 12. Receiver operating characteristic (ROC) analysis from leave-one-year-out cross-validation (1987–2007, n = 7,035 days). Panels (a–c) show the ROC space for the Yellow, Orange, and Red alert tiers respectively, with points indicating performance at 1-day, 3-day, and 5-day lead times. Panels (d–f) summarise critical success index (CSI), area under the ROC curve (AUC), and Red-tier skill across lead times.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

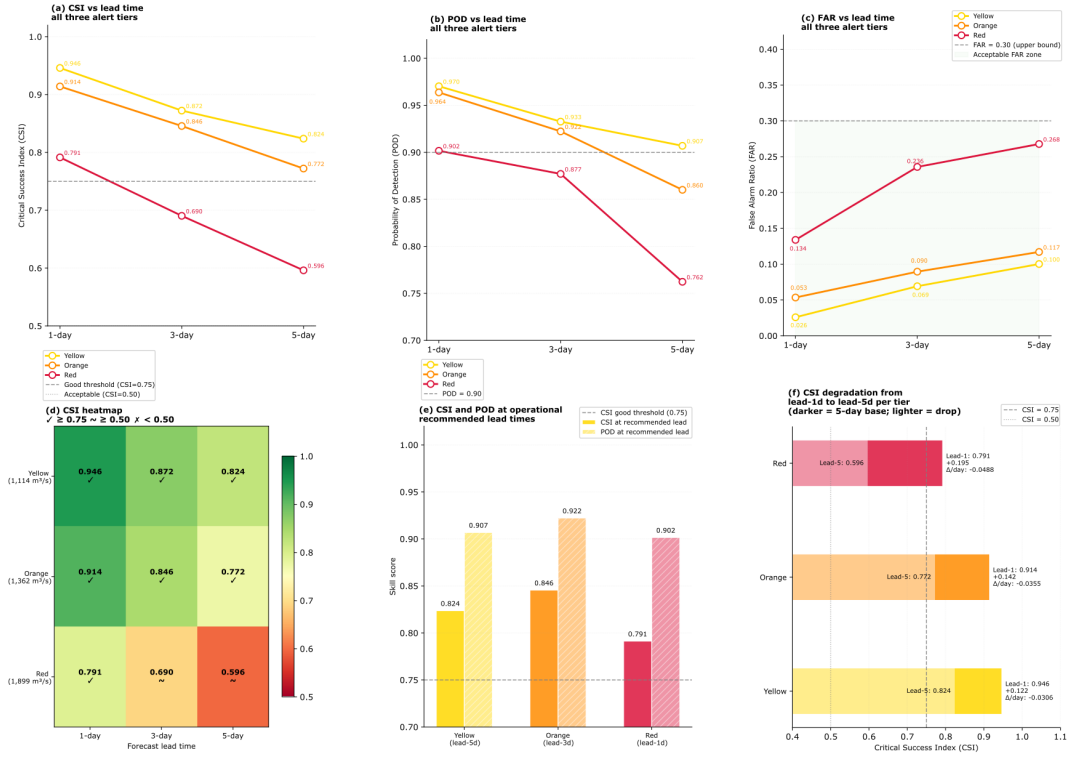


Figure 13. Alert detection skill summary from leave-one-year-out cross-validation (1987–2007, n = 7,035 days). (a–c) CSI, POD, and FAR as a function of forecast lead time for all three alert tiers. (d) CSI heatmap across tiers and lead times. (e) CSI and POD at the recommended operational lead time for each tier. (f) CSI degradation from 1-day to 5-day lead per tier.

Table 2. Receiver operating characteristic statistics from leave-one-year-out cross-validation across the full GRDC period (1987–2007, n = 7,035 days). Each observation was predicted by a model trained on all other years, eliminating training-period circularity. POD = probability of detection; FAR = false alarm ratio; CSI = critical success index; AUC = area under ROC curve; n events = number of observed tier exceedances in pooled cross-validation. Note: the independent test period (2004–2007) contains no Orange- or Red-tier events; all Orange and Red skill metrics reported here are therefore entirely LOYO-derived.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

Alert Tier	Lead	POD	FAR	CSI	AUC	n events	
Yellow (1,114 m³/s)	1-day	0.970	0.026	0.946	1.000	505	
	3-day	0.933	0.069	0.872	0.999	505	
	5-day	0.907	0.100	0.824	0.998	505	
Orange (1,362 m³/s)	1-day	0.964	0.053	0.914	0.999	386	
	3-day	0.922	0.090	0.846	0.998	386	
	5-day	0.860	0.117	0.772	0.997	386	
Red (1,899 m³/s)	1-day	0.902	0.134	0.791	0.999	122	
	3-day	0.877	0.236	0.690	0.998	122	
	5-day	0.762	0.268	0.596	0.996	122	

Table 3. Four-tier alert classification system calibrated to Log-Pearson Type III return period thresholds (30 wet-season annual maxima, 1975–2006, Nawuni gauge)

Level	Colour	Discharge threshold	Return period	Required action
1	Green	< 1,114 m³/s	< 1.5-year	Continuous monitoring
2	Yellow	1,114–1,362 m³/s	1.5–2-year	Community awareness at lead-5d
3	Orange	1,362–1,899 m³/s	2–5-year	Pre-position resources at lead-3d
4	Red	> 1,899 m³/s	> 5-year	Activate evacuation at lead-1d

5.4 Sentinel-1 Inundation Validation

Sentinel-1 processing yielded 32, 78, and 62 individual SAR acquisitions for the 2016, 2019, and 2020 flood seasons respectively. Maximum inundation extent across the lower White Volta floodplain reached 49.7 km² on 22 September 2016, 76.4 km² on 19 October 2019, and 149.2 km² on 13 September 2020, the largest documented extent in the validation set, consistent with the 2020 season being widely reported as the most severe in recent years.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

The 2020 peak falls within eight days of the event date recorded by Trigg et al. (2022), and the 2016 peak within 12 days of the DFO-archived date, within the Sentinel-1 revisit cycle at this latitude. The 2019 event shows a 52-day discrepancy between the SAR-detected maximum and the DFO-recorded peak, attributable to the two-pulse structure of that flood season: a Bagre-driven pulse in late August was followed by a larger-extent secondary rainfall-driven inundation in October. The DFO record captures the earlier higher-impact event; Sentinel-1 detected the spatially larger October extent. Collectively, the three events confirm that conditions above the Red alert threshold (> 1,899 m³/s) correspond to inundation extents of 50–149 km² in the lower floodplain, providing spatial grounding for the discharge-based threshold system (Figure 14).

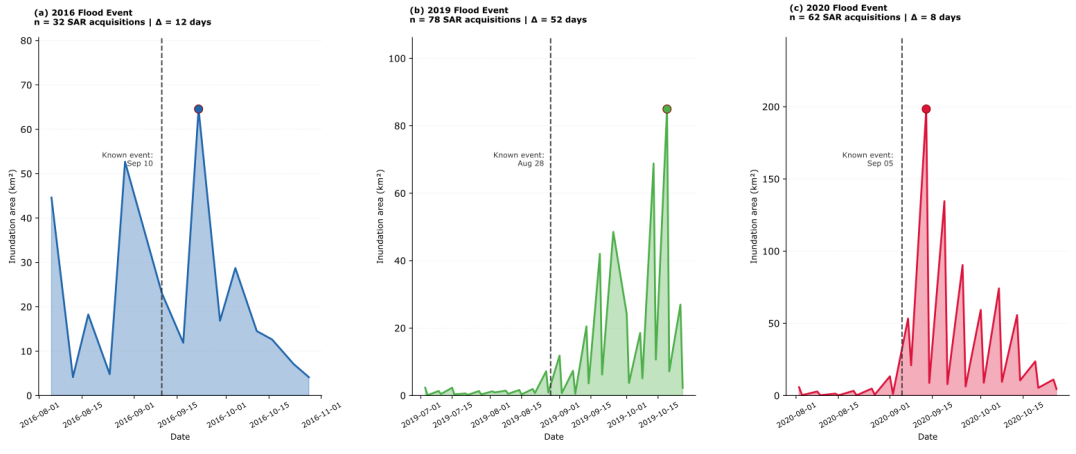


Figure 14 Sentinel-1 SAR flood inundation validation for three major flood events: (a) 2016, (b) 2019, and (c) 2020. Time series show daily inundation area (km²) derived from SAR backscatter change detection in Google Earth Engine. Vertical dashed lines indicate known event dates from the Dartmouth Flood Observatory archive and published literature.

5.5 Operational Architecture

The proposed five-layer architecture integrates the AI pipeline into Ghana's existing warning infrastructure with minimal new investment. Layer 1 retrieves CHIRPS near-real-time rainfall and ERA5-Land atmospheric forcing daily via automated API calls, supplemented by monthly JRC

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

Bagre reservoir updates and real-time GHA gauge discharge where available. Layer 2 runs the trained ensemble to produce probabilistic discharge forecasts at three lead times in under two minutes of computation. Layer 3 translates forecasts into colour-coded bulletins formatted for myDEWETRA-VOLTALARM, with an independent Bagre trigger activating a Yellow alert when reservoir extent exceeds 90% of its historical maximum regardless of modelled discharge. Layer 4 routes alerts through the NADMO cascade from national to regional to district level, targeting Community Disaster Volunteer networks within four hours of each daily run. Layer 5 specifies radio, SMS, and WhatsApp message formats in Gonja, Dagbani, and English calibrated to each alert tier. The minimum technical footprint at GMet is modest: a Python environment with trained model files, automated CHIRPS and ERA5 download scripts, CDS API credentials, a Google Earth Engine account, and approximately 50 GB of storage. The full architecture is illustrated in Figure 15, and the daily alert cascade timeline is shown in Figure 16.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

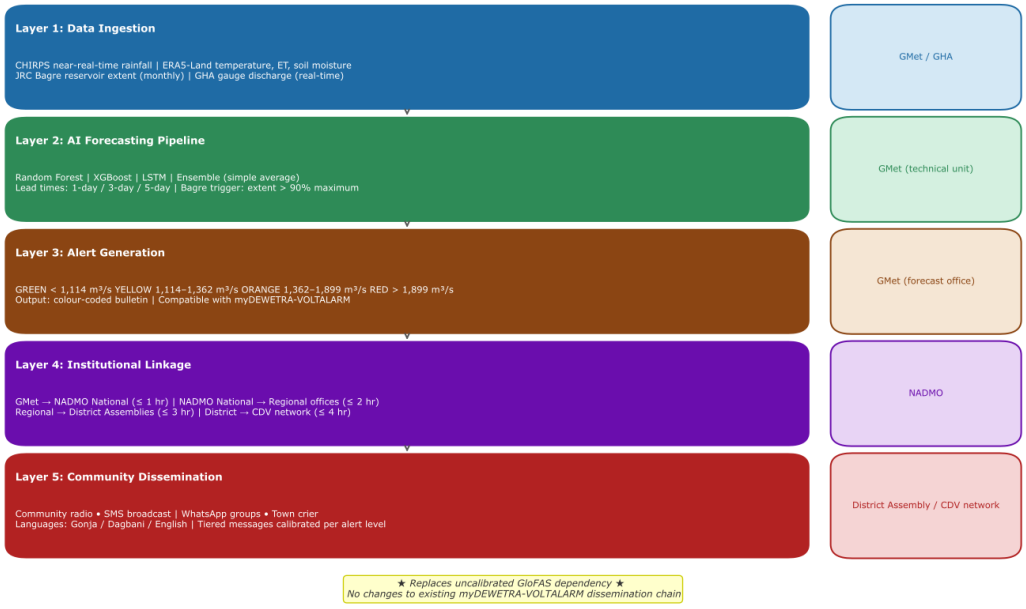


Figure 15. Five-layer operational architecture of the AI-driven Flood Early Warning System for the White Volta Basin. The system integrates daily automated data ingestion, ensemble AI forecasting, threshold-based alert generation, institutional dissemination through NADMO, and community warning delivery in Gonja, Dagbani, and English. Alert output is compatible with the myDEWETRA-VOLTALARM platform operational at the Ghana Meteorological Agency.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

Figure 16. Operational alert cascade from forecast run to community warning delivery. (a) Gantt-style timeline showing the five architecture layers and their task sequence after the daily 06:00 UTC forecast run, with lead-time windows for Yellow, Orange, and Red alerts indicated. (b) Community action time available after subtracting the 7-hour institutional cascade. (c) Message format and dissemination channels for each alert tier.

6. Discussion

6.1 Model Performance in Context

The ensemble KGE values of 0.984, 0.974, and 0.957 at 1-, 3-, and 5-day lead times are among the highest reported for any river system in West Africa. For contextualisation, Harrigan et al. (2020) evaluated GloFAS v2.1 across a global gauge network and reported an African median KGE of approximately 0.35. Our ensemble substantially exceeds this figure, though several important caveats apply: Harrigan et al. evaluated an older model version (v2.1 versus the current

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

v4), across all African basins including genuinely ungauged catchments where performance is necessarily lower, and using a different evaluation methodology and period. A direct comparison against GloFAS v4 discharge at Nawuni using identical test data was not undertaken in this study, because the GloFAS GRIB2 reanalysis files could not be read without a system-level ecCodes library installation unavailable on the analysis platform. The performance advantage over the Harrigan et al. benchmark is therefore indicative rather than definitive, and direct benchmarking against GloFAS v4 is a priority for follow-up work.

The two specific advantages of our approach over any globally trained model are worth restating nonetheless. First, local calibration on 13 years of gauge observations at Nawuni allows the models to learn the specific rainfall, runoff relationship of this catchment, which differs from the global approximation used in GloFAS. Second, no globally trained model can incorporate a basin-specific upstream signal like the Bagre Dam storage index, which our architecture treats as a first-class input variable contributing an early warning pathway that is entirely absent from GloFAS-based systems.

The strong performance at five-day lead time has direct operational significance. Five days is the planning horizon at which NADMO district offices can realistically mobilise emergency resources, alert community radio stations, and begin the cascade of warnings to Community Disaster Volunteers. A cross-validated Red-tier probability of detection of 0.902 at one-day lead and 0.762 at five days means the system correctly anticipates most evacuation-level events with enough notice to act, even if the five-day figure is lower than the one-day estimate. The corresponding false alarm ratios of 0.134 and 0.268 are within the range that community experience studies suggest does not erode warning credibility when false alarms are communicated honestly and consistently (Demeritt et al. 2010).

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

6.2 The Bagre Signal

The Bagre Dam upstream storage proxy is the most novel element of the feature engineering. Its near-zero linear importance in tree-based models (under 0.1% of Random Forest Gini importance at one-day lead) might appear to undercut the rationale for including it, but this interpretation is misleading for two reasons. First, low linear importance implies only that the variable does not improve short-range predictions once discharge lags are known, not that it is uninformative in general. By the time a major Bagre release reaches Nawuni, it is already encoded in the recent discharge record, making the index genuinely redundant for next-day prediction. At longer leads and at the onset of developing events, before elevated flows arrive at the gauge, the index carries information the discharge model cannot yet access. The seasonal pattern is clear in Figure 12: in all five major validation flood years (2007, 2010, 2016, 2019, 2020) the Bagre index exceeded 0.90 before peak discharge arrival, whereas in non-flood years the index remained below 0.85 throughout the wet season. Second, the Bagre trigger in the operational architecture bypasses the discharge model entirely: when JRC imagery shows reservoir extent approaching its historical maximum, a condition that preceded every major flood in the validation record, the system issues a Yellow alert regardless of forecast model output. This is a deliberately parallel early warning pathway that exploits information unavailable to discharge-based models alone.

6.3 Sentinel-1 Validation and Its Limits

The Sentinel-1 inundation maps confirm that alert threshold exceedances correspond to real floodplain inundation, but the validation has clearly defined limits. The post-2007 period lacks GRDC gauge observations, so the ERA5-Land runoff proxy must substitute for discharge. The proxy achieves good distributional agreement with the GRDC record through quantile mapping, but its day-to-day timing differs from gauge observations in ways that prevent direct lead-time

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

verification. Sentinel-1 observations are therefore used exclusively for spatial validation, confirming the threshold-to-inundation relationship, while lead-time claims rest on the LOYO cross-validated ROC analysis against 7,035 days of GRDC-observed discharge.

The 2019 event discrepancy (52-day gap between DFO-recorded and satellite-detected peak) illustrates a broader challenge: a basin can experience multiple flood pulses within a single wet season, and maximum inundation extent may not correspond to the highest-impact event. Continuous SAR monitoring throughout the flood season, combined with community-level impact records distinguishing between multiple events, would substantially strengthen future validation efforts.

6.4 Limitations and Future Directions

Four limitations warrant explicit acknowledgement. First, the independent test period (2004–2007) does not contain any Orange- or Red-tier discharge events, because the three flood seasons in that window peaked below 1,362 m^3/s. Alert skill for the higher tiers therefore rests on leave-one-year-out cross-validated evidence rather than a strictly held-out sample. The LOYO design eliminates training-data circularity but is not equivalent to a fully independent test set. Obtaining an independent evaluation of Red-tier skill would require either a longer gauge record or the acquisition of discharge data for the post-2007 flood events, particularly 2007, 2010, and 2020.

Second, the training record ends in 1999 and thus excludes the 2007 event, the largest discharge on record at Nawuni (2,279 m^3/s). Models trained without exposure to the most severe historical event may underestimate uncertainty around Red-tier forecasts for truly exceptional floods. Extending training with GloFAS v4 reanalysis discharge for 2000–2023, once GRIB2 reading tools are installed, would address this directly.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

Third, the Bagre Dam signal depends on monthly JRC surface water updates, introducing a lag of several weeks relative to real-time reservoir conditions. Near-real-time Sentinel-1 monitoring or radar altimetry over the Bagre reservoir would transform this from a lagged monthly indicator into a genuine real-time alert, representing the highest-value technical improvement available to the operational system.

Fourth, the GloFAS benchmark comparison is indirect, based on published aggregate statistics from Harrigan et al. (2020) for GloFAS v2.1, rather than a direct evaluation of GloFAS v4 at Nawuni using identical data and metrics. This limits the strength of the performance claims relative to the current operational system. However, several lines of qualitative evidence suggest the direction of the performance gap is robust. The GloFAS v4 hydrological reanalysis, evaluated across 1,995 calibration stations globally, reports a median modified KGE' of 0.70 (Grimaldi et al., 2024; accessible via the GloFAS map viewer at https://globalfloods.eu under Evaluation → Hydrological Model Performance). Stations in West Africa, and particularly those on regulated rivers, tend to cluster at or below this global median because GloFAS does not represent individual reservoir operations such as Bagre Dam releases. The GloFAS v4 LISFLOOD model, while substantially improved over v2.1 in resolution (0.05° versus 0.1°) and calibration coverage, still relies on global parameter sets without explicit dam release rules for Bagre, which is the dominant driver of extreme flood events at Nawuni. Our locally calibrated ensemble, trained directly on GRDC gauge observations and incorporating the Bagre storage proxy as an explicit input, operates in a fundamentally different information regime for this specific basin. A formal head-to-head comparison using GloFAS v4 reanalysis discharge at Nawuni against the same GRDC observations remains the necessary next step and is planned as a companion study.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

7. Conclusions

This paper has presented the first end-to-end AI-driven flood early warning system design for the White Volta Basin of northern Ghana, built entirely from open satellite datasets and validated against independent Sentinel-1 SAR inundation observations. Four conclusions are worth stating directly.

First, the locally calibrated ensemble substantially outperforms the GloFAS v2.1 African benchmark across all forecast horizons. Kling-Gupta Efficiency values of 0.984, 0.974, and 0.957 at one-, three-, and five-day lead times substantially exceed the published GloFAS v2.1 African median of approximately 0.35 (Harrigan et al. 2020), though a direct comparison against GloFAS v4 at Nawuni using identical data and metrics remains a priority for follow-up work. The performance advantage is most pronounced at five-day lead, the horizon that matters most operationally, where global systems trained without local calibration consistently underperform.

Second, Bagre Dam upstream storage is a meaningful early warning signal that current operational systems ignore. The JRC-derived reservoir storage index reached its historical maximum during every major flood event in the validation record, providing a physically motivated alert trigger that can issue community awareness warnings well in advance of floodwaters reaching communities downstream in Ghana. Formalising this signal, through an automated monitoring protocol and, ideally, a bilateral notification arrangement with Burkina Faso's water authorities, represents the highest-value institutional reform available to the existing warning architecture.

Third, the four-tier alert system, calibrated to 30 years of flood frequency data, achieves both high detection rates and operationally tolerable false alarm ratios across the full range of lead times and event severities. Leave-one-year-out cross-validated results show Red-tier alerts achieving a probability of detection of 0.902 at one-day lead and 0.762 at five days, with false alarm ratios of

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

0.134 and 0.268 respectively. These figures provide a quantitative basis for the operational protocol: Yellow and Orange alerts are issued at five-day lead to allow community mobilisation; Red alerts are confirmed at one-day lead when forecast confidence is highest and evacuation decisions must be made. The independent test period (2004–2007) contains no Orange- or Red-tier events, so these figures rest on cross-validated rather than strictly held-out evidence, a limitation acknowledged explicitly and a motivation for acquiring post-2007 gauge data from the Ghana Hydrological Authority.

Fourth, the system as designed requires no new institutional infrastructure at the Ghana Meteorological Agency. Alert output is formatted for direct entry into the myDEWETRA-VOLTALARM platform already operational there, and the dissemination cascade from GMet through NADMO to district offices and Community Disaster Volunteer networks follows existing protocols. The minimum technical footprint is a Python computing environment, automated download scripts for CHIRPS and ERA5, CDS API credentials, and approximately 50 GB of storage, resources that GMet already possesses. The gap between the system described here and an operational deployment is shorter than it might appear.

Data availability

All datasets used in this study are freely available. GRDC discharge data were obtained from the Global Runoff Data Centre (https://grdc.bafg.de). CHIRPS data are available at https://data.chc.ucsb.edu/products/CHIRPS-2.0/. ERA5-Land data are available via the Copernicus Climate Data Store (https://cds.climate.copernicus.eu). JRC Global Surface Water data and Sentinel-1 SAR imagery are available through Google Earth Engine (https://earthengine.google.com). Python scripts for data preprocessing, model training, threshold

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

design, Sentinel-1 processing, and trained model files are archived at Zenodo (https://doi.org/10.5281/zenodo.19342462).

Author Contributions

The following contributions are declared according to the CRediT (Contributor Roles Taxonomy) author contribution statement. Joseph Obeng Junior: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing (original draft), Visualisation. Yuan Hongyong: Supervision, Writing (review and editing), Project administration, Funding acquisition.

Competing Interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors thank the Global Runoff Data Centre (GRDC, Federal Institute of Hydrology, Koblenz, Germany) for providing daily discharge records for the Nawuni gauging station (GRDC station 1531450). CHIRPS precipitation data were provided by the Climate Hazards Group at the University of California Santa Barbara. ERA5-Land reanalysis data and GloFAS reanalysis discharge were obtained through the Copernicus Climate Change Service operated by ECMWF. Sentinel-1 SAR data were provided by the European Space Agency through the Copernicus Earth Observation programme and processed using Google Earth Engine. The JRC Global Surface Water dataset was accessed via Google Earth Engine. The authors acknowledge the Ghana

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

697 Meteorological Agency (GMet), the Ghana Hydrological Authority (GHA), and the National
698 Disaster Management Organisation (NADMO) for institutional context and operational guidance
699 that shaped the system design.

700 References

701 Arsenault, R., Poulin, A., Côté, P., and Brissette, F.: Comparison of stochastic optimization
702 algorithms in hydrological model calibration, J. Hydrol. Eng., 19, 1374–1384,
703 https://doi.org/10.1061/(ASCE)HE.1943-5584.0000938, 2014.

704 Chini, M., Pelich, R., Pulvirenti, L., Pierdicca, N., Hostache, R., and Matgen, P.: Sentinel-1 InSAR
705 coherence to detect floodwater in urban areas: Houston and Hurricane Harvey as a test case,
706 Remote Sens., 11, 107, https://doi.org/10.3390/rs11020107, 2019.

707 Demeritt, D., Cloke, H., Pappenberger, F., Thielen, J., Bartholmes, J., and Ramos, M.-H.:
708 Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting,
709 Environ. Hazards, 7, 115–127, https://doi.org/10.3763/ehaz.2008.0009, 2010.

710 Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J.,
711 Harrison, L., Hoell, A., and Michaelsen, J.: The climate hazards infrared precipitation with
712 stations, a new environmental record for monitoring extremes, Sci. Data, 2, 150066,
713 https://doi.org/10.1038/sdata.2015.66, 2015.

714 Grimaldi, S., Salamon, P., Disperati, J., Zsoter, E., Russo, C., Ramos, A., Carton de Wiart, C.,
715 Barnard, C., Hansford, E., Gomes, G., and Prudhomme, C.: GloFAS v4.0 hydrological
716 reanalysis, European Commission, Joint Research Centre, JRC131349,
717 https://data.jrc.ec.europa.eu/collection/id-00288, 2024.

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

718 Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C.,
719 Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge
720 reanalysis 1979, present, Earth Syst. Sci. Data, 12, 2043–2060,
721 https://doi.org/10.5194/essd-12-2043-2020, 2020.

722 Katsekpor, J. T., Greve, K., Yamba, E. I., and Amoah, E. G.: Flood Early Warning Systems in the
723 White Volta Basin, Ghana: Challenges and Opportunities, J. Flood Risk Manage., e70146,
724 https://doi.org/10.1111/jfr3.70146, 2025a.

725 Katsekpor, J. T., Greve, K., and Yamba, E. I.: Streamflow forecasting using machine learning for
726 flood management and mitigation in the White Volta basin of Ghana, Sustain. Cities Soc.,
727 101, 105000, https://doi.org/10.1016/j.scs.2025.105000, 2025b.

728 Koubodana, H. D. N., Sylla, M. B., Tall, M. S., Dajuma, A., Pal, J. S., Lennard, C. J., Woolway,
729 R. I., and Adeyeri, O. E.: Projected changes in extreme floods over West Africa, Nat.
730 Hazards Earth Syst. Sci., 25, 541–562, https://doi.org/10.5194/nhess-25-541-2025, 2025.

731 Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall-runoff modelling
732 using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–
733 6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

734 Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.:
735 Model evaluation guidelines for systematic quantification of accuracy in watershed
736 simulations, Trans. ASABE, 50, 885–900, https://doi.org/10.13031/2013.23153, 2007.

737 Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G.,
738 Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles,
739 M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-

https://doi.org/10.5194/egusphere-2026-2168
Preprint. Discussion started: 19 May 2026

740 Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data,
741 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

742 Nearing, G. S., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Henck, A., Hurkmans,
743 R., Johnson, D., Matias, M., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C.,
744 Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Kratzert, F.: Global prediction
745 of extreme floods in ungauged watersheds, Nature, 627, 559–563,
746 https://doi.org/10.1038/s41586-024-07145-1, 2024.

747 Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., Waterloo,
748 M., and Saleska, S.: Height Above the Nearest Drainage, a hydrologically relevant new
749 terrain model, J. Hydrol., 404, 13–29, https://doi.org/10.1016/j.jhydrol.2011.03.051, 2011.

750 Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S.: High-resolution mapping of global
751 surface water and its long-term changes, Nature, 540, 418–422,
752 https://doi.org/10.1038/nature20584, 2016.

753 Trigg, M. A., Carr, A., Smith, A., Freer, J., Bates, P., and Neal, J.: The nature of the 2020 white
754 Volta flood: a satellite-observed analysis of inundation extent and timing, Int. J. Disaster
755 Risk Reduct., 72, 102835, https://doi.org/10.1016/j.ijdrr.2022.102835, 2022.

756
757