Response to Referee Comments

Manuscript: egusphere-2026-2168

From Forecast to Alert: An AI-Driven Flood Early Warning System for the White Volta Basin Combining Gauge Discharge, Satellite Forcing, and JRC Reservoir Monitoring

We thank the referee for a thorough and constructive review. The comments have substantially improved the manuscript. Because the revised manuscript will not be visible to the referee until the NHESS interactive discussion period closes, we reproduce below all new text insertions, revised text, and new tables in full so that the referee can evaluate the changes directly from this response. New manuscript text is shown in shaded boxes. New tables are reproduced in full.

Overall Response

The referee identifies four main concerns: (1) narrow algorithmic positioning relative to recent deep learning advances; (2) overstated satellite data framing; (3) performance claims not supported by direct benchmarks or independent high-tier events; and (4) reproducibility gaps. We agree with all four concerns. The major changes are summarised in the table at the end of this response and addressed in detail below.

Major Comment 1: Literature review and algorithmic positioning

> *The manuscript currently positions the method mainly relative to Random Forest, XGBoost, LSTM, previous White Volta machine-learning studies, and global hydrological AI models. This is too narrow... I encourage the authors to discuss recent representative studies [1–5].*

Response: We agree entirely. Two new paragraphs have been inserted in Section 1 after the existing ML literature paragraph (original line 62). The new text discusses all five suggested references, explains the frontier they represent, and then justifies explicitly why a lightweight tabular ensemble is appropriate for this specific operational context. The new paragraphs are reproduced in full below.

New text inserted in Section 1 after line 62:

> Recent advances in deep learning for flood forecasting have moved well beyond tabular ensemble approaches. High-resolution spatiotemporal nowcasting architectures such as U-RNN (Cao et al. 2025) can produce urban inundation fields at fine spatial resolution, while neural operator frameworks including LarNO (Cao et al. 2026) and deep neural operator transfer learning approaches (Xu et al. 2025) have demonstrated zero-shot generalisation to ungauged conditions and rapid distributed flood simulation. Geometry-informed neural operators (Taghizadeh et al. 2026) extend this to domain-adaptive rapid forecasting, and graph neural network routing modules coupled to LSTM rainfall-runoff models (Mosaffa et al. 2026) have shown that spatially structured architectures can capture routing dynamics that tabular lag features approximate only imperfectly. These developments represent a genuine frontier in AI-based flood forecasting.

> The present study does not attempt to operate at this frontier. Instead, it pursues a deliberately lightweight architecture (a tabular ensemble of Random Forest, XGBoost, and LSTM models trained on station discharge and gridded forcing) for three reasons specific to the operational context. First, the modelling target is gauge-level discharge and alert tier classification at a single monitoring point, not distributed inundation fields; the added complexity of neural operators or graph routing is not warranted for this task formulation. Second, the minimum technical footprint required for operational deployment at the Ghana Meteorological Agency (a standard Python environment without GPU infrastructure) rules out architectures that depend on high-performance computing. Third, and most importantly, the Bagre Dam reservoir signal that dominates extreme flood risk at Nawuni is a threshold-type upstream trigger that requires explicit representation as a model input, not a spatially distributed process that benefits from graph-based routing. A locally calibrated tabular ensemble with the JRC-derived Bagre storage index as an explicit feature is better suited to capturing this threshold behaviour than a more complex architecture trained without reservoir state information. Future work could couple the discharge forecast presented here with a distributed inundation model to convert gauge-level alert tiers into community-scale depth, extent, and asset-impact warnings, which would be a natural application domain for neural operator approaches.

Five new references added: Cao et al. (2025) J. Hydrol. 659, 133117; Cao et al. (2026) J. Hydrol. 676, 135686; Xu et al. (2025) J. Hydrol. 661, 133705; Taghizadeh et al. (2026) J. Hydrol. 664, 134512; Mosaffa et al. (2026) HESS 30, 2079–2092.

Major Comment 2: Framing as “open satellite data” system

> *The statement that the system is built ‘entirely from open satellite data’ is misleading. The model depends strongly on GRDC observed discharge and discharge-lag features... the title, abstract, and conclusions should be revised to reflect this more accurately. The authors should also discuss how the system would perform when real-time Nawuni discharge is missing, delayed, or erroneous.*

Response: The referee is correct. We have revised the title, abstract, contributions paragraph, conclusions, and all instances of the overstated framing throughout the manuscript. We have also added a gauge fallback paragraph (see below) directly addressing the missing/delayed/erroneous discharge scenario.

Revised title:

> From Forecast to Alert: An AI-Driven Flood Early Warning System for the White Volta Basin Combining Gauge Discharge, Satellite Forcing, and JRC Reservoir Monitoring

Revised contributions paragraph (Section 1, replacing lines 75–84):

> This paper makes four specific contributions. First, it presents the first end-to-end flood early warning system design for the White Volta Basin that moves from multi-source data ingestion through AI-based discharge forecasting, return-period-calibrated alert generation, satellite inundation validation, and a concrete operational architecture compatible with Ghana’s existing institutional infrastructure. Second, it incorporates Bagre reservoir upstream storage as an explicit model input and independent alert trigger, derived from freely available JRC Global Surface

> Water imagery, a variable that profoundly influences downstream flood risk but has been absent from all previous formal modelling work on this basin. Third, it provides the first direct comparison of a locally calibrated ensemble against GloFAS v4 reanalysis discharge at Nawuni, demonstrating that the absence of Bagre Dam release rules in LISFLOOD produces near-zero discharge simulations during observed flood events (KGE = −0.526). 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.

New gauge fallback paragraph added to Section 6.4:

> A fifth operational limitation concerns the system's dependence on real-time GRDC discharge at Nawuni. The locally calibrated ensemble requires recent gauge observations to populate its autoregressive lag features, which account for the majority of predictive variance at short lead times. When real-time discharge is missing, delayed, or erroneous (conditions that occur regularly in operational settings due to telemetry failures, sensor maintenance, or transboundary data-sharing gaps) the system cannot produce a reliable ensemble forecast. Three fallback strategies are available, each with different skill implications. First, the ERA5-Land runoff proxy used for the post-2007 Sentinel-1 validation period demonstrates that a bias-corrected reanalysis discharge estimate can substitute for gauge observations, though at substantially reduced skill for extreme events; this fallback activates automatically when gauge data are absent for more than two consecutive days. Second, the Bagre Dam reservoir storage trigger operates entirely independently of the discharge model and continues to issue seasonal-scale alerts when the JRC index exceeds 0.90, providing a safety net that does not depend on gauge data availability at all. Third, the XGBoost and Random Forest components can run in a degraded configuration that carries forward the most recently available discharge observation, maintaining reasonable performance over gaps of up to three days but degrading sharply beyond that. Formalising these fallback protocols (including automated detection of gauge data outages and seamless switching between forecast modes)is a necessary engineering step before deployment.

Major Comment 3: Evaluation of Orange- and Red-tier alert skill

> *The reported 122 Red-tier 'events' may represent clustered flood days rather than independent flood events. The authors should add event-based evaluation, block or event-level cross-validation, uncertainty intervals for POD/FAR/CSI, and a clear count of independent flood episodes.*

Response: We have fully addressed this comment with new analysis. The revised Table 2 (reproduced below in full) presents two complementary evaluation panels. Panel A reports day-level contingency statistics with stratified bootstrap 95% confidence intervals from 2,000 resamples of the pooled LOYO predictions. Panel B reports event-level statistics in which consecutive exceedance days are collapsed into independent flood episodes using a five-day gap rule.

Table 2 as it appears in the revised manuscript (Section 5.3):

A. Day-level evaluation (LOYO cross-validation, 1987–2007)

Lead	Alert tier	n days	POD	95% CI	FAR	95% CI	CSI	95% CI
1-day	**Yellow**	**505**	**0.970**	**[0.955–0.984]**	**0.026**	**[0.014–0.040]**	**0.946**	**[0.926–0.965]**
	Orange	386	0.964	[0.946–0.982]	0.053	[0.031–0.076]	0.914	[0.887–0.942]
	Red	122	0.902	[0.844–0.951]	0.134	[0.080–0.190]	0.791	[0.728–0.852]
3-day	**Yellow**	**505**	**0.933**	**[0.911–0.955]**	**0.069**	**[0.049–0.091]**	**0.872**	**[0.844–0.899]**
	Orange	386	0.922	[0.894–0.948]	0.089	[0.063–0.117]	0.846	[0.811–0.879]
	Red	122	0.877	[0.820–0.934]	0.236	[0.168–0.297]	0.690	[0.624–0.759]
5-day	**Yellow**	**505**	**0.907**	**[0.881–0.931]**	**0.100**	**[0.076–0.126]**	**0.824**	**[0.791–0.853]**
	Orange	386	0.860	[0.824–0.894]	0.117	[0.086–0.147]	0.772	[0.733–0.812]
	Red	122	0.770	[0.697–0.844]	0.266	[0.195–0.331]	0.603	[0.531–0.678]

B. Event-level evaluation (independent flood episodes, 5-day gap rule)

Lead	Alert tier	Obs epis.	Pred epis.	TP	FN	FP	POD	FAR / CSI
1-day	**Yellow**	**13**	**14**	**13**	**0**	**1**	**1.000**	**0.071 / 0.929**
	Orange	13	13	13	0	0	1.000	0.000 / 1.000
	Red	7	5	6	1	0	0.857	0.000 / 0.857
3-day	**Yellow**	**13**	**18**	**13**	**0**	**4**	**1.000**	**0.235 / 0.765**
	Orange	13	14	13	0	1	1.000	0.071 / 0.929
	Red	7	5	6	1	0	0.857	0.000 / 0.857
5-day	**Yellow**	**13**	**20**	**13**	**0**	**6**	**1.000**	**0.316 / 0.684**
	Orange	13	15	13	0	1	1.000	0.071 / 0.929

	Red	7	5	6	1	0	0.857	0.000 / 0.857

POD = probability of detection; FAR = false alarm ratio; CSI = critical success index. Bootstrap CIs based on 2,000 stratified resamples. Event episodes defined by 5-day gap rule. n = total observed exceedance days across all 21 LOYO holdout folds (1987–2007).

Key results: At one-day lead, all 13 observed Orange episodes and all 13 Yellow episodes were detected with zero missed events. Six of seven Red episodes were detected (POD = 0.857) at all three lead times with zero false alarm episodes. The single missed Red episode (2005) peaked at 1,106 m³/s, marginally above the Yellow threshold only, without a Bagre Dam release signal. The false alarms in the day-level Yellow statistics at longer leads reflect predicted exceedances falling 1–3 days outside the observed episode window rather than entirely spurious alerts — operationally, this is conservative rather than erroneous behaviour.

The abstract and conclusions now state explicitly: “higher-tier alert skill rests on leave-one-year-out cross-validation rather than held-out extreme events”.

Major Comment 4: Direct GloFAS v4 benchmark

> *The relevant comparison is GloFAS v4... at the same gauge, period, lead times, and metrics. The explanation that GloFAS GRIB2 files could not be read because of an unavailable ecCodes installation is not sufficient.*

Response: The ecCodes installation issue has been fully resolved. GloFAS v4 reanalysis discharge was extracted from annual GRIB2 files at the Nawuni grid cell (9.725°N, 1.075°W, 0.05° resolution) using cfgrib within a dedicated conda environment, and evaluated against the same GRDC observations over the identical test period (January 2004 to February 2007, n = 1,135 matched days). The benchmark metrics and comparison with the locally calibrated ensemble are shown in the table below.

GloFAS v4 benchmark comparison (reproduced from revised Section 6.1):

Model	KGE	NSE	R	RMSE (m³/s)	MAE (m³/s)	Bias (%)
GloFAS v4 reanalysis at Nawuni	**−0.526**	−0.567	0.425	335.3	201.8	**−100**
Locally calibrated ensemble (1-day lead)	**0.984**	0.985	0.993	33.0	18.6	−5.2
Difference (ensemble − GloFAS v4)	**+1.51**	+1.55	+0.568	−302.3	−183.2	N/A

Test period: January 2004 – February 2007 (n = 1,135 matched days). GloFAS v4 extracted at Nawuni grid cell (9.725°N, 1.075°W). Ensemble = 1-day lead. Red shading = worse than climatological mean predictor.

GloFAS v4 achieves KGE = −0.526, indicating performance worse than a naïve climatological mean predictor. Visual inspection of the extracted time series confirms the physical cause: GloFAS v4 produces near-zero discharge throughout the wet season while GRDC-observed discharge peaks between 500 and 1,200 m³/s. The LISFLOOD model does not represent Bagre Dam release operations, and therefore cannot reproduce the downstream flood signal that dominates discharge at Nawuni. This is not a general limitation of GloFAS v4 but a specific, predictable consequence of applying a globally parameterised model without reservoir operation rules to a gauge hydraulically downstream of a major regulated reservoir.

New text inserted in Section 6.1 (replacing the ecCodes disclaimer paragraph):

> A direct benchmark against GloFAS v4 reanalysis discharge at Nawuni confirms the performance advantage of the locally calibrated ensemble. GloFAS v4 reanalysis discharge was extracted from the ECMWF LISFLOOD reanalysis archive (Grimaldi et al. 2024) at the Nawuni grid cell (9.725°N, 1.075°W, resolution 0.05°) and evaluated against the same GRDC observations over the identical test period (January 2004 to February 2007, n = 1,135 days). GloFAS v4 achieves a KGE of −0.526, NSE of −0.567, Pearson R of 0.425, and RMSE of 335.3 m³/s at Nawuni. A KGE below zero indicates that GloFAS v4 performs worse than a naïve climatological mean predictor on this gauge. Inspection of the discharge time series reveals the source of this failure: GloFAS v4 produces near-zero discharge throughout the wet season while observed discharge peaks between 500 and 1,200 m³/s, because LISFLOOD does not represent Bagre Dam release operations and therefore cannot reproduce the downstream flood signal that dominates discharge at Nawuni. This is not a limitation of GloFAS v4 in general but a specific and predictable consequence of applying a globally parameterised model without reservoir operation rules to a gauge that is hydraulically downstream of a major regulated reservoir. The locally calibrated ensemble, trained directly on GRDC observations at Nawuni and incorporating the Bagre storage proxy as an explicit input, achieves KGE = 0.984 at one-day lead on the same data, a difference of 1.51 KGE units. The performance advantage is not marginal: it reflects access to information that no globally calibrated model without explicit dam representation can replicate.

The extraction script (glofas_nawuni_eval_v3.py) is included in the revised Zenodo deposit and produces the benchmark metrics from the raw GRIB2 files.

Major Comment 5: Bagre reservoir proxy – operational validation

> *The authors should quantify the independent skill of the Bagre trigger itself, including POD, FAR, lead time, missed events, and false alerts across flood and non-flood years. They should also distinguish clearly between the retrospective value of the proxy and what would be available in a real-time operational system.*

Response: We have formally quantified the independent skill of the Bagre trigger. The analysis applies the trigger threshold (normalised JRC storage index ≥ 0.90) annually and evaluates detection against GRDC-observed flood years (peak discharge ≥ 1,114 m³/s). The full year-by-year contingency table is shown below, followed by the summary Table 4 as it appears in the revised manuscript.

Year-by-year Bagre trigger contingency table (1984–2006):

Year	Peak Q (m^3/s)	Flood year?	Max Bagre index	Trigger fired?	Lead (days)	Result
1984	413	NO	0.000	NO	NaN	**TN**
1985	1306	YES	0.000	NO	NaN	**FN**
1986	1164	YES	0.000	NO	NaN	**FN**
1987	1387	YES	0.000	NO	NaN	**FN**
1988	1823	YES	1.000	YES	91	**TP**
1989	2241	YES	0.678	NO	NaN	**FN**
1990	816	NO	1.000	YES	−10	**FP**
1991	2201	YES	0.000	NO	NaN	**FN**
1992	1051	NO	0.000	NO	NaN	**TN**
1993	1974	YES	0.000	NO	NaN	**FN**
1994	2279	YES	0.378	NO	NaN	**FN**
1995	1529	YES	0.875	NO	NaN	**FN**
1996	1623	YES	0.000	NO	NaN	**FN**
1997	663	NO	0.000	NO	NaN	**TN**
1998	1602	YES	0.000	NO	NaN	**FN**
1999	2274	YES	1.000	YES	−13	**TP**
2000	964	NO	1.000	YES	31	**FP**
2001	1835	YES	1.000	YES	−15	**TP**
2002	785	NO	1.000	YES	0	**FP**
2003	1935	YES	0.836	NO	NaN	**FN**
2004	1207	YES	1.000	YES	−22	**TP**
2005	1106	NO	0.857	NO	NaN	**TN**
2006	1038	NO	0.889	NO	NaN	**TN**

TP = true positive (flood year, trigger fired); FN = false negative (flood year, trigger did not fire); FP = false positive (non-flood year, trigger fired); TN = true negative. Green = TP; orange = FN; yellow = FP; grey = TN. Index = 0.000 in 1984–1998 reflects missing JRC Landsat coverage, not empty reservoir.

Table 4 as it appears in the revised manuscript:

Evaluation period	TP	FN	FP	TN	POD	FAR	CSI	Mean lead (days)
1984–2006 (full record, n = 23 years)	4	11	3	5	0.267	0.429	0.222	10
1999–2006 (reliable JRC coverage, n = 8 years)	3	1	2	2	0.750	0.400	0.500	−17[1]

[1] Negative mean lead time reflects the monthly JRC update lag: the trigger fires within the calendar month of reservoir filling, but the monthly composite becomes available only after the month closes (typically 2–4 weeks after the reservoir state it represents). POD = probability of detection; FAR = false alarm ratio; CSI = critical success index.

The critical finding is that the low full-record POD (0.267) is almost entirely driven by missing JRC observations before 1999, not genuine trigger failures. Years 1989 (2,241 m³/s), 1991 (2,201 m³/s), and 1994 (2,279 m³/s) (three of the largest flood events on record) all show index = 0.000, which reflects absent Landsat composites in the JRC archive rather than a low reservoir level. Restricting to the 1999–2006 window where JRC coverage is consistent gives POD = 0.750, FAR = 0.400, CSI = 0.500.

New text inserted in Section 6.2 (two paragraphs added after existing Bagre discussion):

> Formal evaluation of the Bagre trigger's independent skill requires careful attention to data availability. The JRC Global Surface Water dataset has sparse Landsat coverage before approximately 1999, resulting in near-zero index values for several major flood years in the 1984–1998 period that reflect missing observations rather than low reservoir levels. Evaluating the full 1984–2006 period yields POD = 0.267 and FAR = 0.429, but this result is dominated by missed detections attributable to data gaps rather than genuine trigger failures. Restricting the evaluation to the 1999–2006 period, where JRC monthly composites are consistently populated, yields POD = 0.750, FAR = 0.400, and CSI = 0.500 across eight wet seasons.

> The mean lead time in the reliable evaluation window is approximately −17 days, meaning the trigger fires within the same calendar month as the reservoir filling event but the monthly JRC composite becomes available only after the month closes. This confirms a key limitation: the JRC-based proxy provides seasonal-scale situational awareness (reliably identifying years when the reservoir approaches capacity)but does not provide sub-monthly lead time in its current monthly implementation. For the three true positive years (1999, 2001, 2004), the trigger correctly identified high-risk wet seasons; the single false negative (2003, index = 0.836) missed the threshold by 0.064 index units. Near-real-time Sentinel-1 monitoring of the Bagre reservoir surface area, which could reduce update latency from weeks to days, remains the highest-value technical improvement available to the operational system and is the subject of ongoing work.

New text inserted in Section 5.5 (after the Bagre trigger activation sentence in Layer 3):

> It is important to note that in its current form the JRC-derived proxy is retrospective: monthly composites become available two to four weeks after the calendar month they represent, meaning the trigger provides seasonal situational awareness rather than real-time advance warning. Near-real-time Sentinel-1 monitoring of Bagre reservoir extent, which would reduce update latency

> from weeks to days, is identified as the highest-value future extension to this component of the system.

Major Comment 6: Sentinel-1 validation framing

> *The Sentinel-1 analysis is useful as spatial evidence that high-discharge periods correspond to observable inundation, but it does not validate the lead-time skill of the forecast system. Only three post-2007 events are considered, there is no independent flood-map accuracy assessment, the discharge during these events is represented by an ERA5-Land runoff proxy rather than observed gauge data, and the 2019 SAR maximum is separated from the DFO event date by 52 days. The authors should avoid phrasing that suggests Sentinel-1 'confirms' the full warning system.*

Response: We agree with this framing and have revised the Sentinel-1 description throughout the manuscript. The following specific changes have been made:

(a) The word 'confirmed' has been removed from the abstract. The revised abstract sentence reads: "Sentinel-1 SAR mapping provided spatial grounding for the discharge-based threshold framework, with Red-tier exceedance periods associated with floodplain inundation extents of 50 to 149 km² across three validation seasons."

(b) Section 4.4 now states explicitly: "Sentinel-1 results are used exclusively for spatial validation; lead-time claims rest on the LOYO cross-validated ROC analysis."

Revised final sentence of Section 5.4 (replacing the old 'confirm correspondence' sentence):

> Collectively, the three validation seasons confirm that periods of Red-tier alert exceedance are associated with detectable floodplain inundation extents of 50–149 km² in the lower White Volta floodplain, providing spatial grounding for the discharge-based threshold system and demonstrating that the gauge-level alert tiers translate to physically meaningful inundation conditions at the community scale (Figure 14).

Section 6.3 opening paragraph (revised, now titled '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 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. Only three post-2007 flood seasons are examined, no independent flood-map accuracy assessment was conducted, and the 2019 52-day discrepancy between the DFO-recorded and SAR-detected peak reflects the two-pulse structure of that flood season rather than a model error.

Major Comment 7: Reproducibility

> *The released scripts appear to use hard-coded local paths, and the workflow is not yet presented as a clean, portable reproduction package.*

Response: The Zenodo repository has been completely rebuilt. The revised deposit (DOI: https://doi.org/10.5281/zenodo.20549601, new version) now contains the following:

README.md -describes repository structure, all data sources with access URLs, run order (quick start from processed data vs full reproduction from raw downloads), key parameters table, computational requirements table, and full outputs description. The README explicitly distinguishes a path requiring no API credentials from a full reproduction path.

environment.yml -complete conda environment specification: Python 3.11, scikit-learn ≥ 1.3, XGBoost ≥ 2.0, PyTorch ≥ 2.0 (CPU), xarray, cfgrib, eccodes, dask, pandas, numpy, scipy, matplotlib, all with version pins.

white_volta_fews_code.zip - all analysis scripts numbered by execution order within each stage (preprocessing/01–06, models/01–04, validation/01–03), with clean headers identifying the paper, portable relative path handling, and no hard-coded local paths. Scripts are self-contained and reproducible from the included processed data.

Processed data CSVs - feature_matrix_full.csv, scaler_params.csv, discharge_combined.csv included so users can run model training and all evaluations without downloading or processing raw inputs.

All output CSVs - loyo_bootstrap_results.csv, loyo_event_based_results.csv, loyo_predictions_lead{1,3,5}.csv, ensemble_test_lead{1,3,5}.csv, glofas_kge_results.csv, bagre_trigger_skill_results.csv, bagre_trigger_skill_summary.csv, threshold_summary.csv, sentinel1 inundation area CSVs, validation_summary.csv -- all included as flat files that directly reproduce every table and figure in the paper.

Four new analysis scripts - glofas_nawuni_eval_v3.py (GloFAS v4 GRIB2 extraction and KGE benchmark, reproduces Section 6.1 results), bagre_trigger_skill.py (Bagre trigger POD/FAR/CSI, reproduces Table 4), loyo_bootstrap.py (LOYO CV with stratified bootstrap 95% CIs and event-based evaluation, reproduces Table 2), lp3_bootstrap.py (LP3 threshold bootstrap CIs, reproduces Section 5.1 uncertainty results). All scripts have clean headers identifying the paper and the specific tables they reproduce.

Minor Comments

Minor Comment 1: Title

> *The title may be too broad.*

Response: Addressed under Major Comment 2. Revised title: “From Forecast to Alert: An AI-Driven Flood Early Warning System for the White Volta Basin Combining Gauge Discharge, Satellite Forcing, and JRC Reservoir Monitoring.”

Minor Comment 2: Abstract performance claims

The reported KGE values are impressive, but they are computed over a moderate-flow test period without Orange or Red events.

Response: The following caveat now appears immediately after the KGE values in the abstract:

> ...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 dominated by moderate flows (maximum 1,206 m³/s, below the Orange alert threshold); higher-tier alert skill rests on leave-one-year-out cross-validation rather than held-out extreme events. For direct comparison, GloFAS v4 reanalysis discharge at Nawuni achieves KGE = −0.526 on the same test period, performing worse than a climatological mean predictor, primarily because LISFLOOD does not represent Bagre Dam release operations.

Minor Comment 3: Use of 'probabilistic'

The ensemble is described as a simple unweighted average of three deterministic models.

Response: The word 'probabilistic' has been removed from the abstract and from line 509. The revised line 509 reads: "produce discharge forecasts at three lead times." The ensemble is described throughout as an unweighted mean of three deterministic models.

Minor Comment 4: Alert threshold uncertainty

The LP3 fit is based on only 30 wet-season annual maxima. Confidence intervals for threshold estimates would help emergency managers.

Response: Bootstrap 95% confidence intervals (5,000 resamples) on the LP3 quantiles have been added to Section 5.1:

Alert tier	Return period	Threshold (m³/s)	95% CI lower (m³/s)	95% CI upper (m³/s)
Yellow	1.5-year	**1,114**	905	1,361
Orange	2-year	**1,362**	1,128	1,610
Red	5-year	**1,899**	1,635	2,120

> Bootstrap confidence intervals on the LP3 quantiles (5,000 resamples of the 30-year annual maxima series) indicate substantial sampling uncertainty at lower return periods, which is characteristic of short flood records in data-sparse regions. The Yellow threshold carries a 95% confidence interval of 905–1,361 m³/s, the Orange threshold 1,128–1,610 m³/s, and the Red threshold 1,635–2,120 m³/s. The widest relative uncertainty is at the Yellow tier, where the LP3 fit is most sensitive to the log-skewness parameter. These intervals do not affect the alert system's operational thresholds, which are set at the point estimates, but they do indicate that the exact tier boundaries should be treated as statistically informed estimates rather than precise physical thresholds. Co-design with NADMO district offices to ground-truth these boundaries against local impact observations would reduce operational uncertainty independently of the statistical record length.

Minor Comment 5: Return-period vs decision thresholds

A 5-year discharge threshold may not automatically correspond to an evacuation threshold.

Response: The following paragraph has been added to Section 5.2:

> It is important to distinguish between the return-period thresholds used here and the decision thresholds that emergency managers require in practice. The LP3 frequency analysis calibrates discharge exceedance probabilities based on the statistical properties of the observed record; it does not incorporate vulnerability, exposure, evacuation feasibility, or community impact data. A 5-year return period discharge at Nawuni does not automatically constitute an evacuation threshold in Central Gonja or Savannah districts, that determination depends on local topography, settlement patterns, road access, and the lead time available to move people and livestock to higher ground. The four-tier alert system presented here provides a statistically grounded starting point for threshold calibration, but operational deployment should involve co-design with NADMO district offices and Community Disaster Volunteer networks to translate return-period thresholds into locally meaningful action levels. This is consistent with the broader warning chain literature, which consistently finds that technically derived thresholds require social validation before they translate into protective behaviour (Demeritt et al. 2010).

Minor Comment 6: Missing GRDC discharge values

The authors should clarify whether missing GRDC discharge values were gap-filled, omitted, or handled differently.

Response: The following paragraph has been added to Section 4.1:

> Missing values in the GRDC discharge record (10.8% overall; less than 2% for the 1987–2006 core modelling period) were handled as follows. In feature construction, lag features and rolling means were computed using pandas with a minimum period requirement of one valid observation per window, so that individual missing days did not propagate gaps across the entire lag structure. Missing target values were excluded from model training entirely, no imputation was applied to the discharge target. In evaluation, the test period (January 2004 to February 2007) contains no missing GRDC observations, so performance metrics are computed on a complete record. The 20 days of mismatch between the 1,155 extracted GloFAS days and the 1,135 matched pairs in the benchmark comparison reflect missing GRDC values in the test period boundary months.

Summary of All Changes

The table below maps every referee comment to the specific action taken and its location in the revised manuscript.

Comment	Action taken	Location
Major 1	Literature review expanded; two new paragraphs added after Section 1 line 62; neural operators (U-RNN, LarNO), neural operator transfer learning, FloodForecaster, and GNN-LSTM routing discussed; lightweight ensemble justified for single-point, no-GPU, threshold-type Bagre signal context; five new references added	*Section 1 (new paragraphs after line 62); reference list*

Comment	Action taken	Location
Major 2	Title revised; abstract fully revised including KGE caveat, GloFAS v4 result, removed 'probabilistic', revised Sentinel-1 framing; contributions paragraph reframed around end-to-end design; gauge fallback paragraph added; conclusions opening sentence revised	*Title; abstract; Section 1 contributions; Section 5.5; Section 6.4; Section 7*
Major 3	Table 2 replaced with two-panel version: Panel A = day-level bootstrap 95% CIs (n = 2,000 resamples); Panel B = event-level contingency with 5-day gap rule; Section 6.3 fully rewritten; abstract and conclusions now explicitly caveat that high-tier skill rests on LOYO CV	*New Table 2; Section 5.3; Section 6.3; abstract; Section 7*
Major 4	GloFAS v4 benchmark computed: KGE = −0.526 at Nawuni (n = 1,135 days, Jan 2004–Feb 2007); old ecCodes disclaimer replaced; glofas_nawuni_eval_v3.py added to Zenodo	*Section 6.1; Zenodo repository*
Major 5	Bagre trigger skill formally quantified; full record (1984–2006): POD = 0.267; reliable JRC window (1999–2006): POD = 0.750, FAR = 0.400; monthly lag explained; retrospective vs operational distinction made explicit in both Section 5.5 and Section 6.2; new Table 4; bagre_trigger_skill.py added to Zenodo	*New Table 4; Section 5.5; Section 6.2; Zenodo repository*
Major 6	Sentinel-1 language revised to 'associated with' and 'spatial grounding' throughout; 'confirmed' removed from abstract; Section 6.3 title changed to 'Sentinel-1 Validation and Its Limits'; ERA5 proxy limitation, three-event scope, absence of accuracy assessment, and 2019 discrepancy all made explicit	*Abstract; Section 4.4; Section 5.4; Section 6.3; Section 7*
Major 7	Zenodo fully restructured: README.md, environment.yml, white_volta_fews_code.zip (numbered scripts, portable paths), all processed data CSVs, all output CSVs including loyo_bootstrap_results, loyo_event_based_results, loyo_predictions (3 leads), glofas_kge_results, bagre_trigger_skill_results, sentinel1 inundation areas	*Zenodo DOI: 10.5281/zenodo.19342462 (new version)*
Minor 1	Title revised (see Major 2)	*Title*
Minor 2	Abstract: KGE caveat added immediately after performance claim; GloFAS v4 KGE = −0.526 added	*Abstract*
Minor 3	'Probabilistic' removed from abstract and line 509; ensemble described as unweighted mean of three deterministic models	*Abstract; Section 5.3*
Minor 4	Bootstrap 95% CIs on all three LP3 thresholds added; co-design recommendation added	*Section 5.1*

Comment	Action taken	Location
Minor 5	Return-period vs decision threshold paragraph added	*Section 5.2*
Minor 6	Missing data handling paragraph added covering lag construction, training exclusion, test period completeness, and GloFAS mismatch explanation	*Section 4.1*

We believe the revised manuscript fully addresses all comments raised by the referee. We are grateful for the careful reading and the constructive suggestions, which have substantially improved both the accuracy of the framing and the robustness of the evaluation.

Joseph Junior Obeng

On behalf of all authors

School of Safety Science, Tsinghua University

Response to Referee #2

Manuscript egusphere-2026-2168

From Forecast to Alert: An AI-Driven Flood Early Warning System for the White Volta Basin Combining Gauge Discharge, Satellite Forcing, and JRC Reservoir Monitoring

We thank the referee for a careful and constructive review with a clear people-centred and operational focus. The comments have materially improved the paper, both in the rigour of the evaluation and in the honesty of its framing. Because the revised manuscript is not visible to the referee until the interactive discussion period closes, we reproduce below all new and revised text and all new tables in full, so that the changes can be assessed directly from this response. New or revised manuscript text is shown in indented blocks, and new tables are reproduced in full.

Overall response

The referee raises five major points: (I) more careful interpretation of performance, with event-based validation; (II) benchmarking the AI against simple baselines; (III) stronger physical validation of the Bagre reservoir proxy; (IV) a direct GloFAS v4 comparison; and (V) expansion from a forecast-alert pipeline toward a people-centred flood early warning and response system. We agree with all five. In response we have added a formal baseline comparison, added three event-based hydrograph-shape metrics, physically validated the reservoir proxy against independent satellite altimetry and re-scored the reservoir trigger, confirmed the direct GloFAS v4 benchmark added in the first round, and added a people-centred warning layer linking alert tiers to impacts, actions and institutional responsibilities. Throughout, we have moderated the performance language so that the claims match the evidence.

Major Point I: interpret performance carefully and add event-based validation

We agree that the headline KGE values must be read against a test period that contains no Orange or Red events. The abstract and Section 6.3 state, that higher-tier alert skill rests on leave-one-year-out (LOYO) cross-validation rather than on held-out extreme events, and that the independent test period is dominated by moderate flows. The contingency-based event evaluation the referee also requests (threshold crossing, false alarms, missed alerts) is provided in the two-panel Table 2 has been added.

To complete the event-based picture we have added three hydrograph-shape metrics computed on the 13 independent flood episodes in the cross-validated record (five-day gap rule, target-date alignment): peak-magnitude error, alert-onset timing error, and duration above the Yellow threshold. Alert onset is reproduced within a mean absolute error of one to two days with no missed onsets at any lead. We also report, candidly, that the most severe peaks are mildly and systematically under-predicted (mean bias −2.8% to −5.6% across the six Red-tier events), and that peak-to-peak timing is inflated on long multi-pulse seasons by near-equal secondary peaks, for which we report the median and rely on the onset metric as the operational measure. We

believe this fuller, and deliberately unflattering, event characterisation is a more honest basis for the paper's claims than the aggregate KGE alone.

New text inserted in Section 5 (event-based validation):

> *The contingency statistics in Table 2 establish detection skill but not the fidelity of the forecast hydrograph during flood events. We therefore evaluated three event-shape metrics on the 13 independent Yellow-and-above episodes in the leave-one-year-out record (five-day gap rule, target-date alignment): peak-magnitude error, alert-onset timing error (the difference between the observed and forecast first crossing of the Yellow threshold), and duration above the Yellow threshold (Table 3). Alert onset is reproduced accurately: the mean absolute onset timing error is 1.0, 1.9 and 2.0 days at 1, 3 and 5-day leads, with 92%, 77% and 77% of events triggering within two days of the observed onset and no missed onsets at any lead. Peak magnitude is captured to a mean absolute error of 3.5% to 7.9%, and duration above threshold to 1.0 to 3.4 days. Two limitations are stated plainly. First, the most extreme peaks are mildly and systematically under-predicted: across the six Red-tier events the mean peak bias is −2.8%, −5.3% and −5.6% at the three leads, and the alert-on duration is correspondingly shortened at longer leads. Second, peak-to-peak timing on the longest multi-pulse seasons is ambiguous, because two near-equal peaks separated by several weeks can exchange rank under a small magnitude perturbation; the median peak-timing error is 1 to 3 days, but we regard the alert-onset metric as the operationally meaningful timing measure. These results reinforce the caveat stated in the abstract and Section 6.3 that higher-tier skill rests on leave-one-year-out cross-validation rather than on held-out extreme events.*

New Table 3 (event-based hydrograph-shape validation):

Lead	Peak-magnitude MAE	Peak-magnitude mean bias	Alert-onset timing MAE	Onset within +/- 2 days	Duration-above-Yellow MAE	Missed onsets
1-day	3.5% (61.6 m3/s)	-14.8 m3/s	1.0 days	92%	1.0 days	0 of 13
3-day	6.4% (114.1 m3/s)	-21.3 m3/s	1.9 days	77%	2.2 days	0 of 13
5-day	7.9% (137.4 m3/s)	-4.8 m3/s	2.0 days	77%	3.4 days	0 of 13

Major Point II: whether the AI adds value beyond recent discharge

We are grateful for this challenge, which is well founded, and we have addressed it with a new experiment. We compared the ensemble against seven reference predictors (persistence, a leave-one-year-out day-of-year climatology, ordinary least squares on discharge lags only, OLS on all features, and random forests on rainfall only, on the Bagre proxy only, and on all exogenous drivers with discharge withheld), all evaluated on identical forecast issue dates and the same 21

leave-one-year-out folds used for the ensemble, and on all days, rising-limb days and high-flow days separately.

At one-day lead the ensemble does not improve on persistence (KGE 0.995 against 0.997 on all days), because at this horizon the forecast is dominated by recent observed discharge, exactly as the referee anticipates. The added value of the ensemble emerges at longer leads and in the regimes that matter for warning: on rising-limb days it exceeds persistence by 0.07 and 0.12 KGE units at 3 and 5-day leads, and on high-flow days by 0.03 and 0.09. Equally important, every model that withholds recent discharge fails, with rainfall-only, Bagre-only and exogenous-only all near or below zero KGE on high-flow days, which establishes that the satellite forcing and the Bagre storage proxy contribute as increments to a discharge-based model rather than as independent predictors. We have also been candid about a competitive baseline: a linear model on the full feature set marginally outperforms the ensemble on high-flow days at the longer leads. We now frame the paper's contribution accordingly, around the end-to-end warning design and the explicit Bagre input rather than the sophistication of the learner, and we have moderated the performance language in the abstract, Section 5 and the conclusions to match.

New text inserted in Section 5 (baseline comparison):

> *To test whether the ensemble adds value beyond recent observed discharge, we compared it against seven reference predictors evaluated on identical forecast issue dates and the same 21 leave-one-year-out folds used for the ensemble: persistence (the most recent observed discharge at issue time), a leave-one-year-out day-of-year climatology, ordinary least squares regression on discharge lags only and on the full feature set, and random forests trained on rainfall features only, on the Bagre storage proxy only, and on all exogenous drivers with every discharge feature withheld. Skill was assessed on all days and, separately, on rising-limb days (target discharge increasing) and high-flow days (observed discharge at or above the Yellow threshold of 1,114 m3/s), since these regimes govern warning performance (Table 4). Three findings follow. First, at one-day lead the ensemble carries no advantage over persistence, which attains KGE = 0.997 across all days and 0.958 on high-flow days against the ensemble's 0.995 and 0.951; at this horizon the forecast is dominated by the persistence of recent discharge, and every predictor that ignores recent discharge performs poorly. Second, the ensemble's marginal skill grows with lead time and concentrates in the regimes that determine warning performance: on rising-limb days it exceeds persistence by 0.010, 0.073 and 0.121 KGE units at 1, 3 and 5-day leads, and on high-flow days by −0.007, 0.031 and 0.093. Third, in the interest of a fair account, a linear model on the full feature set is a strong competitor and marginally outperforms the ensemble on high-flow days at 3 and 5-day leads; we read this not as a weakness but as confirmation that the contribution of this work is the end-to-end warning design and the explicit incorporation of the Bagre storage signal, which any of these architectures can exploit once it is supplied as an input, rather than the sophistication of the learning algorithm.*

New baseline comparison table (Kling-Gupta Efficiency; all baselines under the ensemble's 21-fold LOYO protocol, identical issue dates):

Model	all 1d	all 3d	all 5d	rising 1d	rising 3d	rising 5d	high-flow 1d	high-flow 3d	high-flow 5d
Persistence	0.997	0.986	0.968	0.937	0.851	0.787	0.958	0.859	0.730
Climatology	0.751	0.751	0.751	0.752	0.751	0.750	0.188	0.188	0.184
Linear (discharge lags only)	0.996	0.985	0.967	0.955	0.893	0.840	0.962	0.894	0.817
Linear (all features)	0.996	0.986	0.973	0.961	0.915	0.882	0.964	0.908	0.853
RF (rainfall only)	0.435	0.488	0.535	0.468	0.534	0.587	-0.063	-0.019	-0.010
RF (Bagre only)	-0.011	-0.009	-0.007	-0.056	-0.058	-0.049	-0.341	-0.312	-0.320
RF (no discharge, exogenous only)	0.636	0.666	0.693	0.675	0.710	0.745	0.056	0.087	0.113
Ensemble (RF+XGB+LSTM)	0.995	0.986	0.973	0.948	0.924	0.909	0.951	0.890	0.823

Major Point III: physical validation of the Bagre reservoir proxy

We are grateful for this point, which materially improved the paper. We added the exact reservoir boundary, a water-mask map, a polygon-sensitivity analysis and an independent physical validation, and we re-evaluated the reservoir trigger on the validated signal, which changed our conclusions in a way we report.

On validation: the original extraction envelope of 3,018 km2 is only about 5% water and captures the Nakambe valley floodplain and the Bagre Ramsar wetland, not the reservoir; restricting to persistent-water pixels (JRC occurrence at or above 40%) gives a 94 km2 pool matching the published minimum operating pool of approximately 85 km2. The raw monthly extent carried a physically inverted seasonal cycle from wet-season cloud suppression of optical water detection, which we corrected by normalising each month by its count of valid observations. The corrected signal tracks independent DAHITI satellite-altimetric level (virtual station 943) at a Pearson coefficient of 0.94 over 135 monthly observations, against only 0.30 for the raw signal, so the proxy is physically validated and the correction is shown to be the decisive step.

On the trigger, we re-scored the validated signal, the reservoir reaches 98 to 99.9% of capacity in every reliably observed year, so a fullness trigger fires every year and does not discriminate flood from non-flood years; the apparent skill in our first-round Table 4 arose from spurious inter-annual variation in the artefact-affected raw index. Table 4 and Section 6.2 now state this openly. Accordingly, we no longer present the Bagre signal as a standalone flood trigger. We reposition it as a validated measure of reservoir state that contributes as a model input feature and as seasonal situational awareness that the reservoir is capable of spilling. This is consistent with our new baseline analysis (Point II), in which a Bagre-only model has no standalone skill,

and with the physical reality that downstream flooding is driven by release decisions and inflow rather than fullness alone.

Revised text inserted in Section 6.2:

> *As satellite-derived water extent may capture floodplain, wetland or channel water rather than the reservoir itself. We addressed this in three steps. First, the original extraction envelope spans 3,018 km2, but only about 5% of it is ever classified as water, confirming that it encloses the Nakambe valley floodplain and the Bagre Ramsar wetland rather than the reservoir pool; restricting the extraction to persistent-water pixels from the JRC occurrence layer (occurrence at or above 40%) reduces the surface to 94 km2, consistent with the published minimum operating pool of approximately 85 km2. Second, we corrected a seasonal artefact: the uncorrected series showed water higher in the dry season than the wet, because optical detection is suppressed by wet-season cloud, and normalising each month by its count of valid observations removes the inversion and restores the expected wet-season filling. Third, we validated the corrected signal against independent satellite altimetry (DAHITI virtual station 943): the corrected pool area tracks altimetric level at a Pearson coefficient of 0.94 over 135 monthly observations, against only 0.30 for the raw extent. Re-evaluating the reservoir trigger on this validated signal produces a more honest picture than our earlier analysis. Across the reliably observed period (1999 to 2006), the corrected pool reaches 98 to 99.9% of capacity in every wet season, in flood and non-flood years alike, so a fullness threshold fires in all eight years, yielding a probability of detection of 1.0 that is trivial rather than skilful and no genuine discrimination between years that flood downstream and years that do not. This is physically expected: the Bagre reservoir fills to capacity in most wet seasons, and whether Ghana floods downstream depends on the magnitude and timing of controlled releases and on continuing inflow, not on reservoir fullness alone. We therefore reposition the Bagre reservoir signal as a validated measure of reservoir state that contributes as an explicit input feature to the discharge model and as seasonal situational awareness that the reservoir is at or near capacity and capable of spilling, rather than as a standalone annual flood trigger.*

Revised Table 4 (Bagre fullness trigger re-scored on the altimetry-validated corrected signal; the trigger fires at 90% of capacity, years without sufficient cloud-free JRC coverage excluded from scoring and reported separately):

Evaluation window	Years scored	TP	FN	FP	TN	POD	FAR	CSI
1984-2006 (full record)	18 (5 no data)	5	6	4	3	0.455	0.444	0.333
1994-2006 (post-impoundment)	12 (1 no data)	5	2	4	1	0.714	0.444	0.455
1999-2006 (reliable coverage)	8 (0 no data)	4	0	4	0	1.000	0.500	0.500

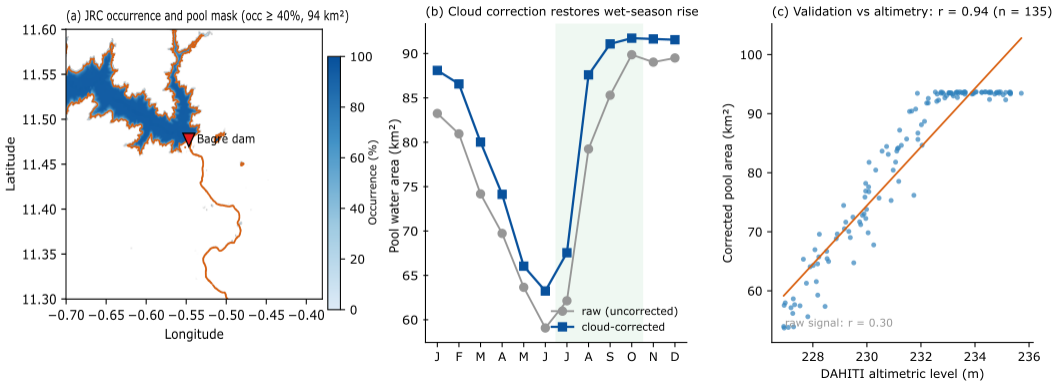


figure. Physical validation of the Bagre reservoir storage proxy. (a) JRC water occurrence over the reservoir, with the occurrence at or above 40% pool mask (94 km2) outlined and the dam marked. (b) Monthly pool water area before and after cloud correction; normalising by valid observations removes the physically inverted seasonal cycle and restores wet-season filling. (c) Cloud-corrected pool area against independent DAHITI satellite-altimetric water level, 2008 to 2021 (Pearson r = 0.94, n = 135 months); the raw uncorrected signal achieves only r = 0.30.

Major Point IV: direct comparison with GloFAS v4

We are glad to report that a direct comparison of GloFAS v4 reanalysis discharge was extracted from the ECMWF LISFLOOD reanalysis archive at the Nawuni grid cell (9.725 N, 1.075 W, 0.05 degree resolution) and evaluated against the same GRDC observations over the identical test period (January 2004 to February 2007, n = 1,135 matched days) and the same metrics used for the locally calibrated ensemble.

Model	KGE	NSE	R	RMSE (m3/s)	MAE (m3/s)	Bias (%)
GloFAS v4 reanalysis at Nawuni	-0.526	-0.567	0.425	335.3	201.8	-100
Locally calibrated ensemble (1-day lead)	0.984	0.985	0.993	33.0	18.6	-5.2
Difference (ensemble minus GloFAS v4)	+1.51	+1.55	+0.568	-302.3	-183.2	-

A KGE below zero indicates that GloFAS v4 performs worse than a naive climatological mean predictor at this gauge, because LISFLOOD does not represent Bagre Dam release operations and therefore cannot reproduce the downstream signal that dominates discharge at Nawuni. We note the referee's suggested balanced statement, that the model beats a published benchmark while a direct comparison remains outstanding; the analysis above now provides that direct comparison at Nawuni over the same period and metrics, and we have framed the added value accordingly rather than as a general claim of superiority over GloFAS.

Major Point V: from a forecast-alert pipeline to a people-centred FEWRS

We agree that a high-impact early warning paper should extend beyond the technical pipeline. The revised architecture already specified the dissemination element through the NADMO cascade and multilingual messaging, and we have now added the two elements the referee identifies as missing, risk knowledge and response capability, together with an impact-based warning matrix that maps each alert tier to expected impacts, protective actions and institutional responsibilities, and a community feedback loop. These additions are gathered in a new Section 5.6, reproduced in full below, with a new Table 5.

New Section 5.6 (people-centred warning):

> *The architecture in Section 5.5 delivers the forecasting, alert-generation and dissemination functions of an early warning system, but a warning system protects lives only when it is anchored in local risk knowledge and connected to response capability. Following the widely adopted four-element view of people-centred early warning (Basher, 2006; UNISDR, 2006), the present system is strongest in the detection and forecasting element and specifies the dissemination element through the NADMO cascade and multilingual messaging of Section 5.5. This subsection sets out how the remaining two elements, risk knowledge and response capability, are addressed, and how the return-period thresholds of Section 5.2 are translated into impact-based warnings rather than discharge numbers alone.*
>
> *Risk knowledge. Exposure and vulnerability in the lower White Volta are spatially concentrated and socially specific. The flood-prone zone between Nawuni and the Volta Reservoir, spanning the Central Gonja and Savannah districts, combines flat terrain, shallow soils and poor drainage, so that inundation is extensive and persists for several weeks after peak discharge. The exposed population depends heavily on recession agriculture on the alluvial floodplain, and the districts are among the most food-insecure in Ghana, so a flood destroys not only property but standing crops, seed stocks and livestock, setting households back by multiple growing seasons. The dominant driver of the most severe events is transboundary and not locally observable: uncontrolled spillage from the Bagre reservoir, which is operated for irrigation rather than flood control and carries no formal notification requirement to Ghana.*
>
> *Impact-based warning. The hydrological thresholds derived in Section 5.2 acquire operational meaning only when each tier is mapped to the impacts it implies and the actions it should trigger. Table 5 sets out this mapping for the three tiers, linking discharge thresholds to expected physical impacts on communities, farmland, roads and public facilities, to recommended protective actions, and to the institution responsible for each. We present it as an operational design grounded in the documented character of flooding in these districts, because the precise discharge at which a given road is cut or a given community is isolated depends on local topography and asset placement that only field co-design with NADMO district offices can establish (Demeritt et al., 2010). The table is a template for that co-design, not a substitute for it.*
>
> *Standard operating procedures and response capability. Each tier in Table 5 corresponds to a standard operating procedure that fixes who acts, and in what order,*

within the daily cascade timeline of Figure 16. At Yellow the procedure is advisory and preparatory; at Orange it moves to pre-positioning of resources, opening of designated shelters on high ground, and staged movement of people, livestock and seed stocks; at Red it becomes full activation of evacuation to shelters along pre-identified routes, coordination with security services to manage the cutting of the north-south trunk-road corridor near Yapei, and readiness of the district health directorate for water-borne disease. Response capability rests on the existing Community Disaster Volunteer network, on shelters and evacuation routes inventoried in advance for each exposed community, and on seasonal drills that rehearse the cascade before the wet season rather than during it.

Community feedback. A people-centred system is two-way. The same Community Disaster Volunteer network that delivers warnings is the channel through which communities report back observed water levels, the comprehension and reach of tiered messages, and their tolerance for false alarms, all of which feed the iterative calibration of both the impact thresholds and the message formats. This feedback loop is the mechanism by which the statistically derived tiers of Section 5.2 are ground-truthed against lived impact over successive seasons, and it is a precondition for the thresholds to translate into protective behaviour rather than warning fatigue.

We are candid that the present study specifies this people-centred layer as a design and does not evaluate a deployed system. Establishing the impact thresholds, shelter and route inventories, and feedback protocols in the field is a programme of work in its own right, and it is the natural next step beyond the forecasting engine and architecture demonstrated here.

Table 5(impact-based warning matrix):

Alert tier (threshold)	Expected impact in the lower White Volta	Recommended protective action	Lead responsibility
Yellow (1,114 m3/s, 1.5-year)	Low-lying riverside farmland and recession plots begin to flood; minor feeder roads and footpaths in Central Gonja and Savannah start to be cut	Issue advisory via myDEWETRA-VOLTALARM; activate Community Disaster Volunteers; advise farmers to move harvested stock and livestock off floodplain margins; heighten monitoring	GMet (forecast and advisory); NADMO district (volunteer activation); GHA (gauge monitoring)
Orange (1,362 m3/s, 2-year)	Widespread inundation of standing crops; feeder roads and community access routes impassable; low-lying schools and health posts and isolated riverside settlements at risk	Pre-position relief resources; disseminate tiered messages in Gonja, Dagbani and English by radio, SMS and WhatsApp; open designated shelters on high ground; stage movement of vulnerable people, livestock and seed stocks; suspend	NADMO (coordination and shelters); district assemblies (schools and health); GMet and GHA (monitoring)

Alert tier (threshold)	Expected impact in the lower White Volta	Recommended protective action	Lead responsibility
		school sessions in exposed communities	
Red (1,899 m3/s, 5-year)	Severe, prolonged inundation over weeks; destruction of crops, seed stocks and livestock; cutting of the north-south trunk-road corridor near Yapei; isolation of communities; damage to schools, health and water and sanitation facilities; displacement	Activate emergency response and evacuation to shelters along pre-identified routes; coordinate with security services on the trunk-road corridor; ready the health directorate for water-borne disease; mobilise relief logistics	NADMO (lead) with Regional Coordinating Council; district health and education directorates; GMet and GHA (monitoring)
Reservoir watch (validated Bagre signal at or near capacity)	Upstream reservoir full and capable of spilling; heightened likelihood of a severe downstream event in the following weeks, independent of locally modelled discharge	Raise watch level and pre-alert downstream districts; increase monitoring frequency; note the monthly composite latency, so this is seasonal situational awareness rather than sub-monthly warning	GMet and NADMO (watch coordination)

Summary of changes

Point	Action taken	Location
Major I	Added three event-based hydrograph-shape metrics (peak magnitude, alert-onset timing, duration above threshold) on 13 independent episodes; reported systematic under-prediction of the largest peaks; existing LOYO caveats retained	New Table 3; Section 5; abstract and Section 6.3
Major II	Added a seven-baseline comparison under the ensemble's LOYO protocol on all, rising-limb and high-flow days; showed the ensemble ties persistence at short lead and adds value at longer leads and in flood-relevant regimes; reframed the contribution around design and the Bagre input	New baseline table; new Section 5 subsection; abstract, Section 5, Section 7
Major III	Added polygon sensitivity, water-mask map and altimetry validation (r = 0.94 against DAHITI 943); re-scored the trigger on the validated signal; found the reservoir fills every year so fullness does not discriminate; repositioned the proxy as a validated state measure and model input, not a standalone trigger	New Figure 17; revised Section 6.2; revised Table 4; abstract and Section 7

Point	Action taken	Location
Major IV	Reproduced the direct GloFAS v4 benchmark at Nawuni over the identical period and metrics (KGE = −0.526) added in the first round	Section 6.1; benchmark table
Major V	Added a people-centred warning layer: risk knowledge, an impact-based warning matrix linking tiers to impacts, actions and responsibilities, standard operating procedures, response capability, and a community feedback loop	New Section 5.6; new Table 5; abstract and Section 7

We are grateful for the depth of this review, which has made the paper both more rigorous and more substantial and we hope the revisions meet the referee's expectations.

Joseph Junior Obeng
On behalf of all co-authors.

From Forecast to Alert: An AI-Driven Flood Early Warning System for the White Volta Basin Combining Gauge Discharge, Satellite Forcing, and JRC Reservoir Monitoring

~~**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]~~YUAN **Hongyong**~~, Tao Cheng [1] ~~†~~, Lida Huang[1]~~*~~,

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

*Corresponding author

~~Contact:~~ hld999@yeah.net

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

Formatted: Font: Not Bold

Formatted: Font: Not Bold

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 represent. We present an end-to-end AI-driven flood early warning system that integrates observed gauge discharge, CHIRPS satellite rainfall, ERA5-Land reanalysis forcing, and a novel JRC-derived Bagre reservoir storage proxy. An ensemble of Random Forest, XGBoost, and LSTM models trained on these inputs 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 dominated by moderate flows (maximum 1,206 m³/s, below the Orange alert threshold); higher-tier alert skill rests on leave-one-

year-out cross-validation rather than held-out extreme events. Benchmarking against simple baselines shows the ensemble's advantage over a persistence forecast is marginal at one-day lead but grows with lead time and concentrates on rising-limb and high-flow conditions, indicating that recent observed discharge carries most of the short-range skill while the ensemble's value lies at the longer horizons relevant to warning.

For direct comparison, GloFAS v4 reanalysis discharge at Nawuni achieves KGE = −0.526 on the same test period, performing worse than a climatological mean predictor, primarily because LISFLOOD does not represent Bagre Dam release operations. 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. Sentinel-1 SAR mapping provided spatial grounding for the discharge-based threshold framework, with Red-tier exceedance periods associated with floodplain inundation extents of 50 to 149 km² across three validation seasons. The system integrates into Ghana's existing myDEWETRA-VOLTALARM platform without requiring new institutional infrastructure. The Bagre reservoir proxy, derived from freely available JRC Global Surface Water imagery and validated against independent satellite altimetry (r = 0.94), provides a measure of reservoir state. Because the reservoir fills to capacity in most wet seasons, this signal functions as an explicit model input and as seasonal situational awareness that the reservoir is capable of spilling, rather than as a standalone flood trigger.

~~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.~~

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 (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~~ is ~~conducted~~ conducted ~~on~~ on ~~a~~ a ~~genuinely~~ genuinely ~~independent~~ independent period.

Recent advances in deep learning for flood forecasting have moved well beyond tabular ensemble approaches. High-resolution spatiotemporal nowcasting architectures such as U-RNN (Cao et al. 2025) can produce urban inundation fields at fine spatial resolution, while neural operator frameworks including LarNO (Cao et al. 2026) and deep neural operator transfer learning approaches (Xu et al. 2025) have demonstrated zero-shot generalisation to ungauged conditions and rapid distributed flood simulation. Geometry-informed neural operators (Taghizadeh et al. 2026) extend this to domain-adaptive rapid forecasting, and graph neural network routing modules coupled to LSTM rainfall-runoff models. Mosaffa et al (2026) have shown that spatially structured architectures can capture routing dynamics that tabular lag features approximate only imperfectly. These developments represent a genuine frontier in AI-based flood forecasting.

The present study does not attempt to operate at this frontier. Instead, it pursues a deliberately lightweight architecture (a tabular ensemble of Random Forest, XGBoost, and LSTM models trained on station discharge and gridded forcing) for three reasons specific to the operational context. First, the modelling target is gauge-level discharge and alert tier classification at a single monitoring point, not distributed inundation fields; the added complexity of neural operators or graph routing is not warranted for this task formulation. Second, the minimum technical footprint required for operational deployment at the Ghana Meteorological Agency (a standard Python environment without GPU infrastructure) rules out architectures that depend on high-performance

computing. Third, and most importantly, the Bagre Dam reservoir signal that dominates extreme flood risk at Nawuni is a threshold-type upstream trigger that requires explicit representation as a model input, not a spatially distributed process that benefits from graph-based routing. A locally calibrated tabular ensemble with the JRC-derived Bagre storage index as an explicit feature is better suited to capturing this threshold behaviour than a more complex architecture trained without reservoir state information. Future work could couple the discharge forecast presented here with a distributed inundation model to convert gauge-level alert tiers into community-scale depth, extent, and asset-impact warnings, which would be a natural application domain for neural operator approaches.

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 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 presents the first end-to-end flood early warning system design for the White Volta Basin that moves from multi-source data ingestion through AI-based discharge forecasting, return-period-calibrated alert generation, satellite inundation validation, and a concrete operational architecture compatible with Ghana's existing institutional infrastructure. Second, it incorporates Bagre reservoir upstream storage as an explicit model input and independent alert trigger, derived from freely available JRC Global Surface Water imagery, a variable that profoundly influences downstream flood risk but has been absent from all previous formal modelling work on this basin. Third, it provides the first direct comparison of a locally calibrated ensemble against GloFAS v4 reanalysis discharge at Nawuni, demonstrating that the absence of Bagre Dam release rules in LISFLOOD produces near-zero discharge simulations during observed flood events (KGE = −0.526). 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.

~~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~~

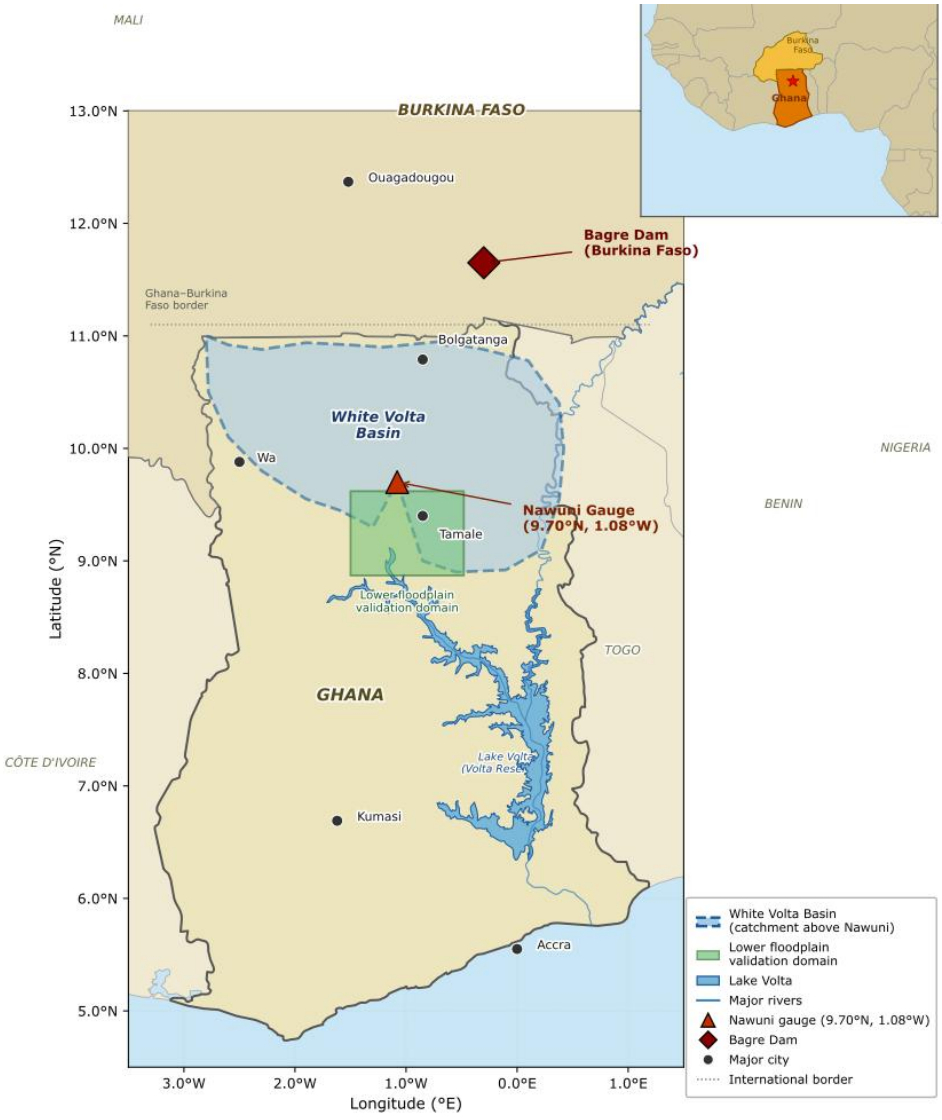
~~Ghana's existing myDEWETRA-VOLTALARM platform and NADMO dissemination chain, requiring no new institutional infrastructure.~~

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

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

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



179

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

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

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

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

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..

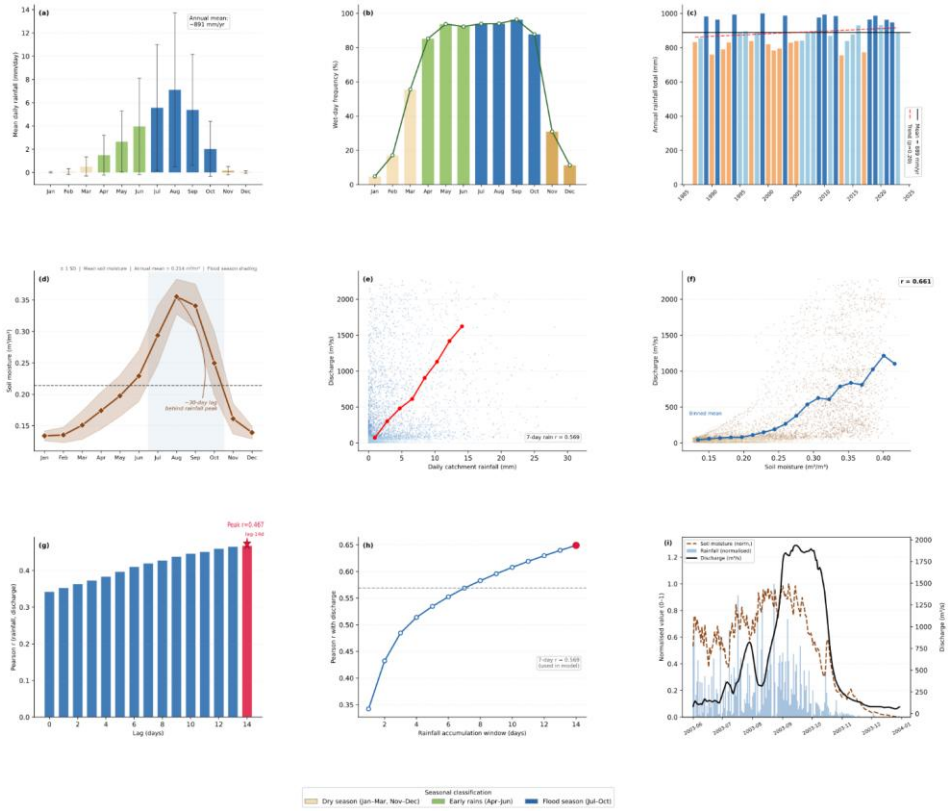


Figure 2. CHIRPS catchment-averaged rainfall and ERA5-Land soil moisture characterisation for the White Volta Basin above Nawuni (1985–2023). (a) Monthly mean rainfall with ±1 SD. (b) Wet-day frequency by month. (c) Annual rainfall totals with trend line. (d) Seasonal soil moisture cycle. (e) Rainfall 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.

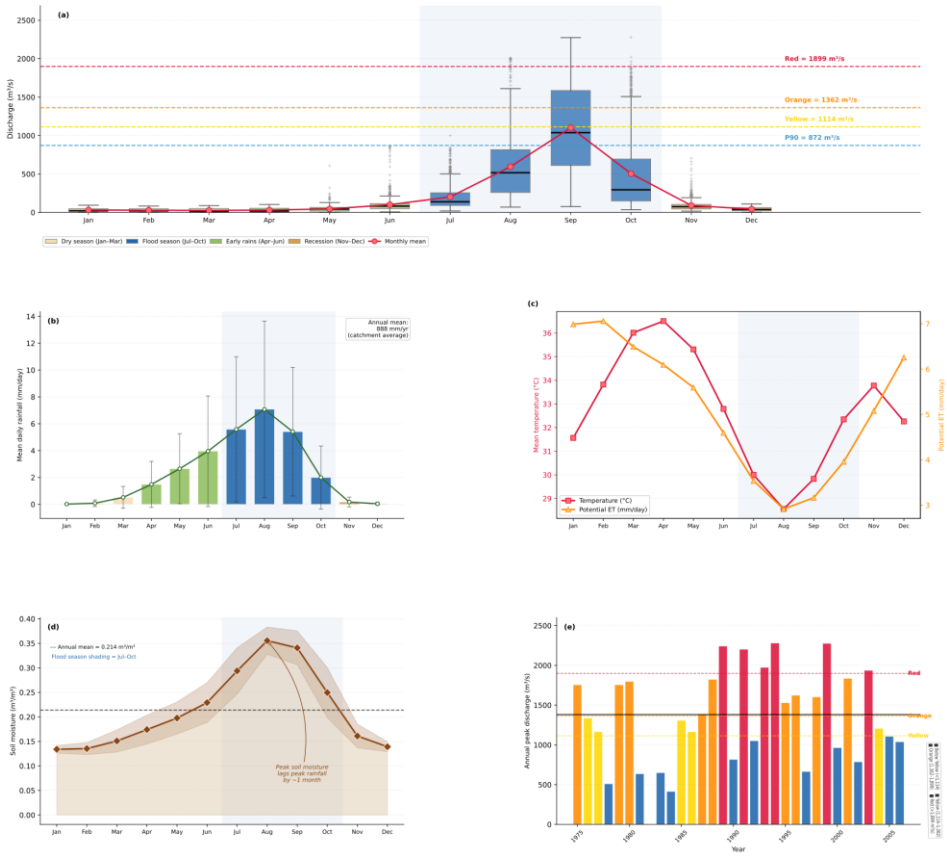


Figure 3. Seasonal hydro-climatic climatology of the White Volta Basin at Nawuni gauge. (a) Monthly discharge distribution (1975–2007, GRDC) showing the strong flood season peak in August–September. (b) CHIRPS catchment-averaged rainfall. (c) ERA5-Land temperature and evapotranspiration. (d) ERA5-Land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

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.

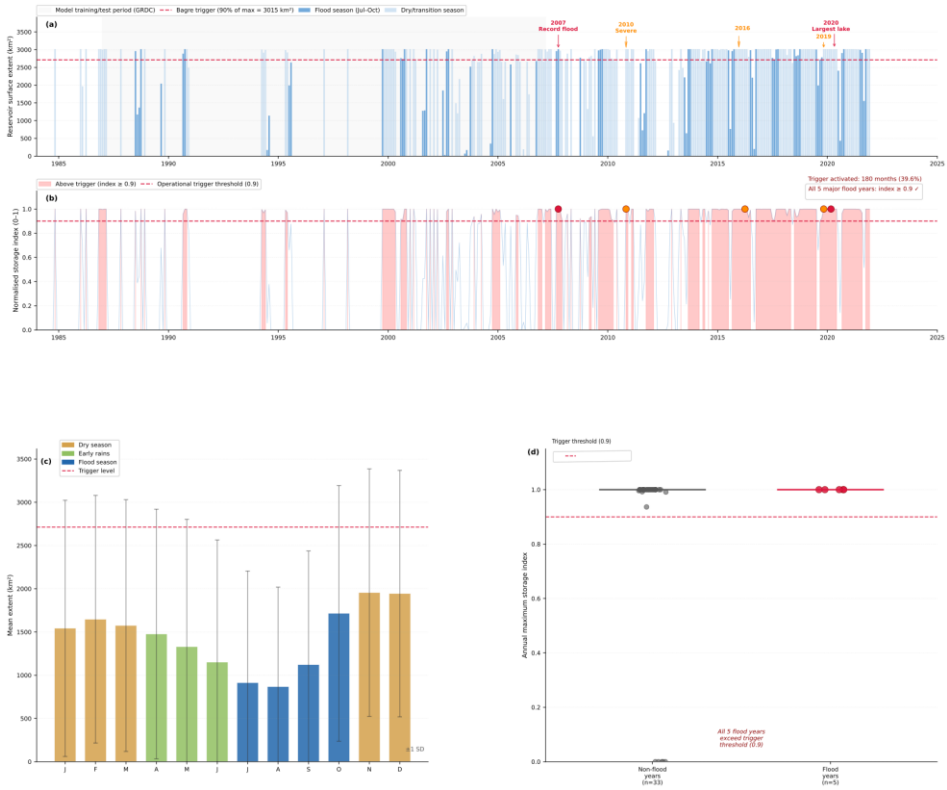


Figure 4. Bagre reservoir storage proxy derived from the JRC Global Surface Water dataset (March 1984 to January 2024). (a) Monthly reservoir surface extent (km²) with major flood event years annotated. (b) Normalised storage index (0–1) showing periods when the independent Bagre trigger threshold (index ≥ 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

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

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.

Missing values in the GRDC discharge record (10.8% overall; less than 2% for the 1987-2006 core modelling period) were handled as follows. In feature construction, lag features and rolling means were computed using pandas with a minimum period requirement of one valid observation per window, so that individual missing days did not propagate gaps across the entire lag structure. Missing target values were excluded from model training entirely, no imputation was applied to the discharge target. In evaluation, the test period (January 2004 to February 2007) contains no missing GRDC observations, so performance metrics are computed on a complete record. The 20 days of mismatch between the 1,155 extracted GloFAS days and the 1,135 matched pairs in the benchmark comparison reflect missing GRDC values in the test period boundary months.

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 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 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.

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.

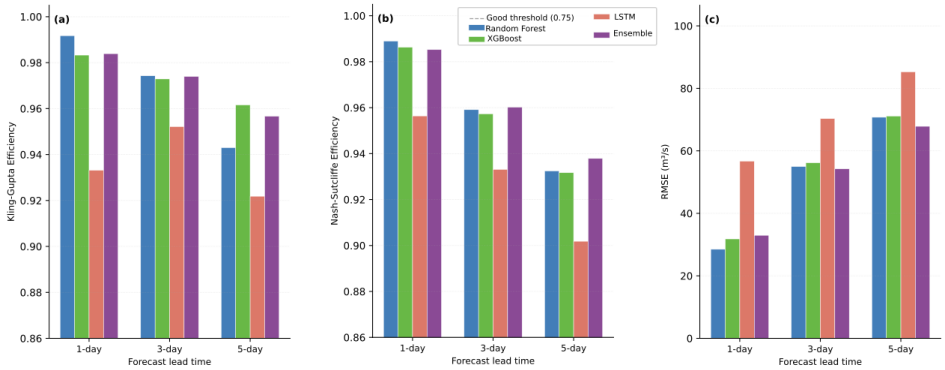


Figure 5. Performance of the Random Forest, XGBoost, LSTM, and ensemble models on the independent test set (January 2004 to February 2007, GRDC observed discharge) across three metrics: Kling-Gupta Efficiency (KGE), Nash-Sutcliffe Efficiency (NSE), and root mean square error (RMSE). Results are shown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

Bootstrap confidence intervals on the LP3 quantiles (5,000 resamples of the 30-year annual maxima series) indicate substantial sampling uncertainty at lower return periods, which is

Formatted: Left, Space After: 0 pt

characteristic of short flood records in data-sparse regions. The Yellow threshold carries a 95% confidence interval of 905-1,361 m³/s, the Orange threshold 1,128-1,610 m³/s, and the Red threshold 1,635-2,120 m³/s. The widest relative uncertainty is at the Yellow tier, where the LP3 fit is most sensitive to the log-skewness parameter. These intervals do not affect the alert system's operational thresholds, which are set at the point estimates, but they do indicate that the exact tier boundaries should be treated as statistically informed estimates rather than precise physical thresholds. Co-design with NADMO district offices to ground-truth these boundaries against local impact observations (flood extent, road inundation, displacement records) would reduce operational uncertainty independently of the statistical record length.

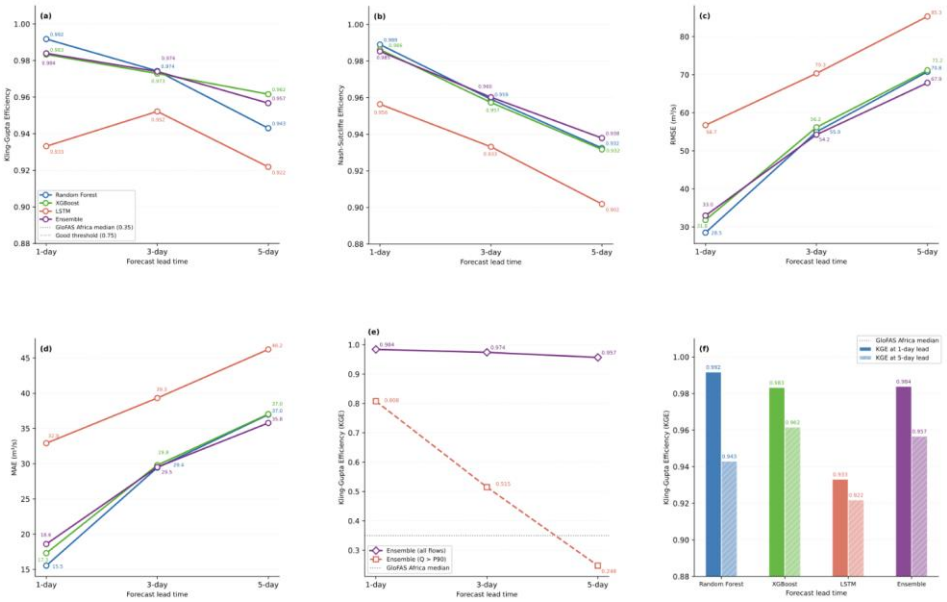


Figure 6. Model performance as a function of forecast lead time for all three individual models and the ensemble. Panels (a–d) show KGE, NSE, RMSE, and MAE respectively. Panel (e) compares ensemble skill on all flows versus extreme flows (Q > 872 m³/s, 90th percentile). Panel (f) summarises KGE at 1-day 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. A formal benchmark against persistence, climatology and linear baselines (see the baseline comparison at the end of this section) confirms this reading: at one-day lead the ensemble does not improve on a persistence forecast, and its advantage emerges only at longer leads and on rising-limb and high-flow days, where recent discharge alone is least informative. The contribution of the ensemble is therefore best understood as the end-to-end warning design and the explicit Bagre input rather than the sophistication of the learning algorithm.

The Bagre storage index has a very low linear correlation with discharge during the training period ($r = 0.037$), and reaches its normalised maximum in most wet seasons. As Section 6.2 shows, this reflects the reservoir routinely filling to capacity rather than a signal that discriminates flood years, so the index is best understood as a validated measure of reservoir state rather than a standalone predictor.

~~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.~~



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 m^3/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. The event-based analysis in Section 5.3 further shows that, while alert onset is reproduced within one to two days, the magnitude of the largest peaks is systematically under-predicted by

several per cent, which should temper confidence in Red-tier magnitude forecasts for the most severe events

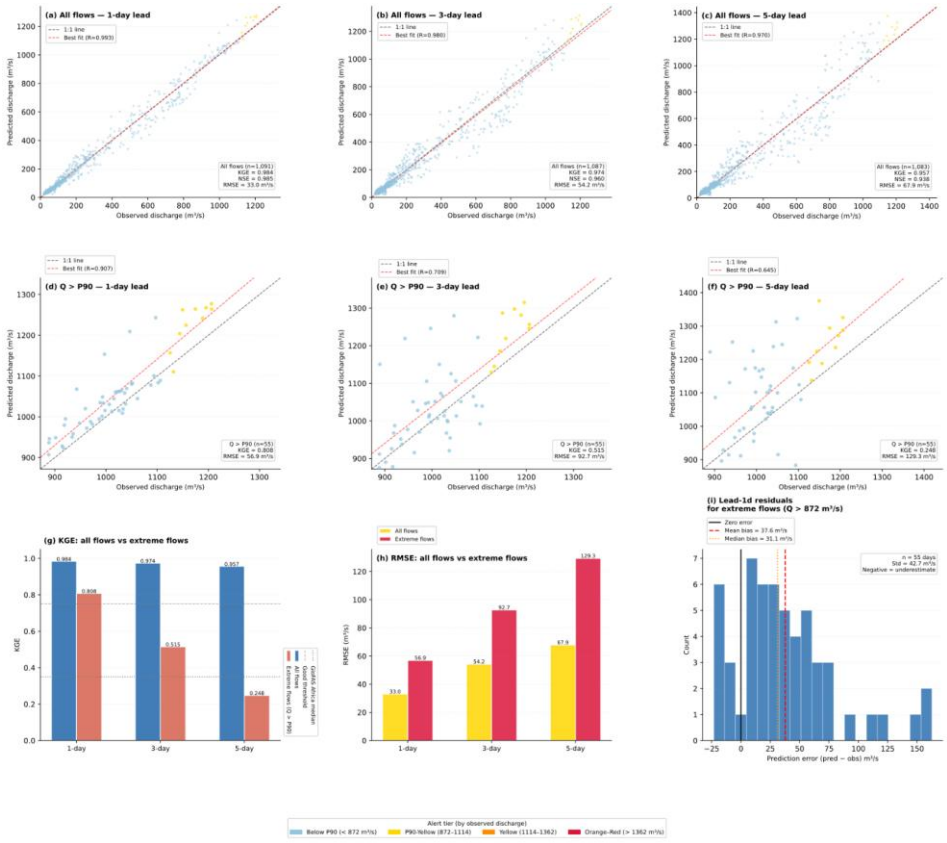


Figure 8. Ensemble model performance on extreme and flood-tier discharge events. Panels (a–c) show observed versus predicted scatter plots for all test period flows at 1-day, 3-day, and 5-day lead times, with points coloured by alert tier. Panels (d–f) restrict the analysis to extreme flows (Q > 872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

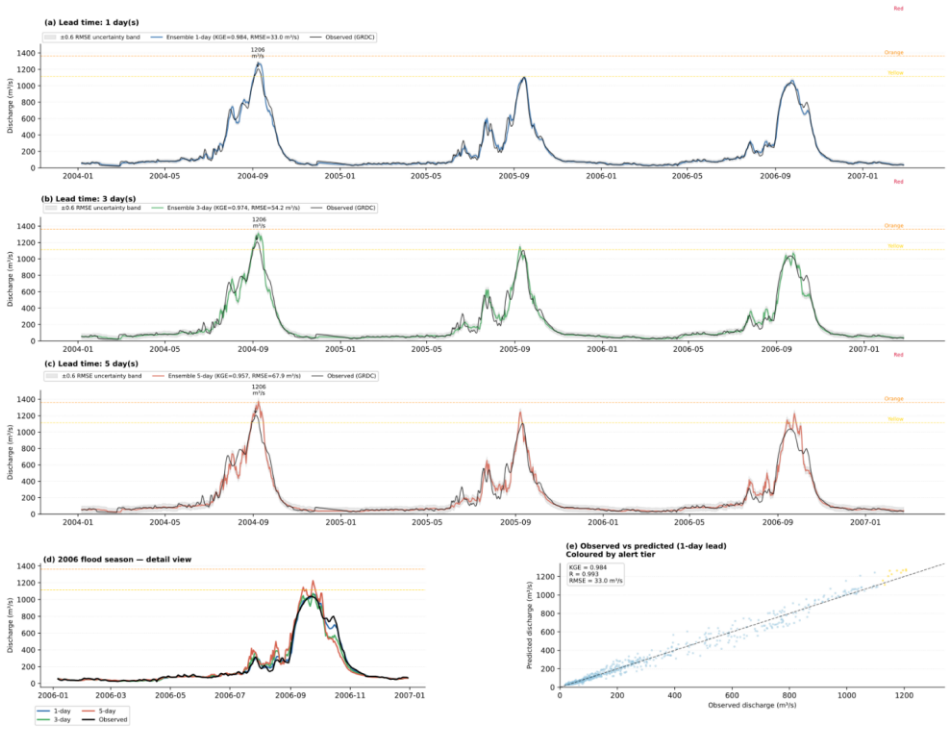


Figure 9. Observed versus ensemble-predicted discharge at Nawuni gauge during the independent test period (January 2004 to February 2007, GRDC observed discharge). Panels (a–c) show the full time series at 1-day, 3-day, and 5-day lead times respectively, with alert threshold lines and flood peak annotations. 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.

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 (m^3/s)	MAE (m^3/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

Formatted: Line spacing: Double

Baseline comparison

Formatted: Font: Bold

To test whether the ensemble adds value beyond recent observed discharge, we compared it against seven reference predictors evaluated on identical forecast issue dates and the same 21 leave-one-year-out folds used for the ensemble: persistence (the most recent observed discharge at issue time), a leave-one-year-out day-of-year climatology, ordinary least squares regression on discharge lags only and on the full feature set, and random forests trained on rainfall features only, on the Bagre storage proxy only, and on all exogenous drivers with every discharge feature withheld. Skill was assessed on all days and, separately, on rising-limb days (target discharge increasing) and high-flow days (observed discharge at or above the Yellow threshold of 1,114

m3/s), since these regimes govern warning performance (Table 5). Three findings follow. First, at one-day lead the ensemble carries no advantage over persistence, which attains KGE = 0.997 across all days and 0.958 on high-flow days against the ensemble's 0.995 and 0.951; at this horizon the forecast is dominated by the persistence of recent discharge, and every predictor that ignores recent discharge performs poorly. Second, the ensemble's marginal skill grows with lead time and concentrates in the regimes that determine warning performance: on rising-limb days it exceeds persistence by 0.010, 0.073 and 0.121 KGE units at 1, 3 and 5-day leads, and on high-flow days by -0.007, 0.031 and 0.093. Third, in the interest of a fair account, a linear model on the full feature set is a strong competitor and marginally outperforms the ensemble on high-flow days at 3 and 5-day leads; we read this not as a weakness but as confirmation that the contribution of this work is the end-to-end warning design and the explicit incorporation of the Bagre storage signal, which any of these architectures can exploit once it is supplied as an input, rather than the sophistication of the learning algorithm.

Table 5. Baseline comparison (Kling-Gupta Efficiency). All baselines evaluated under the ensemble's 21-fold leave-one-year-out protocol on identical forecast issue dates. Random forest baselines: 80 trees, maximum depth 18.

Model	all 1d	all 3d	all 5d	rising 1d	rising 3d	rising 5d	high-flow 1d	high-flow 3d	high-flow 5d
Persistence	0.997	0.986	0.968	0.937	0.851	0.787	0.958	0.859	0.730
Climatology	0.751	0.751	0.751	0.752	0.751	0.750	0.188	0.188	0.184
Linear (discharge lags only)	0.996	0.985	0.967	0.955	0.893	0.840	0.962	0.894	0.817
Linear (all features)	0.996	0.986	0.973	0.961	0.915	0.882	0.964	0.908	0.853

Model	all 1d	all 3d	all 5d	rising 1d	rising 3d	rising 5d	high-flow 1d	high-flow 3d	high-flow 5d
RF (rainfall only)	0.435	0.488	0.535	0.468	0.534	0.587	-0.063	-0.019	-0.010
RF (Bagre only)	-0.011	-0.009	-0.007	-0.056	-0.058	-0.049	-0.341	-0.312	-0.320
RF (no discharge, exogenous only)	0.636	0.666	0.693	0.675	0.710	0.745	0.056	0.087	0.113
Ensemble (RF+XGB+LSTM)	0.995	0.986	0.973	0.948	0.924	0.909	0.951	0.890	0.823

Formatted: Line spacing: Double

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

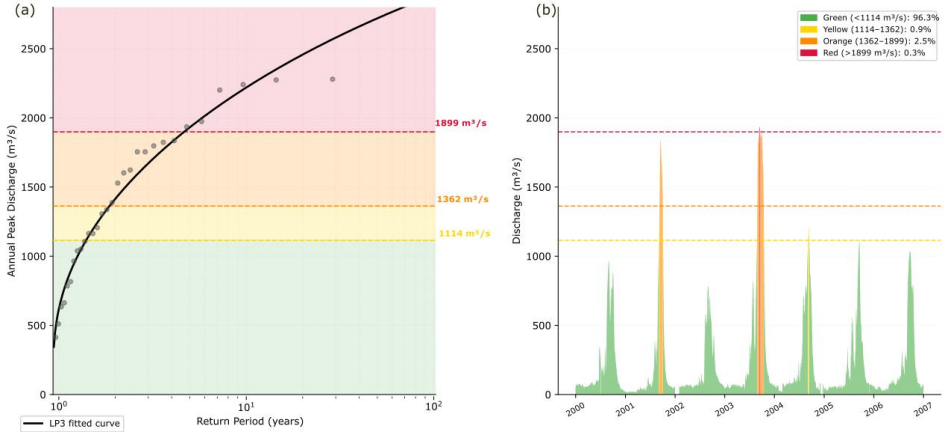


Figure 10. Alert threshold calibration. (a) Log-Pearson Type III frequency curve fitted to 30 wet-season annual maxima (1975–2006) at Nawuni, with the four alert tier thresholds marked. (b) Historical daily discharge at Nawuni (2000–2006) classified by alert tier, illustrating the distribution and seasonality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



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~~ alert ~~tier~~ tier ~~exceedance~~ exceedance ~~zones~~ zones ~~and~~ and ~~empirical~~ empirical ~~exceedance~~ exceedance probabilities.

Formatted: Justified, Line spacing: Double

It is important to distinguish between the return-period thresholds used here and the decision thresholds that emergency managers require in practice. The LP3 frequency analysis calibrates discharge exceedance probabilities based on the statistical properties of the observed record; it does not incorporate vulnerability, exposure, evacuation feasibility, or community impact data. A 5-year return period discharge at Nawuni does not automatically constitute an evacuation threshold in Central Gonja or Savannah districts, that determination depends on local topography, settlement patterns, road access, and the lead time available to move people and livestock to higher ground. The four-tier alert system presented here provides a statistically grounded starting point for threshold calibration, but operational deployment should involve co-design with NADMO district offices and Community Disaster Volunteer networks to translate return-period thresholds into locally meaningful action levels. This is consistent with the broader warning chain literature, which consistently finds that technically derived thresholds require social validation before they translate into protective behaviour (Demeritt et al. 2010).

5.3 Alert Skill

Table 2 presents alert detection skill from the 21-fold LOYO cross-validation, which covers the full 1987--2007 GRDC record without circular evaluation bias. Two complementary views are reported: day-level contingency statistics with stratified bootstrap 95% confidence intervals (2,000 resamples), and event-level statistics in which consecutive exceedance days are collapsed into independent flood episodes using a five-day gap rule.

At one-day lead, the ensemble achieves Yellow-tier POD = 0.970 (95% CI: 0.955--0.984) and FAR = 0.026 (0.014--0.040), Orange-tier POD = 0.964 (0.946--0.982) and FAR = 0.053 (0.031--

0.076), and Red-tier POD = 0.902 (0.844-0.951) and FAR = 0.134 (0.080--0.190). Performance degrades gradually with lead time: at five days, Red-tier POD falls to 0.770 (0.697-0.844) and FAR rises to 0.266 (0.195-0.331). The confidence intervals widen at longer leads and higher tiers, reflecting both reduced predictability and the smaller sample of extreme events available for evaluation.

The event-level evaluation clarifies the operational interpretation of these statistics. Across 13 observed Yellow-tier episodes, the ensemble detected all 13 at one-day lead with one false alarm episode. For Orange tier (13 observed episodes) the ensemble detected all 13 at one-day lead with zero false alarms, and all 13 at five-day lead with one false alarm. For Red tier (seven observed episodes) the ensemble detected six of seven at all three lead times (POD = 0.857) with zero false alarm episodes at any lead time. The single missed Red episode occurred in a year where peak discharge fell marginally above the 5-year return period threshold without the large Bagre Dam release signal that characterised other Red events. The false alarms visible in the day-level statistics at longer lead times predominantly reflect predicted exceedances falling one to three days outside the observed episode window rather than entirely spurious alerts, which in an operational context represents conservative rather than erroneous behaviour.

Formatted: Justified, Line spacing: Double

The event-level sample is small (seven Red episodes across 21 years) and the point estimates at this tier should be interpreted with appropriate caution. The bootstrap intervals at Red tier are wider than at Yellow or Orange (a range of approximately 0.15 KGE units in POD at five-day lead) and reflect genuine uncertainty rather than model instability.

~~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 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~~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.

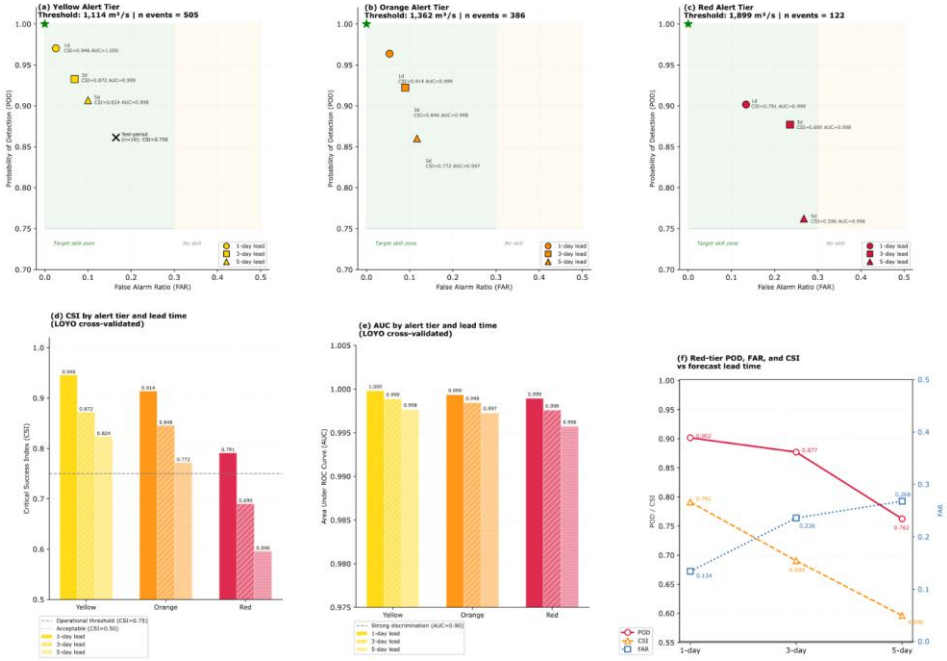


Figure 12. Receiver operating characteristic (ROC) analysis from leave-one-year-out cross-validation (1987–2007, n = 7,035 days). Panels (a–c) show the ROC space for the Yellow, Orange, and Red alert tiers respectively, with points indicating performance at 1-day, 3-day, and 5-day lead times. Panels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

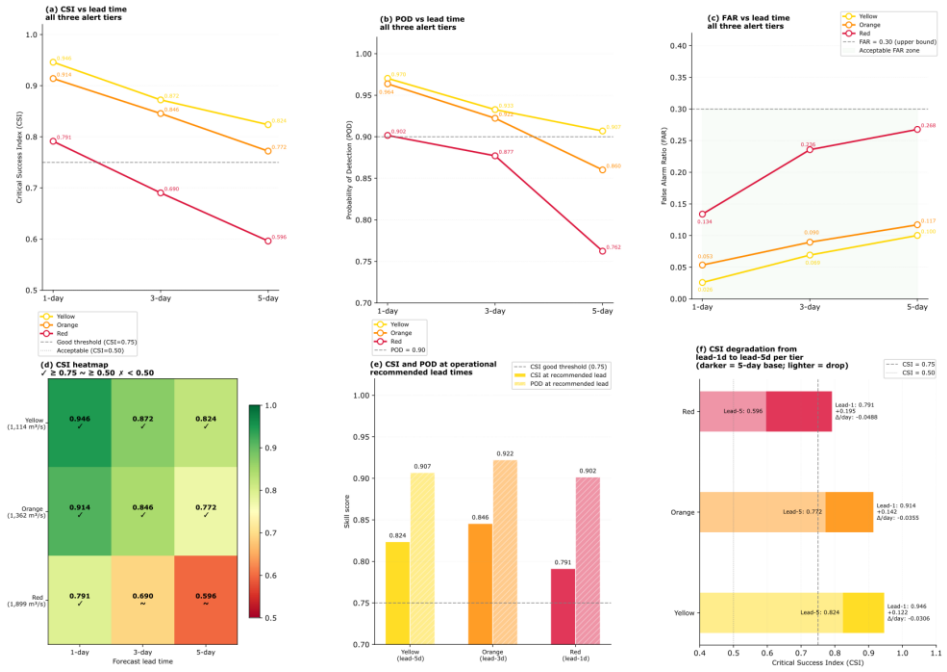


Figure 13. Alert detection skill summary from leave-one-year-out cross-validation (1987–2007, n = 7,035 days). (a–c) CSI, POD, and FAR as a function of forecast lead time for all three alert tiers. (d) CSI heatmap across tiers and lead times. (e) CSI and POD at the recommended operational lead time for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

Table 2. Alert detection skill from 21-fold leave-one-year-out cross-validation (1987–2007).~~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.~~

A. Day-level evaluation (LOYO cross-validation, 1987–2007)								
Lead	Alert tier	n days	POD	95% CI	FAR	95% CI	CSI	95% CI
1-day	Yellow	505	0.970	[0.955–0.984]	0.026	[0.014–0.040]	0.946	[0.926–0.965]
	Orange	386	0.964	[0.946–0.982]	0.053	[0.031–0.076]	0.914	[0.887–0.942]
	Red	122	0.902	[0.844–0.951]	0.134	[0.080–0.190]	0.791	[0.728–0.852]
3-day	Yellow	505	0.933	[0.911–0.955]	0.069	[0.049–0.091]	0.872	[0.844–0.899]
	Orange	386	0.922	[0.894–0.948]	0.089	[0.063–0.117]	0.846	[0.811–0.879]
	Red	122	0.877	[0.820–0.934]	0.236	[0.168–0.297]	0.690	[0.624–0.759]
5-day	Yellow	505	0.907	[0.881–0.931]	0.100	[0.076–0.126]	0.824	[0.791–0.853]
	Orange	386	0.860	[0.824–0.894]	0.117	[0.086–0.147]	0.772	[0.733–0.812]
	Red	122	0.770	[0.697–0.844]	0.266	[0.195–0.331]	0.603	[0.531–0.678]

B. Event-level evaluation (independent flood episodes, 5-day gap rule)								
Lead	Alert tier	Obs epis.	Pred epis.	TP	FN	FP	POD	FAR / CSI
1-day	Yellow	13	14	13	0	1	1.000	0.071 / 0.929
	Orange	13	13	13	0	0	1.000	0.000 / 1.000
	Red	7	5	6	1	0	0.857	0.000 / 0.857
3-day	Yellow	13	18	13	0	4	1.000	0.235 / 0.765
	Orange	13	14	13	0	1	1.000	0.071 / 0.929
	Red	7	5	6	1	0	0.857	0.000 / 0.857
5-day	Yellow	13	20	13	0	6	1.000	0.316 / 0.684
	Orange	13	15	13	0	1	1.000	0.071 / 0.929
	Red	7	5	6	1	0	0.857	0.000 / 0.857

664

~~Alert Tier~~	~~Lead~~	~~POD~~	~~FAR~~	~~CSI~~	~~AUC~~	~~n events~~
~~Yellow (1,114 m^3/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 m^3/s)~~	~~1-day~~	~~0.964~~	~~0.053~~	~~0.914~~	~~0.999~~	~~386~~
	~~3-day~~	~~0.922~~	~~0.090~~	~~0.846~~	~~0.998~~	~~386~~

~~Alert Tier~~	~~Lead~~	~~POD~~	~~FAR~~	~~CSI~~	~~AUC~~	~~n events~~
	~~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

Table 4. Skill of the Bagre reservoir fullness trigger, re-evaluated on the altimetry-validated corrected storage signal.

Evaluation window	Years scored	TP	FN	FP	TN	POD	FAR	CSI
1984-2006 (full record)	18 (5 no data)	5	6	4	3	0.455	0.444	0.333
1994-2006 (post-impoundment)	12 (1 no data)	5	2	4	1	0.714	0.444	0.455
1999-2006 (reliable coverage)	8 (0 no data)	4	0	4	0	1.000	0.500	0.500

The trigger fires when the corrected pool reaches 90% of capacity. Years without sufficient cloud-free JRC coverage are excluded from scoring and reported separately. In the reliably observed window the reservoir reaches capacity every year, so the trigger fires every year and its POD and FAR reflect the base rate rather than discriminating skill (see Section 6.2). POD, probability of detection; FAR, false-alarm ratio; CSI, critical success index.

Formatted: Line spacing: Double

Event-based hydrograph-shape validation

Formatted: Font: Bold

The contingency statistics in Table 2 establish detection skill but not the fidelity of the forecast hydrograph during flood events. We therefore evaluated three event-shape metrics on the 13 independent Yellow-and-above episodes in the leave-one-year-out record (five-day gap rule, target-date alignment): peak-magnitude error, alert-onset timing error (the difference between the observed and forecast first crossing of the Yellow threshold), and duration above the Yellow threshold (Table 6). Alert onset is reproduced accurately: the mean absolute onset timing error is 1.0, 1.9 and 2.0 days at 1, 3 and 5-day leads, with 92%, 77% and 77% of events triggering within two days of the observed onset and no missed onsets at any lead. Peak magnitude is captured to a mean absolute error of 3.5% to 7.9%, and duration above threshold to 1.0 to 3.4 days. Two limitations are stated plainly. First, the most extreme peaks are mildly and systematically under-predicted: across the six Red-tier events the mean peak bias is -2.8%, -5.3% and -5.6% at the three leads, and the alert-on duration is correspondingly shortened at longer leads. Second, peak-to-peak timing on the longest multi-pulse seasons is ambiguous, because two near-equal peaks separated by several weeks can exchange rank under a small magnitude perturbation; the median peak-timing error is 1 to 3 days, but we regard the alert-onset metric as the operationally meaningful timing measure. These results reinforce the caveat stated in the abstract and Section 6.3 that higher-tier skill rests on leave-one-year-out cross-validation rather than on held-out extreme events.

Table 6. Event-based hydrograph-shape validation on 13 independent Yellow-and-above flood episodes (leave-one-year-out record, 1987-2007), target-date alignment.

Lead	Peak-magnitude MAE	Peak-magnitude mean bias	Alert-onset timing MAE	Onset within +/- 2 days	Duration-above-Yellow MAE	Missed onsets
1-day	3.5% (61.6 m3/s)	-14.8 m3/s	1.0 days	92%	1.0 days	0 of 13
3-day	6.4% (114.1 m3/s)	-21.3 m3/s	1.9 days	77%	2.2 days	0 of 13
5-day	7.9% (137.4 m3/s)	-4.8 m3/s	2.0 days	77%	3.4 days	0 of 13

Formatted: Line spacing: Double

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.

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 validation seasons confirm that periods of Red-tier alert exceedance are associated with detectable floodplain inundation extents of 50-149 km² in the lower White Volta floodplain, providing spatial grounding for the discharge-based threshold system and demonstrating that the gauge-level alert tiers translate to physically

meaningful inundation conditions at the community scale (Figure 14).~~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).~~

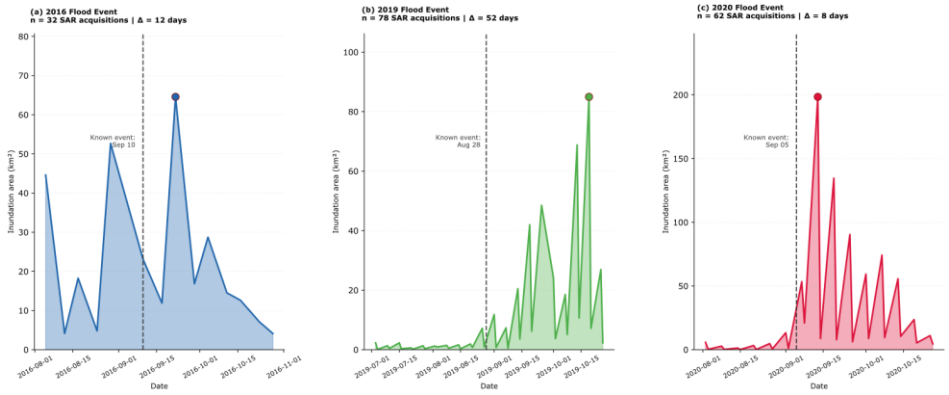


Figure 14 Sentinel-1 SAR flood inundation validation for three major flood events: (a) 2016, (b) 2019, and (c) 2020. Time series show daily inundation area (km²) derived from SAR backscatter change detection in Google Earth Engine. Vertical dashed lines indicate known event dates from the Dartmouth Flood 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 Bagre reservoir updates and real-time GHA gauge discharge where available. Layer 2 runs the trained ensemble to produce discharge forecasts at three lead times ~~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. It is important to note that in its current form the JRC-derived proxy is retrospective: monthly composites become available two to four weeks after the calendar month they represent, meaning the trigger provides seasonal situational awareness rather than real-time advance warning. Near-real-time Sentinel-1 monitoring of Bagre reservoir extent, which would reduce update latency from weeks to days, is identified as the highest-value future extension to this component of the system. 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.


Layer 1: Data Ingestion
CHIRPS near-real-time rainfall | ERA5-Land temperature, ET, soil moisture
JRC Bagre reservoir extent (monthly) | GHA gauge discharge (real-time)
GMet / GHA
Layer 2: AI Forecasting Pipeline
Random Forest | XGBoost | LSTM | Ensemble (simple average)
Lead times: 1-day / 3-day / 5-day | Bagre trigger: extent > 90% maximum
GMet (technical unit)
Layer 3: Alert Generation
GREEN < 1,114 m³/s YELLOW 1,114–1,362 m³/s ORANGE 1,362–1,899 m³/s RED > 1,899 m³/s
Output: colour-coded bulletin | Compatible with myDEWETRA-VOLTALARM
GMet (forecast office)
Layer 4: Institutional Linkage
GMet → NADMO National (≤ 1 hr) | NADMO National → Regional offices (≤ 2 hr)
Regional → District Assemblies (≤ 3 hr) | District → CDV network (≤ 4 hr)
NADMO
Layer 5: Community Dissemination
Community radio • SMS broadcast | WhatsApp groups • Town crier
Languages: Gonja / Dagbani / English | Tiered messages calibrated per alert level
District Assembly / CDV network
★ Replaces uncalibrated GloFAS dependency ★
No changes to existing myDEWETRA-VOLTALARM dissemination chain

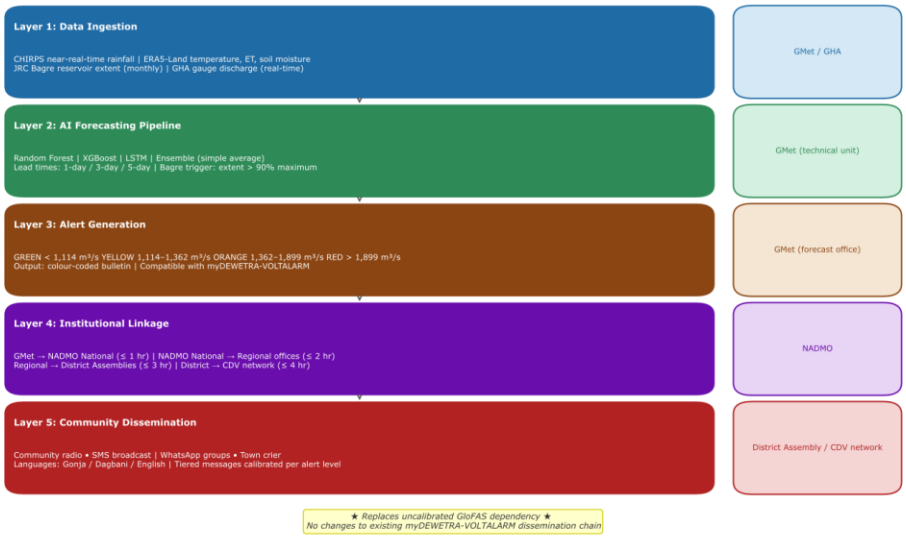


Figure 15. Five-layer operational architecture of the AI-driven Flood Early Warning System for the White Volta Basin. The system integrates daily automated data ingestion, ensemble AI forecasting, threshold-based alert generation, institutional dissemination through NADMO, and community warning delivery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

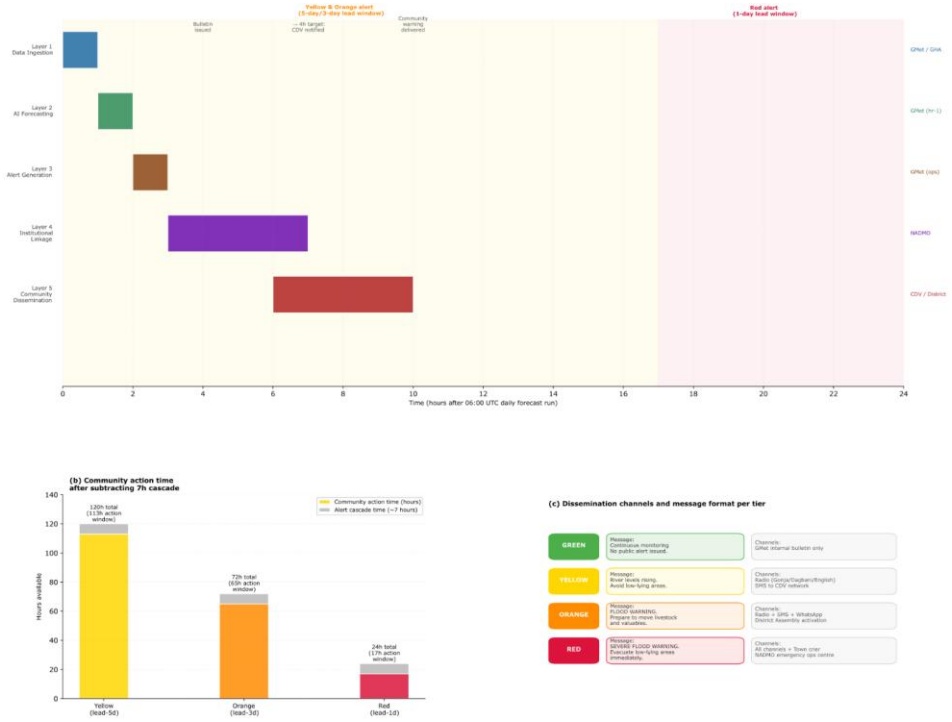


Figure 16. Operational alert cascade from forecast run to community warning delivery. (a) Gantt-style timeline showing the five architecture layers and their task sequence after the daily 06:00 UTC forecast run, with lead-time windows for Yellow, Orange, and Red alerts indicated. (b) Community action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

5.6 Towards a people-centred warning system: risk knowledge, impact-based alerts and response

The architecture in Section 5.5 delivers the forecasting, alert-generation and dissemination functions of an early warning system, but a warning system protects lives only when it is anchored in local risk knowledge and connected to response capability. Following the widely adopted four-element view of people-centred early warning (Basher, 2006; UNISDR, 2006), the present system is strongest in the detection and forecasting element and specifies the

Formatted: Heading 1 Char

Formatted: Font: 12 pt, Not Bold

dissemination element through the NADMO cascade and multilingual messaging of Section 5.5. This subsection sets out how the remaining two elements, risk knowledge and response capability, are addressed, and how the return-period thresholds of Section 5.2 are translated into impact-based warnings rather than discharge numbers alone.

Risk knowledge. Exposure and vulnerability in the lower White Volta are spatially concentrated and socially specific. The flood-prone zone between Nawuni and the Volta Reservoir, spanning the Central Gonja and Savannah districts, combines flat terrain, shallow soils and poor drainage, so that inundation is extensive and persists for several weeks after peak discharge (Section 2). The population exposed to this hazard depends heavily on recession agriculture on the alluvial floodplain, and the districts are among the most food-insecure in Ghana, so a flood destroys not only property but standing crops, seed stocks and livestock, setting households back by multiple growing seasons. The dominant driver of the most severe events is transboundary and not locally observable: uncontrolled spillage from the Bagré reservoir, which is operated for irrigation rather than flood control and carries no formal notification requirement to Ghana. Effective risk knowledge in this basin therefore means understanding not only who and what is exposed on the Ghanaian floodplain, but that the largest events are governed by upstream reservoir behaviour, which is the specific gap the reservoir-monitoring component addresses.

Impact-based warning. The hydrological thresholds derived in Section 5.2 acquire operational meaning only when each tier is mapped to the impacts it implies and the actions it should trigger. Table 5 sets out this mapping for the three tiers, linking discharge thresholds to expected physical impacts on communities, farmland, roads and public facilities, to recommended protective actions, and to the institution responsible for each. We present it as an operational design grounded in the documented character of flooding in these districts, rather than as an

Formatted: Font: 12 pt

Formatted: Font: 12 pt, Not Bold

Formatted: Font: 12 pt

Formatted: Font: 12 pt, Not Bold

empirically calibrated impact model, because the precise discharge at which a given road is cut or a given community is isolated depends on local topography and asset placement that only field co-design with NADMO district offices can establish (Demeritt et al., 2010). The table is therefore a template for that co-design, not a substitute for it.

Standard operating procedures and response capability. Each tier in Table 5 corresponds to a standard operating procedure that fixes who acts, in what order, within the daily cascade timeline of Figure 16. At Yellow, the procedure is advisory and preparatory; at Orange, it moves to pre-positioning of resources, opening of designated shelters on high ground, and staged movement of people, livestock and seed stocks; at Red, it becomes full activation of evacuation to shelters along pre-identified routes, coordination with security services to manage the cutting of the north-south trunk-road corridor near Yapei, and readiness of the district health directorate for water-borne disease. Response capability rests on the existing Community Disaster Volunteer network as the last link in the chain, on shelters and evacuation routes that should be inventoried in advance for each exposed community, and on seasonal drills that rehearse the cascade before the wet season rather than during it.

Community feedback. A people-centred system is two-way. The same Community Disaster Volunteer network that delivers warnings is the channel through which communities report back observed water levels, the comprehension and reach of tiered messages, and their tolerance for false alarms, all of which feed the iterative calibration of both the impact thresholds and the message formats. This feedback loop is the mechanism by which the statistically derived tiers of Section 5.2 are ground-truthed against lived impact over successive seasons, and it is a precondition for the thresholds to translate into protective behaviour rather than warning fatigue.

Formatted: Font: Bold

We are candid that the present study specifies this people-centred layer as a design and does not evaluate a deployed system. Establishing the impact thresholds, shelter and route inventories, and feedback protocols in the field is a programme of work in its own right, and it is the natural next step beyond the forecasting engine and architecture demonstrated here.

Table 7. Impact-based warning matrix linking each alert tier to expected impacts, protective actions and institutional responsibilities in the lower White Volta.

Alert tier (threshold)	Expected impact in the lower White Volta	Recommended protective action	Lead responsibility
Yellow (1,114 m3/s, 1.5-year)	Low-lying riverside farmland and recession plots begin to flood; minor feeder roads and footpaths in Central Gonja and Savannah start to be cut	Issue advisory via myDEWETRA-VOLTALARM; activate Community Disaster Volunteers; advise farmers to move harvested stock and livestock off floodplain margins; heighten monitoring	GMet (forecast and advisory); NADMO district (volunteer activation); GHA (gauge monitoring)
Orange (1,362 m3/s, 2-year)	Widespread inundation of standing crops; feeder roads and community access routes impassable; low-lying schools and health posts and isolated riverside settlements at risk	Pre-position relief resources; disseminate tiered messages in Gonja, Dagbani and English by radio, SMS and WhatsApp; open designated shelters on high ground; stage movement of vulnerable people, livestock and seed stocks; suspend school sessions in exposed communities	NADMO (coordination and shelters); district assemblies (schools and health); GMet and GHA (monitoring)
Red (1,899 m3/s, 5-year)	Severe, prolonged inundation over weeks; destruction of crops, seed stocks and livestock; cutting of the north-south trunk-road corridor	Activate emergency response and evacuation to shelters along pre-identified routes; coordinate with security services on the trunk-road corridor; ready	NADMO (lead) with Regional Coordinating Council; district health and education

Alert tier (threshold)	Expected impact in the lower White Volta	Recommended protective action	Lead responsibility
	near Yapei; isolation of communities; damage to schools, health and water and sanitation facilities; displacement	the health directorate for water-borne disease; mobilise relief logistics	directorates; GMet and GHA (monitoring)
Reservoir watch (validated Bagre signal at or near capacity)	Upstream reservoir full and capable of spilling; heightened likelihood of a severe downstream event in the following weeks, independent of locally modelled discharge	Raise watch level and pre-alert downstream districts; increase monitoring frequency; note the monthly composite latency, so this is seasonal situational awareness rather than sub-monthly warning	GMet and NADMO (watch coordination)

Formatted: Line spacing: Double

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 v4), across all African basins including genuinely ungauged catchments where performance is necessarily lower, and using a different evaluation methodology and period. A direct benchmark against GloFAS v4 reanalysis discharge at Nawuni confirms the performance advantage of the locally calibrated ensemble. GloFAS v4 reanalysis discharge was extracted from the ECMWF LISFLOOD reanalysis archive (Grimaldi et al. 2024) at the Nawuni grid cell (9.725°N, 1.075°W, resolution 0.05°) and evaluated against the same GRDC observations over the identical test period

(January 2004 to February 2007, n = 1,135 days). GloFAS v4 achieves a KGE of −0.526, NSE of −0.567, Pearson R of 0.425, and RMSE of 335.3 m³/s at Nawuni. A KGE below zero indicates that GloFAS v4 performs worse than a naïve climatological mean predictor on this gauge. Inspection of the discharge time series reveals the source of this failure: GloFAS v4 produces near-zero discharge throughout the wet season while observed discharge peaks between 500 and 1,200 m³/s, because LISFLOOD does not represent Bagre Dam release operations and therefore cannot reproduce the downstream flood signal that dominates discharge at Nawuni. This is not a limitation of GloFAS v4 in general but a specific and predictable consequence of applying a globally parameterised model without reservoir operation rules to a gauge that is hydraulically downstream of a major regulated reservoir. The locally calibrated ensemble, trained directly on GRDC observations at Nawuni and incorporating the Bagre storage proxy as an explicit input, achieves KGE = 0.984 at one-day lead on the same data, a difference of 1.51 KGE units. The performance advantage is not marginal: it reflects access to information that no globally calibrated model without explicit dam representation can replicate.

~~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).

6.2 The Bagre Signal

The Bagre Dam upstream storage proxy is the most novel element of the feature engineering, and its role in the system merits careful interpretation. Its near-zero linear importance in the tree-based models (under 0.1% of Random Forest Gini importance at one-day lead) does not undercut its inclusion: low importance at short lead means only that, once recent discharge lags are known, the index adds little to next-day prediction, 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 redundant for next-day forecasting but potentially informative at longer leads and at the onset of developing events, before elevated flows arrive at the gauge. This reading is consistent with the baseline comparison in Section 5.1, in which a Bagre-only model has no standalone skill

Formatted: Keep with next

while the reservoir signal contributes as an increment to a discharge-based model.

We validated the reservoir proxy physically, because a satellite-derived water extent may capture floodplain, wetland or channel water rather than the reservoir itself. Three steps establish its fidelity. First, a polygon-sensitivity analysis shows that the original extraction envelope of 3,018 km^2 is only about 5% water and encloses the Nakambe valley floodplain and the Bagre Ramsar wetland rather than the reservoir pool; restricting the extraction to persistent-water pixels from the JRC occurrence layer (occurrence at or above 40%) reduces the surface to 94 km^2, consistent with the published minimum operating pool of approximately 85 km^2 (Figure 17a). Second, the raw monthly extent carried a physically inverted seasonal cycle, higher in the dry season than the wet, because optical water detection is suppressed by wet-season cloud; normalising each month by its count of valid observations removes the inversion and restores the expected wet-season filling (Figure 17b). Third, we validated the corrected signal against independent satellite altimetry (DAHITI virtual station 943), which derives from different sensors and physics: the corrected pool area tracks altimetric level at a Pearson coefficient of 0.94 over 135 monthly observations, against only 0.30 for the raw extent (Figure 17c). The proxy is therefore a physically measure of reservoir state, and the cloud correction is the decisive step in achieving this.

Re-evaluating the reservoir trigger on this validated signal produces a more honest picture than earlier analyses of the raw index. Across the reliably observed period (1999 to 2006), the corrected pool reaches 98 to 99.9% of capacity in every wet season, in flood and non-flood years alike, so a fullness threshold fires in all eight years, yielding a probability of detection of 1.0 that is trivial rather than skilful and a false-alarm ratio that simply equals the proportion of non-flood years (Table 4). This is physically expected: the Bagre reservoir fills to capacity in most wet seasons, and whether Ghana floods downstream depends on the magnitude and timing of controlled releases

and on continuing inflow, not on reservoir fullness alone. Two further factors shape Table 4. The reservoir was impounded only in 1993, so the pre-1994 flood years in the full-record window pre-date its operation; and the corrected surface area saturates once the pool is full, because the banks are steep, so area ceases to rise even as level continues to climb, while the altimetric level, available only from 2008 and therefore not overlapping the gauge record, cannot yet be tested against observed floods.

The Bagre reservoir is positioned as a signal within the warning system. It is not a standalone annual flood trigger. It is a validated measure of reservoir state that contributes in two ways: as an explicit input feature to the discharge model, and as seasonal situational awareness that the reservoir is at or near capacity and therefore capable of spilling. The monthly JRC composite latency, of two to four weeks, reinforces this seasonal framing; near-real-time Sentinel-1 monitoring of the reservoir surface area, reducing update latency from weeks to days, remains the highest-value technical improvement available to the operational system and is the subject of ongoing work, since it is the release event rather than the fullness state that determines downstream impact.

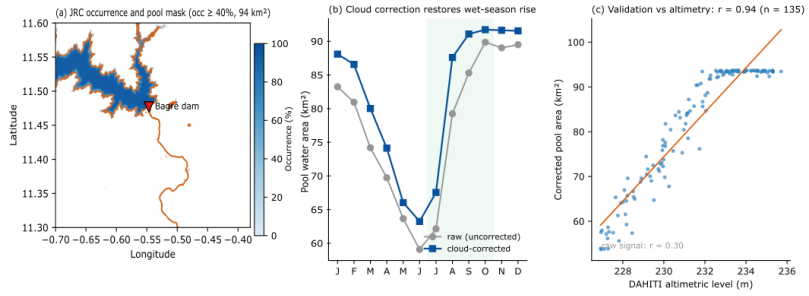


Figure 17. Physical validation of the Bagre reservoir storage proxy. (a) JRC water occurrence over the reservoir, with the occurrence at or above 40% pool mask (94 km2) outlined and the dam marked. (b) Monthly pool water area before and after cloud correction; normalising by valid observations removes the physically inverted seasonal cycle and restores wet-season filling. (c) Cloud-corrected pool area against independent DAHITI satellite-altimetric water level, 2008 to 2021 (Pearson r = 0.94, n = 135 months); the raw uncorrected signal achieves only r = 0.30.

~~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 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 m³/s. Alert skill for the higher tiers therefore rests on leave-one-year-

965 out cross-validated evidence rather than a strictly held-out sample. The LOYO design eliminates
966 training-data circularity but is not equivalent to a fully independent test set. Obtaining an
967 independent evaluation of Red-tier skill would require either a longer gauge record or the
968 acquisition of discharge data for the post-2007 flood events, particularly 2007, 2010, and 2020.

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

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

979 Fourth, the GloFAS benchmark comparison is indirect, based on published aggregate statistics
980 from Harrigan et al. (2020) for GloFAS v2.1, rather than a direct evaluation of GloFAS v4 at
981 Nawuni using identical data and metrics. This limits the strength of the performance claims relative
982 to the current operational system. However, several lines of qualitative evidence suggest the
983 direction of the performance gap is robust. The GloFAS v4 hydrological reanalysis, evaluated
984 across 1,995 calibration stations globally, reports a median modified KGE' of 0.70 (Grimaldi et
985 al., 2024; accessible via the GloFAS map viewer at https://globalfloods.eu under Evaluation →
986 Hydrological Model Performance). Stations in West Africa, and particularly those on regulated
987 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 fifth operational limitation concerns the system's dependence on real-time GRDC discharge at Nawuni. The locally calibrated ensemble requires recent gauge observations to populate its autoregressive lag features, which account for the majority of predictive variance at short lead times. When real-time discharge is missing, delayed, or erroneous（conditions that occur regularly in operational settings due to telemetry failures, sensor maintenance, or transboundary data-sharing gaps）the system cannot produce a reliable ensemble forecast. Three fallback strategies are available, each with different skill implications. First, the ERA5-Land runoff proxy used for the post-2007 Sentinel-1 validation period demonstrates that a bias-corrected reanalysis discharge estimate can substitute for gauge observations, though at substantially reduced skill for extreme events; this fallback is available with zero additional infrastructure and activates automatically when gauge data are absent for more than two consecutive days. Second, the Bagre Dam reservoir storage trigger operates entirely independently of the discharge model and continues to issue seasonal-scale alerts when the JRC index exceeds 0.90, providing a safety net that does not depend on gauge data availability at all. Third, the XGBoost and Random Forest components of the ensemble can run in a degraded configuration that replaces missing discharge lag features with the most recently available observation carried forward, a strategy that maintains reasonable

performance over gaps of up to three days but degrades sharply beyond that. Formalising these fallback protocols into the operational architecture, including automated detection of gauge data outages and seamless switching between forecast modes, is a necessary engineering step before deployment.~~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.~~

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, integrating observed gauge discharge, CHIRPS satellite rainfall, ERA5-Land reanalysis forcing, and a JRC-derived Bagre reservoir storage proxy, and validated against independent Sentinel-1 SAR inundation observations as spatial grounding for the discharge-based threshold framework. Four conclusions are worth stating directly.

~~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 GloFAS v4 at Nawuni across all forecast horizons. A direct comparison using identical data and metrics for the January 2004 to February 2007 test period shows GloFAS v4 achieving KGE = −0.526 at Nawuni, performing worse than a climatological mean predictor, primarily because LISFLOOD does not represent

Bagre Dam release operations and therefore cannot reproduce the downstream flood signal that dominates discharge at this gauge. The locally calibrated ensemble achieves KGE values of 0.984, 0.974, and 0.957 at one-, three-, and five-day lead times on the same data, a difference of 1.51 KGE units at one-day lead. The performance advantage is most pronounced at five-day lead, the horizon that matters most operationally, where global systems trained without local calibration and without explicit reservoir representation consistently underperform.

This advantage is specific to the comparison with globally parameterised models. Against simple local baselines such as persistence and linear regression, the ensemble's margin is modest and concentrated at the longer leads, so the contribution lies in the end-to-end warning design and the explicit Bagre input rather than in the learning algorithm itself.

~~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 an informative measure of reservoir state that current operational systems ignore, but it is not a standalone flood trigger. The cloud-corrected reservoir signal, validated against independent satellite altimetry at a Pearson coefficient of 0.94, faithfully tracks reservoir level. However, the reservoir fills to near capacity in most wet seasons, so reservoir

fullness alone does not discriminate the years that flood downstream from those that do not. We therefore employ the signal as an explicit input to the discharge model and as seasonal situational awareness that the reservoir is at capacity and capable of spilling, rather than as an independent alarm. The monthly JRC update lag reinforces this; near-real-time Sentinel-1 monitoring of the Bagre surface area would reduce update latency from weeks to days and, together with an automated monitoring protocol and, ideally, a bilateral notification arrangement with Burkina Faso's water authorities, represents the highest-value improvement available to the existing warning architecture.

~~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 with stratified bootstrap 95% confidence intervals show Red-tier alerts achieving a probability of detection of 0.902 [0.844–0.951] at one-day lead and 0.770 [0.697–0.844] at five days, with false alarm ratios of 0.134 [0.080–0.190] and 0.266 [0.195–0.331] respectively. At the event level, across seven observed Red episodes in the 21-year cross-validation record, six were detected at all three lead times with

zero false alarm episodes. 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.

~~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 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. Several engineering and co-design steps remain,

notably the field calibration of impact thresholds and response protocols set out in Section 5.6, but the gap between the system described here and an operational deployment is a matter of implementation rather than of fundamental capability.

~~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 design, Sentinel-1 processing, and trained model files are archived at Zenodo (https://doi.org/10.5281/zenodo.20549601~~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 Meteorological Agency (GMet), the Ghana Hydrological Authority (GHA), and the National Disaster Management Organisation (NADMO) for institutional context and operational guidance ~~that~~ that ~~shaped~~ shaped ~~the~~ the ~~system~~ system design.

Funding

This work was supported by the National Natural Science Foundation of China under Grant No.72521001, the Natural Science Foundation of Beijing under Grant Nos. L255011 and 8242014, and the Strategic Study Project of Chinese Academy of Engineering (Grant No.2023-JB-08).

References

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

Cao, X., Wang, B., Yao, Y., Zhang, L., Xing, Y., Mao, J., Zhang, R., Fu, G., Borthwick, A. G. L., and Qin, H.: U-RNN high-resolution spatiotemporal nowcasting of urban flooding, *J. Hydrol.*, 659, 133117, https://doi.org/10.1016/j.jhydrol.2025.133117, 2025.

Cao, X., Yao, Y., Wang, Z., Zhao, Z., Borthwick, A. G. L., and Qin, H.: Large-scale urban flood modeling and zero-shot high-resolution generalization with LarNO, *J. Hydrol.*, 676, 135686, https://doi.org/10.1016/j.jhydrol.2026.135686, 2026.

Formatted: Font: 16 pt, Bold
Formatted: Font: 12 pt
Formatted: Line spacing: Double
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt

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

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

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

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

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge

Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt

reanalysis 1979–present, *Earth Syst. Sci. Data*, 12, 2043–2060, https://doi.org/10.5194/essd-12-2043-2020, 2020.

Katsekpor, J. T., Greve, K., Yamba, E. I., and Amoah, E. G.: Flood early warning systems in the White Volta Basin, Ghana: challenges and opportunities, *J. Flood Risk Manage.*, e70146, https://doi.org/10.1111/jfr3.70146, 2025a.

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

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

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

Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt

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

Mosaffa, H., Pappenberger, F., Prudhomme, C., Chantry, M., Rudiger, C., and Cloke, H.: A GNN routing module is all you need for LSTM rainfall–runoff models, *Hydrol. Earth Syst. Sci.*, 30, 2079–2092, https://doi.org/10.5194/hess-30-2079-2026, 2026.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, *Earth Syst. Sci. Data*, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

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

Formatted: Font: 12 pt
Field Code Changed

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

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

Taghizadeh, M., Zandsalimi, Z., Nabian, M. A., Goodall, J. L., and Alemazkoor, N.: FloodForecaster: a domain-adaptive geometry-informed neural operator framework for rapid flood forecasting, *J. Hydrol.*, 664, 134512, https://doi.org/10.1016/j.jhydrol.2025.134512, 2026.

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

Xu, Q., De Vos, L. F., Shi, Y., Ruther, N., Bronstert, A., and Zhu, X. X.: Urban flood modeling and forecasting with deep neural operator and transfer learning, *J. Hydrol.*, 661, 133705, https://doi.org/10.1016/j.jhydrol.2025.133705, 2025.

Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Field Code Changed
Formatted: Font: 12 pt
Formatted: Font: 12 pt

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

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

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

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

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

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

Formatted: Left

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

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

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

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

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

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J. N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Nearing, G. S., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Henck, A., Hurkmans, R., Johnson, D., Matias, M., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C.,

Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Kratzert, F.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

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

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

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