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