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

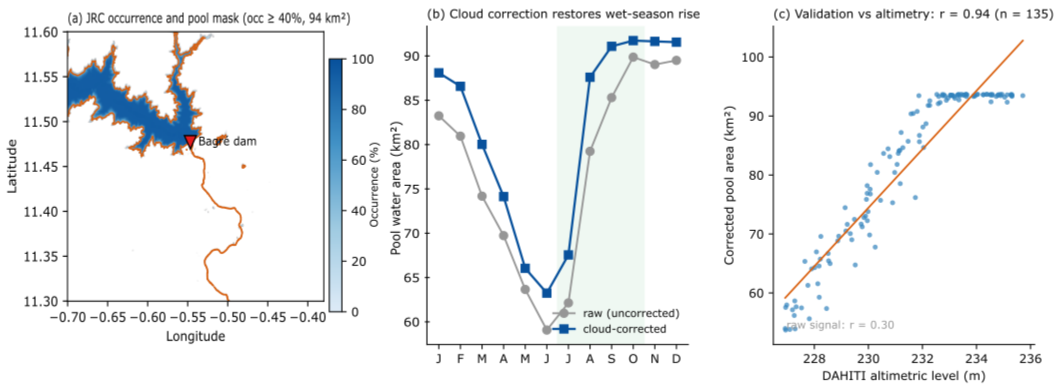

figure Physical validation of the Bagre reservoir storage proxy. (a) JRC water occurrence over the reservoir, with the occurrence at or above 40% pool mask (94 km2) outlined and the dam marked. (b) Monthly pool water area before and after cloud correction; normalising by valid observations removes the 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.