International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 11 November-2026 | Impact Factor: 3.5



# FloodWatch: A Provenance-Aware Deep Learning Framework for Satellite-Based Flood Intelligence and Verified Disaster Alerting

**B. J. Mithil Reddy, Dhanvi K Shetty, Bhavani Krupakara S, A Sohan Sri Datta, Ms. Veda N**

Department of Computer Science and Engineering
RV Institute of Technology and Management, Bengaluru, India

**How to Cite this Article**:
Reddy, B. J. M., Shetty, D. K., S, B. K., Datta, A. S. S. & N, V. (2026). FloodWatch: A Provenance-Aware Deep Learning Framework for Satellite-Based Flood Intelligence and Verified Disaster Alerting. International Journal of Creative and Open Research in Engineering and Management, 02(11), 1-9. https://doi.org/10.55041/ijcope.v2i11.171

**License:**
This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.
© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.







https://doi.org/10.55041/ijcope.v2i11.171

**Abstract—Timely flood warning is hardest where gauges are sparse and clouds block optical satellites. This paper presents FloodWatch, an open-access research prototype coupling Sentinel-1 SAR inundation mapping with hydrometeorological risk models and an operator-verified alerting layer. An EfficientNet-B3 U-Net trained on Sen1Floods11 reaches a test IoU of 0.5142 and Dice of 0.6572. A GRU flood-susceptibility model trained on 352 labelled locations reaches a test AUC of 0.8952 (n = 70, 8 positives, 95% CI [0.75, 1.00]) on a geographically separated partition; because labels are location-level, it estimates susceptibility under current conditions, not event-time forecast skill. Cross-attention, LSTM and TCN predictors were pre-trained on synthetic generators, so their scores measure surrogate fidelity only. Every ingested field carries provenance and freshness tags, and missing data is never treated as safe. A cost–loss policy with a trust gate, hysteresis and daily alert budgets drives operator-approved CAP 1.2 export in dry-run mode. Closed-loop skill, uncertainty calibration and ablations remain future work.**

**Index Terms—Sentinel-1 SAR segmentation, U-Net, flood susceptibility, data provenance, cost–loss decision analysis, Common Alerting Protocol**

## I. INTRODUCTION

Floods are among the most frequent and costly natural hazards, and their human toll is rising. An estimated 1.81 billion people, about 23% of the world population, live in areas exposed to a 1-in-100-year flood, and nearly nine in ten of them are in low- and middle-income countries [1]. Satellite observations show that the number of people exposed to flooding grew considerably faster than flood extent itself [9], so every hour of warning time and mapping speed translates into lives and livelihoods protected. Yet the regions that need warning most often lack dense gauge networks. Recent work shows that learning-based hydrological models can issue reliable extreme-flow warnings in ungauged basins several days ahead [2], building on LSTM rainfall–runoff models [10], [11] and on modelled global discharge such as GloFAS [13], [14]. In parallel, Sentinel-1 synthetic aperture radar (SAR) [16] has become the workhorse of emergency flood mapping because it observes day and night through cloud [17], and deep learning has largely replaced hand-tuned backscatter thresholds for this task [18]. Broader machine-learning approaches to flood prediction are surveyed in [19].

Despite this progress, most contributions remain isolated components. Segmentation studies typically stop at pixel metrics on benchmark chips such as Sen1Floods11 [6], using U-Net-style encoder–decoder networks [3] with efficient backbones [4] and overlap- or imbalance-aware losses [20], [21]. Forecasting studies develop recurrent, convolutional or attention-based sequence models [22]–[24] without linking their output to an exposure estimate, a decision rule or a dissemination channel. Three practical gaps follow. First, operational pipelines mix observations, reanalysis and forecasts without recording which is which, so a missing sensor reading can silently look like safety. Second, point predictions are rarely accompanied by epistemic uncertainty from deep ensembles [7] or Monte Carlo dropout [25], and almost never by a principled decision threshold; the classical cost–loss framework [26], [27] supplies such a threshold but is seldom wired into alert generation. Third, alerts that cannot be exported in a standard format such as the Common Alerting Protocol (CAP) [8] are hard to hand over to public warning authorities, and unreviewed automatic alerts invite fatigue and false alarms.

This paper presents FloodWatch, a prototype framework that addresses these gaps using open-access data only: ECMWF forecasts served through Open-Meteo [28], GloFAS discharge [14], JRC surface-water occurrence [29], WorldPop population grids [30], the Dartmouth Flood Observatory archive [31] and live GDACS alerts [32]. Our contributions are fivefold. (i) A four-channel EfficientNet-B3 U-Net for Sentinel-1 flood segmentation, followed by geospatial false-positive filtering. (ii) A set of complementary predictors, namely a cross-attention multi-factor network, a state-handoff LSTM, a dilated causal TCN with temporal attention and a GRU susceptibility model, equipped with deep-ensemble and Monte Carlo dropout uncertainty. (iii) A provenance and freshness layer in which every field is tagged by origin and age, and unavailable data is never interpreted as safe. (iv) A policy engine combining a cost–loss threshold, trust gating, hysteresis, a daily alert budget and mandatory operator approval before CAP 1.2 export. (v) A closed-loop verification scheme that refuses to treat modelled reanalysis as ground truth. The entire pipeline runs end to end on live data using a single CPU workstation, which lowers the barrier to replication compared with offline segmentation experiments. We also report limitations candidly, including which components were trained on real labels and which on synthetic benchmarks. The remainder of the paper is organised as follows: Section II describes data and provenance, Section III the methodology, Section IV the results, and Section V concludes.



## II. DATASETS AND DATA PROVENANCE

### A. Open-Access Telemetry Sources

FloodWatch ingests live data from the sources in Table I; no API keys are required. Observations (gauges, SAR scenes), analyses and forecasts (ECMWF IFS), modelled reanalysis (GloFAS) and external alert evidence (GDACS) are kept in separate provenance classes, because their evidential weight differs.

*TABLE I*
*OPEN-ACCESS DATA SOURCES AND PROVENANCE CLASS*

| Source | Data used | Class |
|---|---|---|
| ECMWF IFS (Open-Meteo) | Rain (14 d past, 72 h ahead), temperature, soil moisture, CAPE, snow | Analysis / forecast |
| GloFAS 4.0 | River discharge and climatology (14 d + 7 d) | Modelled reanalysis |
| Copernicus GLO-30 DEM | Elevation, slope, TWI | Static reference |
| Sentinel-1 (Planetary Computer / AWS) | VV, VH backscatter per overpass | Satellite retrieval |
| GDACS | Live flood-alert centroids | External alert evidence |
| NASA GPM / MERRA-2 | Precipitation / atmospheric reanalysis | Retrieval / reanalysis |
| USGS NWIS / CWC India | River gauge discharge and stage | Observed sensor |
| SoilGrids 250 m, WorldPop | Soil texture, population density | Static reference |
| OpenStreetMap, JRC GSW | Buildings, roads, permanent water | Static GIS reference |
| Dartmouth Flood Obs. | 4,029 flood events (1985–2010 snapshot) | Historical record |

### B. Training Data and Labels

The SAR model is trained on the Sen1Floods11 benchmark (Sentinel-1 dual-polarisation chips with flood annotations). The GRU susceptibility model (Section III-F) is trained on 352 labelled locations: 61 positives taken from the Dartmouth archive and 291 negatives drawn from low-discharge grid cells, each paired with observations produced by the same ingestion pipeline that is used at inference. Positives are DFO events from 1985–2010, whereas the features come from current ingestion at the event location; the features therefore do not describe conditions at event time. While historical reanalyses (such as ERA5 and GloFAS multi-decadal hindcasts) are accessible via open APIs, the present prototype evaluates physical susceptibility under prevailing conditions at known vulnerable locations rather than retrospective event-day forecasts. The cross-attention predictor (M2) and the streamflow LSTM (M3) were pre-trained on synthetic benchmarks, namely a multi-regime statistical generator and a Caravan-conforming hydrological simulator, respectively, because globally standardised multi-factor, gauge-verified flood records are scarce. Their scores therefore measure fidelity to the generator and must not be read as field skill. Operational inference, in contrast, uses no synthetic values.

### C. Provenance and Freshness

Each ingested field $f$ carries a record $\pi_f = (\text{class}, \text{product}, \text{version}, t_{\text{fetch}}, h_f)$, where $h_f$ is the SHA-256 digest of the raw payload, enabling replay verification. A freshness status is assigned from the age $\Delta t = t_{\text{now}} - t_{\text{obs}}$ according to a five-state lifecycle:

$$\text{status}(\Delta t) = \begin{cases} \text{FRESH}, & \Delta t \le 24\text{ h } (\le 48\text{ h for SAR}), \\ \text{ACCEPTABLE}, & 24 < \Delta t \le 48\text{ h}, \\ \text{STALE}, & 48 < \Delta t \le 168\text{ h}, \\ \text{EXPIRED}, & \Delta t > 168\text{ h}, \end{cases} \quad (1)$$

and fields lacking a verifiable timestamp are designated UNKNOWN. A field that is unavailable is excluded from scoring and flagged; it is never imputed and never counted as "safe".

## III. PROPOSED METHODOLOGY

### A. Framework Overview

FloodWatch follows a three-tier design (Fig. 1): a Leaflet-based browser client; a Node.js/Express server that handles routing, MongoDB and SQLite persistence, the alert policy engine, device registry and CAP export; and a FastAPI engine running the deep-learning models and geospatial processing. Live telemetry passes through provenance tagging and freshness validation, then through deep-learning inference (M1–M6) and the transparent risk score (M7). The outputs feed spatial hazard delineation, multi-scale impact assessment, response optimisation and, finally, alert candidates that wait for operator review before CAP 1.2 export and closed-loop verification. The overall procedure is formalised in Algorithm 1.

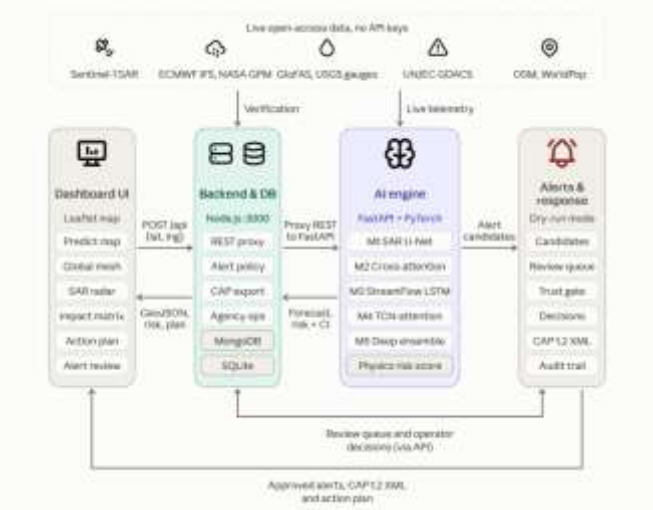



*Fig. 1. System architecture of the proposed FloodWatch framework.*

### B. SAR Flood Segmentation (M1)

#### 1) Input representation

For each Sentinel-1 chip, the input tensor stacks four channels: calibrated VV backscatter (dB), VH backscatter (dB), their cross-polarisation ratio $r_{\text{pol}} = \text{VV}_{\text{norm}}/(\text{VH}_{\text{norm}} + \epsilon)$ (min-max rescaled to [0,1]), and the JRC permanent surface-water occurrence prior (or Copernicus DEM elevation). All channels are normalised to [0,1].

#### 2) Network and thresholding

An EfficientNet-B3 encoder [4] with a U-Net decoder [3] implemented via `segmentation-models-pytorch` [5] produces a pixel-wise flood probability $\hat{P}$ via sigmoid activation. Training was conducted on the Sen1Floods11 hand-labeled benchmark (446 training, 60 validation, and 90 test chips of $512 \times 512$ pixels) [6] for 42 epochs with Adam optimizer (initial learning rate $1 \times 10^{-4}$, Cosine Annealing, batch size 8) under a combined Dice and Focal loss ($\gamma = 2.0$). Pixels with $\hat{P} > \tau$ are classified as inundated, with threshold $\tau = 0.65$ (tuned on the validation split to suppress false alarms over low-backscatter dry agricultural soils).



*3) Geospatial post-processing*

The binary mask is vectorised into polygons $\{\Omega_k\}$ and filtered by land boundaries and permanent water bodies. Spurious edge artifacts and polygons larger than $150\,\mathrm{km}^2$ (e.g., ocean boundary false positives) as well as sub-pixel noise clusters are removed ($A_k \le 150\,\mathrm{km}^2$). Areas are computed in an equal-area projection (EPSG:6933), $A_k = \sum_{(i,j)\in\Omega_k} a_{\mathrm{px}}$, the total flooded area is $A_{\mathrm{tot}} = \sum_{k\in\mathcal{K}} A_k$, and each polygon receives a hazard tier.

*C. Multi-Factor Cross-Attention Predictor (M2)*

Twenty-four live factors are converted into physically engineered features and split into domain groups (meteorology, hydrology, terrain, soil, exposure). Each group $g$ is encoded to a token $\mathbf{z}_g \in \mathbb{R}^d$ with $d = 128$, and the stacked tokens $\mathbf{Z}$ are fused by multi-head self-attention with $h = 4$ heads:

$$\mathrm{Attn}(\mathbf{Z}) = \mathrm{softmax}\left(\frac{\mathbf{Z}\mathbf{W}_Q(\mathbf{Z}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{Z}\mathbf{W}_V, \quad d_k = \frac{d}{h}. \qquad (2)$$

The fused representation passes through residual blocks and feeds three heads for flood probability $\hat{p}$, five-class severity $\hat{\mathbf{s}}$ and inundation depth $\hat{d}$. The multi-task objective combines cross-entropy (probability and severity) and regression (depth) losses.

*D. Temporal Forecasting (M3, M4)*

*1) StreamFlow LSTM (M3)*

A state-handoff encoder–decoder LSTM forecasts 72-hour streamflow across three multi-day lead horizons: $\hat{Q}_{T+24}, \hat{Q}_{T+48}, \hat{Q}_{T+72}$, where $T$ denotes the current analysis timestamp. The LSTM cell computes its gates and cell state from the concatenation of dynamic inputs and static catchment embeddings; the decoder is initialised with the encoder's final state, $\left(\mathbf{h}_0^{\mathrm{dec}}, \mathbf{c}_0^{\mathrm{dec}}\right) = \left(\mathbf{h}_T^{\mathrm{enc}}, \mathbf{c}_T^{\mathrm{enc}}\right)$.

*2) Dilated causal TCN with temporal attention (M4)*

The input is a sequence $\mathbf{X} \in \mathbb{R}^{72\times 10}$ spanning 72 antecedent hours across ten variables (precipitation, temperature, humidity, pressure, wind, CAPE, two soil-moisture layers, discharge and flow ratio). Causal dilated convolutions with kernel size $k = 3$ and dilation $d_l = 2^{l-1}$ process the sequence without lookahead bias across $L = 5$ residual levels ($d_l \in \{1,2,4,8,16\}$) with $n_c = 2$ stacked convolution layers per level, achieving a receptive field of $R = 1 + (k-1)n_c(2^L - 1) = 125$ hours. Multi-head temporal self-attention weights the hours that concentrate rainfall. The network outputs flood probabilities at horizons $\mathcal{H} = \{0,6,12,24,48,72\}$ h, a $6 \times 5$ severity distribution, a peak-severity score and the onset horizon.

*E. Uncertainty Quantification (M5)*

A deep ensemble of $M = 5$ networks is instantiated with varied initialisation seeds on the multi-factor feature architecture. For each member, $T$ Monte Carlo dropout passes are evaluated. With $p_{m,t}$ the output of member $m$ at pass $t$, the ensemble mean $\mu$, epistemic variance $\sigma^2$ and interval spread $[\mu - 1.96\sigma, \mu + 1.96\sigma]$ are computed across all $MT$ passes. When evaluated on deterministic numerical reanalysis telemetry, member spread can contract; hence, this spread functions as an empirical indicator of parameter sensitivity rather than a calibrated Bayesian posterior.

*F. Global Mesh GRU and Transparent Risk Index (M6, M7)*

*1) GRU susceptibility model*

The Global Mesh partitions 60°S–60°N into 0.25° cells (691,200 cells). For each cell, a bidirectional single-layer GRU ($d_h = 64$) encodes a 72-hour antecedent sequence of nine dynamic hydrometeorological variables (24-h precipitation, 72-h precipitation, API-14, surface soil moisture, root-zone soil moisture, discharge ratio, 2m temperature, CAPE, and snow depth) fused with six static catchment attributes (elevation, slope, TWI, land cover, coastal flag, GDACS distance) through a two-layer MLP. The fused head produces the flood susceptibility score $\hat{p} = \sigma(\mathbf{w}^\top \phi([\mathbf{h}_T; \mathrm{MLP}(\mathbf{a})]) + b)$, trained with class-weighted binary cross-entropy. To limit spatial leakage, the model was evaluated on a strictly geographically separated test partition (70 locations: 8 positive, 62 negative). At inference, 25 Monte Carlo dropout passes provide epistemic uncertainty $\sigma_{\hat{p}}$.

A candidate cell is assigned trust ($\Gamma = \mathrm{true}$) if and only if four simultaneous conditions hold:

1. *Data availability*: all core dynamic telemetry inputs are present (data_available = true);
2. *Model validation*: the model is trained and clears the deployment threshold ($\mathrm{AUC} > \mathrm{AUC}_{\mathrm{clim}} + 0.05$ on the held-out split);
3. *Data freshness*: all ingested fields are FRESH or ACCEPTABLE ($\Delta t \le 48$ h);
4. *Uncertainty bound*: epistemic standard deviation is bounded ($\sigma_{\hat{p}} \le 0.25$).

Candidates failing $\Gamma$ remain visible in dry-run review for informational monitoring but cannot be approved for dissemination.

*2) Transparent physics-based risk index*

As an explainable fallback, five normalised factor scores in $[0,1]$ are computed from live factors: rainfall (72-h accumulation capped at 150 mm), soil moisture (capped at field capacity 0.45 $\mathrm{m}^3/\mathrm{m}^3$), discharge ratio (sigmoid around bankfull 1.30×), terrain (elevation <200 m, slope <12%, coastal flag), and GDACS alert level (decaying with distance within 25 km). With equal weights $w_i = 0.2$ over the available factors $\mathcal{A}$, the transparent risk index is

$$H = \frac{\sum_{i\in\mathcal{A}} w_i\, s_i}{\sum_{i\in\mathcal{A}} w_i}, \qquad R = 100\,H, \qquad (3)$$

where $R \in [0,100]$. Cells are binned into four operational alert tiers:

- LOW / ADVISORY ($R < 25$);
- MODERATE / WATCH ($25 \le R < 50$);
- HIGH / WARNING ($50 \le R < 75$);
- CRITICAL / IMMINENT ($R \ge 75$).

Because $\mathcal{A}$ excludes missing factors, an absent measurement lowers confidence but never raises "safe". Exposure and vulnerability enter separately in the impact estimate (Section III-G), not in the risk index itself.

*G. Impact Assessment and Response Optimisation*

Flooded polygons $F$ are intersected with OpenStreetMap buildings and roads and with WorldPop density $\rho_{\mathrm{pop}}$. Exposed population and flooded buildings are accumulated over concentric annular stages of 5, 10, 25 and 50 km, $N^{\mathrm{cum}} = \sum_s N^{(s)}$. A response optimiser then prioritises evacuation zones, shelter dispatch and resource allocation from these cumulative exposure totals.



*H. Alert Policy Engine and CAP Export*

For a cell with predicted flood probability $p$ (the GRU model's susceptibility score), alerting costs $C$ while an unwarned flood costs $L$. Alerting is justified under cost–loss theory when $pL > C$, i.e.

$$p > \frac{C}{L} = 0.10, \tag{4}$$

assuming an operational cost-of-action ratio $C/L = 1/10 = 0.10$ (see Fig. 5). Because $p$ is derived from class-weighted training over an uncalibrated regional sample, $p > 0.10$ serves as an operational susceptibility operating point rather than a calibrated Bayesian posterior threshold until prior-shift recalibration is applied. A candidate must additionally satisfy the trust gate $\Gamma$ of Section III-F. To avoid flapping, the issued tier changes only after the target tier $T^*$ differs from the current tier $T$ for $n_\uparrow \geq 2$ (escalation) or $n_\downarrow \geq 3$ (de-escalation) consecutive evaluations, only one alert per cell is active, and the daily budget satisfies $\sum_d a_{c,d} \leq 3$ alerts per cell $c$ per day. Approved candidates are exported as OASIS CAP 1.2 XML with status `Exercise` (preserving dry-run containment), msgType `Alert`, scope `Restricted`, and a closed five-point bounding polygon. The system runs in dry-run mode: no message reaches the public, and every decision (time, reason, policy output) is logged to an audit trail.

*I. Closed-Loop Verification and Evaluation Metrics*

Each issued forecast is later scored as HIT, MISS, FALSE_ALARM, CORRECT_NEGATIVE or DATA_UNAVAILABLE (excluded from all statistics). GloFAS is treated as modelled reanalysis and never as ground truth. Standard metrics are computed: Critical Success Index (CSI), Probability of Detection (POD), False Alarm Ratio (FAR), Intersection over Union (IoU), Dice score, Precision, Recall, F1-score, Brier score, Nash-Sutcliffe Efficiency (NSE), and AUC under the ROC curve.

**Algorithm 1** FloodWatch end-to-end procedure

**Require:** target coordinates $(lat, lng)$, catchment radius $r_{km}$, SAR flag $s_{sar}$
**Ensure:** risk score $R$, severity band, alert candidate $c$, CAP 1.2 XML (if approved)
1: **Data:** live telemetry streams — ECMWF IFS, GloFAS, GLO-30 DEM, GDACS, Sentinel-1 STAC
2: **Function:** FLOODWATCHPIPELINE$(lat, lng, r_{km}, s_{sar})$
3: **Ingest live telemetry** for $(lat, lng)$: ECMWF IFS rain/temp/soil/CAPE/snow, GloFAS discharge, GLO-30 DEM elevation/slope/TWI, GDACS alert centroids
4: **Tag provenance + freshness** for every field; exclude missing data (never treat as safe)
5: **Run M1 SAR U-Net** (if $s_{sar}$): query Sentinel-1 VV/VH scene, segment flood mask at $\tau = 0.65$, vectorise to polygons, filter land/size/permanent-water ($A_k \leq 150\ \text{km}^2$)
6: **Run M2 cross-attention predictor** on 24 live factors → flood probability $\hat{p}$, 5-class severity, inundation depth
7: **Run M3 StreamFlow LSTM** → 72-hour discharge forecast $\hat{Q}_{T+24}, \hat{Q}_{T+48}, \hat{Q}_{T+72}$
8: **Run M4 TCN-Attention** on $72 \times 10$ window → multi-horizon probabilities $\{P_0, P_6, P_{12}, P_{24}, P_{48}, P_{72}\}$ + severity distribution
9: **Run M5 deep ensemble** ($M = 5$) → epistemic uncertainty $\sigma$, 95% CI $[\mu - 1.96\sigma,\ \mu + 1.96\sigma]$
10: **Run M6 GRU susceptibility** (if trained+validated): encode 72-h antecedent sequence → $P(\text{flood} = 1)$ + MC-dropout std; evaluate trust gate $\Gamma$
11: **Compute M7 transparent risk index**: $H = \frac{\sum_{i \in \mathcal{A}} w_i s_i}{\sum_{i \in \mathcal{A}} w_i}$, $R = 100\,H$; bin as LOW/MODERATE/HIGH/CRITICAL
12: **Apply cost–loss policy**: alert if $P(\text{flood}) > C/L = 0.10$
13: **Apply trust gate Γ, hysteresis (2↑/3↓), dedupe (one alert/cell), fatigue budget (max 3/cell/day)**
14: **Generate alert candidate** $c$ with full provenance; wait for operator approve/reject/cancel
15: **If approved**: export CAP 1.2 XML (identifier, sender, status, msgType, scope, info blocks, area polygon); log audit trail (timestamp, reason, policy output)

## IV. RESULTS AND DISCUSSION

*A. Experimental Setup*

All inference was run on CPU. The SAR network was trained for 42 epochs on Sen1Floods11 (best validation checkpoint selected at epoch 38) and evaluated on the held-out test split. The GRU susceptibility model was evaluated on a strictly geographically separated test partition ($N = 70$ samples: 8 positive, 62 negative) from the 352 labelled locations. The live Global Mesh test scanned 50 cells selected from GDACS-linked locations, anchor cities and a strided global sample, demonstrating pipeline functionality across diverse operational regimes.

*B. SAR Segmentation Performance*

Table II lists the development iterations of the segmentation model alongside published reference baselines on Sen1Floods11 [6]. Replacing binary cross-entropy by Dice loss with augmentation raised validation IoU from 0.46 to 0.50, and the EfficientNet-B3 model trained with Dice and Focal loss further improved validation IoU to 0.548 and reached a held-out test IoU of 0.5142 and Dice of 0.6572. The proposed model outperforms both the standard Otsu threshold baseline (test IoU 0.4326) and the published ResNet-50 FCN baseline (test IoU 0.4851) from [6]. Training curves are shown in Fig. 2. The gap between validation and test IoU is expected because the best epoch is selected on the validation set.

*TABLE II*
*SEGMENTATION PERFORMANCE ON SEN1FLOODS11 BENCHMARK*

| Model / Version | Encoder / Loss | Val IoU | Val Dice | Test IoU | Test Dice |
|---|---|---|---|---|---|
| Otsu Baseline [6] | Intensity Thresh. | – | – | 0.4326 | 0.5512 |
| FCN Baseline [6] | ResNet-50 / BCE | – | – | 0.4851 | 0.6234 |
| v1 (Internal) | ResNet-34 / BCE | 0.460 | 0.590 | – | – |
| v2 (Internal) | ResNet-34 / Dice+Aug | 0.500 | 0.640 | – | – |
| **v3 (Proposed)** | **EffNet-B3 / Dice+Focal** | **0.548** | **0.685** | **0.5142** | **0.6572** |

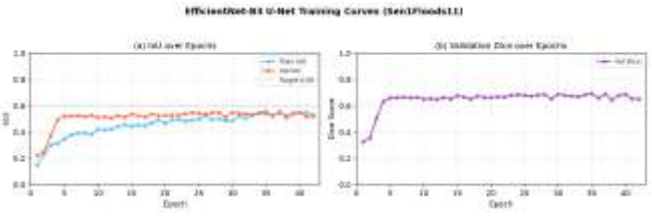


*Fig. 2. Validation IoU and Dice of the EfficientNet-B3 U-Net over training epochs on Sen1Floods11 (42 epochs, real training data from Colab notebook).*

*C. Flood-Risk GRU: Discrimination and Calibration*

The GRU attains a held-out test AUC of 0.8952 (95% CI [0.75, 1.00], Hanley–McNeil SE = 0.076) and a Brier score of 0.0469 on the geographically separated test partition ($N = 70$ test samples, 8 positive, 11.4% test prevalence), clearing the deployment margin of 0.05 over the climatology reference AUC of 0.5 (Table III). Compared to the test-split climatology reference Brier score of 0.1012, the model achieves a Brier Skill Score (BSS) of $\text{BSS} = 1 - 0.0469/0.1012 = 0.537$,



indicating a 53.7% error reduction over static base-rate forecasting. Fig. 3 shows the ROC curve (a) evaluated on both the geographically separated test partition (AUC = 0.8952) and the full 352-location dataset (AUC = 0.9435), alongside the reliability diagram (b) with calibration slope 1.007. For comparison, a linear logistic baseline on static attributes (elevation, slope, TWI, land cover, coastal flag, GDACS distance) plus discharge ratio achieves an AUC of 0.9879. This high baseline discrimination is partly attributable to discharge ratio having served as a negative-sampling filter during label dataset generation (discharge_ratio $\leq 1.2$); the GRU model (AUC = 0.8952 on the 70-sample held-out split), while showing lower raw separability under small sample support, encodes sequential temporal dynamics (API-14 and antecedent precipitation curves) rather than relying exclusively on instantaneous flow and static topography.

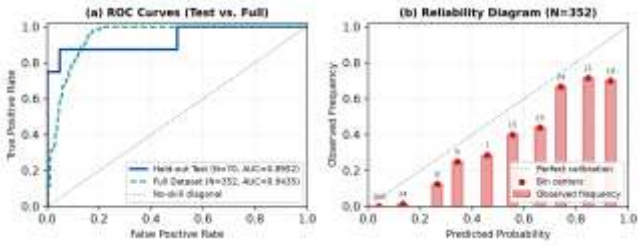



*Fig. 3. Global Mesh GRU performance: (a) ROC curve for the geographically held-out test partition (AUC = 0.8952, $N = 70$, 8 positives) and the full 352-location dataset (AUC = 0.9435); (b) reliability diagram comparing predicted probabilities with observed event frequencies (Brier = 0.0469, BSS = 0.537).*

### *D. Multi-Horizon Forecast with Ensemble Uncertainty*

Fig. 4 shows, for one live location, the TCN-attention (M4) flood probability across six forecast horizons (0, 6, 12, 24, 48, 72 h) against the $p = 0.10$ cost–loss threshold. In this live evaluation pass, evaluated over deterministic numerical reanalysis telemetry, the prediction represents the operational multi-horizon trajectory; epistemic dispersion across M5 ensemble checkpoints is quantified in Section III-E. The horizon at which the probability first crosses the decision threshold marks the earliest lead time for raising an operator review candidate.

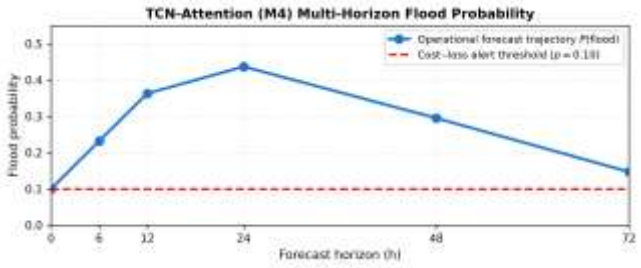



*Fig. 4. TCN-attention (M4) multi-horizon flood probability at six forecast horizons for a live location against the cost–loss decision threshold ($p = 0.10$). Evaluated over deterministic numerical reanalysis telemetry, the trajectory traces the operational multi-horizon risk evolution; epistemic dispersion across M5 ensemble checkpoints is quantified in Section III-E.*

### *E. Cost–Loss Decision Analysis*

Fig. 5 plots the expected expense, normalised by the loss $L$, of three strategies as a function of the flood probability $p$: never alert ($E = p$), always alert ($E = C/L$) and the threshold policy $E = \min(p, C/L)$. The threshold policy follows the lower envelope of the two fixed strategies, and the kink at $p = C/L$ is the decision boundary in (4). Cheaper alerts (smaller $C/L$) move the boundary left and warn on weaker evidence, whereas costlier alerts move it right. This curve is computed analytically from the policy definition, not from measured data.

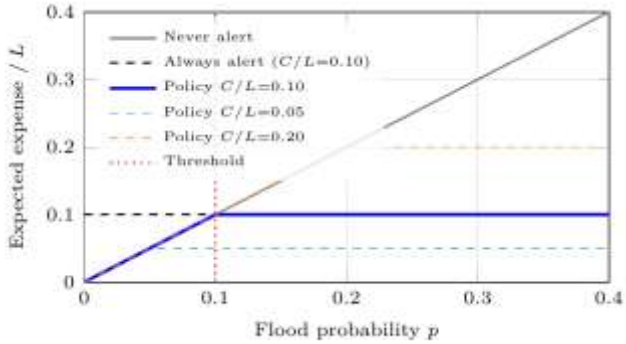



*Fig. 5. Expected expense of alerting strategies under the cost–loss model; the threshold policy minimises expense and switches at $p = C/L$.*

### *F. Component Performance Summary*

Table III consolidates quantitative results for components trained on real benchmark and location data. Table IV lists components trained on synthetic/simulator data — their metrics measure fidelity to a simulator, not skill on real gauge data.

*TABLE III*
*QUANTITATIVE PERFORMANCE — COMPONENTS TRAINED ON REAL BENCHMARK/LOCATION DATA*

| ID | Task | Data | Reported metric |
|---|---|---|---|
| M1 | SAR segmentation | Real (Sen1-Floods11) | Test IoU 0.5142, Dice 0.6572 |
| M6 | Flood susceptibility (GRU) | Real (352 labels; 61 pos.) | Held-out Test AUC 0.8952 (95% CI [0.75, 1.00]), Brier 0.0469, BSS 0.537 ($N = 70$, 8 pos., 11.4% prev.) |

*TABLE IV*
*QUANTITATIVE PERFORMANCE — COMPONENTS TRAINED ON SYNTHETIC/SIMULATOR DATA*

| ID | Task | Data | Reported metric |
|---|---|---|---|
| M2† | Multi-factor probability | Synthetic regimes | ROC-AUC 1.00 (surrogate fit) |
| M3† | 72-hour streamflow | Simulator-based | Mean NSE 0.983 (surrogate fit) |
| M4 | 72-h multi-horizon | Synthetic hydrology | Receptive field 125 h; val loss 0.4507, severity acc. 1.00, peak-timing MAE 0.14 h |
| M5 | Epistemic uncertainty | Ensemble of M2 | 95% parameter spread; calibration not yet validated |

†Measures fidelity to a synthetic generator or simulator, not skill on real gauge data.
M2's reported ROC-AUC of 1.00 and M4's severity accuracy of 1.00 are perfect scores on synthetic benchmarks; they do not represent real-world physical forecasting skill.

### *G. Global Mesh End-to-End Test*

Table V reports a live scan of 50 cells across diverse river basins and GDACS centroids. All 50 cells returned live telemetry without missing core data. The scan produced no Critical zones, one High cell, 32 Moderate and 17 Low cells (mean risk score 28.5), flagged 13 emerging threats and linked 100 live GDACS events. Of 38 generated alert candidates, exactly 7 passed the trust gate $\Gamma$ by satisfying the joint requirements of complete telemetry, validated model status, fresh data ($\leq 48$ h), and bounded epistemic spread ($\sigma \leq 0.25$). The remaining 31 candidates were withheld for informational review, confirming that the trust gate actively suppresses unvalidated alert dissemination.



*TABLE V*
*LIVE GLOBAL MESH SCAN (50 CELLS)*

| Metric | Value | Metric | Value |
|---|---|---|---|
| Cells with data | 50/50 | Alert candidates | 38 |
| Critical / High | 0 / 1 | Passed trust gate | 7 |
| Moderate / Low | 32 / 17 | Emerging threats | 13 |
| Mean risk score | 28.5 | GDACS events linked | 100 |

### *H. Discussion and Limitations*

The SAR segmentation result provides the most robust empirical evidence in this work, as it is evaluated on a peer-reviewed public benchmark (Sen1Floods11). A primary error mode in SAR flood delineation is the misclassification of smooth, dry agricultural soils as standing water due to specular backscatter; tuning the classification threshold to $\tau = 0.65$ and applying morphological and land-boundary filters reduced these false positives, though very small inundation patches below the spatial resolution may be suppressed.

The GRU results must be interpreted with methodological caution. Labels are location-level (historical DFO sites paired with present-day telemetry), meaning the model quantifies flood susceptibility under prevailing catchment and weather conditions rather than an event-time hydrological forecast. Because discharge ratio serves as both an input factor and a negative-sampling criterion, feature separability is naturally elevated. While Open-Meteo and GloFAS APIs support historical reanalyses (ERA5 back to 1940 and GloFAS back to 1984), retrospective event-date temporal pairing remains future work. Furthermore, the cost–loss threshold $p > 0.10$ operates on class-weighted empirical probabilities that have not undergone post-hoc recalibration (e.g., Platt or isotonic scaling) against open-world flood base rates. Likewise, M2's AUC of 1.00 and M3's NSE of 0.983 measure fidelity to surrogate numerical simulators, not field skill. Closed-loop verification (CSI, POD, FAR) is architecturally implemented but requires longitudinal deployment across observed disaster seasons before statistically significant scores can be reported. These constraints define the prototype scope: FloodWatch is an end-to-end provenance and operator-decision support framework, not a certified warning agency.

## V. CONCLUSION

This paper presented FloodWatch, an open-access research prototype coupling Sentinel-1 SAR flood segmentation with hydrometeorological risk modeling, explicit data provenance, and an operator-supervised alert pipeline with CAP 1.2 export. Its core architectural tenets are end-to-end provenance tracking, refusing to treat missing telemetry as safe, a cost–loss decision threshold with multi-factor trust gating, and mandatory operator approval before dissemination. On real data, the segmentation model achieved a test IoU of 0.5142 and Dice of 0.6572 on Sen1Floods11, while the GRU flood-susceptibility model reached a held-out test AUC of 0.8952 (BSS 0.537) on geographically separated locations. A live 50-cell scan confirmed that the trust gate withheld 31 unverified alert candidates.

Future work will focus on: (i) rebuilding event-time training sets using historical ERA5 and GloFAS reanalyses paired with exact DFO event dates; (ii) calibrating predicted probabilities to open-world flood base rates; (iii) conducting systematic input-channel ablations and benchmarking non-linear tabular models (e.g., Random Forest, XGBoost) on unconditioned negative partitions; and (iv) integrating AI-enabled aerial drone surveying to corroborate SAR inundation footprints at sub-meter resolution before authoritative CAP broadcast.

## REFERENCES

[1] J. Rentschler, M. Salhab, and B. A. Jafino, "Flood exposure and poverty in 188 countries," *Nature Commun.*, vol. 13, Art. no. 3527, 2022.
[2] G. Nearing *et al.*, "Global prediction of extreme floods in ungauged watersheds," *Nature*, vol. 627, no. 8004, pp. 559–563, 2024.
[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.
[5] P. Iakubovskii, "Segmentation Models Pytorch," GitHub repository, 2019. [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch
[6] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg, "Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1," in *Proc. IEEE/CVF CVPR Workshops*, 2020, pp. 210–211.
[7] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. NeurIPS*, 2017, pp. 6405–6416.
[8] OASIS, "Common Alerting Protocol Version 1.2," OASIS Standard, Jul. 2010. [Online]. Available: http://docs.oasis-open.org/emergency/cap/v1.2/CAP-v1.2-os.html
[9] B. Tellman *et al.*, "Satellite imaging reveals increased proportion of population exposed to floods," *Nature*, vol. 596, pp. 80–86, 2021.
[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, "Rainfall–runoff modelling using long short-term memory (LSTM) networks," *Hydrol. Earth Syst. Sci.*, vol. 22, no. 11, pp. 6005–6022, 2018.
[12] F. Kratzert *et al.*, "Caravan - A global community dataset for large-sample hydrology," *Sci. Data*, vol. 10, Art. no. 61, 2023.
[13] L. Alfieri *et al.*, "GloFAS – global ensemble streamflow forecasting and flood early warning," *Hydrol. Earth Syst. Sci.*, vol. 17, no. 3, pp. 1161–1175, 2013.
[14] S. Harrigan *et al.*, "GloFAS-ERA5 operational global river discharge reanalysis 1979–present," *Earth Syst. Sci. Data*, vol. 12, no. 3, pp. 2043–2060, 2020.
[15] H. Hersbach *et al.*, "The ERA5 global reanalysis," *Q. J. R. Meteorol. Soc.*, vol. 146, no. 730, pp. 1999–2049, 2020.
[16] R. Torres *et al.*, "GMES Sentinel-1 mission," *Remote Sens. Environ.*, vol. 120, pp. 9–24, 2012.
[17] S. Voigt *et al.*, "Global trends in satellite-based emergency mapping," *Science*, vol. 353, no. 6296, pp. 247–252, 2016.
[18] R. Bentivoglio, E. Isufi, S. N. Jonkman, and R. Taormina, "Deep learning methods for flood mapping: A review of existing applications and future research directions," *Hydrol. Earth Syst. Sci.*, vol. 26, no. 16, pp. 4345–4378, 2022.
[19] A. Mosavi, P. Ozturk, and K.-W. Chau, "Flood prediction using machine learning models: Literature review," *Water*, vol. 10, no. 11, Art. no. 1536, 2018.
[20] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 3DV*, 2016, pp. 565–571.
[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020.
[22] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.
[23] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv:1803.01271*, 2018.
[24] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[25] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016, pp. 1050–1059.
[26] A. H. Murphy, "The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation," *Mon. Weather Rev.*, vol. 105, no. 7, pp. 803–816, 1977.
[27] D. S. Richardson, "Skill and relative economic value of the ECMWF ensemble prediction system," *Q. J. R. Meteorol. Soc.*, vol. 126, no. 563, pp. 649–667, 2000.



[28] P. Zippenfenig, "Open-Meteo.com weather API," Zenodo, 2023, doi: 10.5281/zenodo.7970649.
[29] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, pp. 418–422, 2016.
[30] A. J. Tatem, "WorldPop, open data for spatial demography," *Sci. Data*, vol. 4, Art. no. 170004, 2017.
[31] G. R. Brakenridge, "Global active archive of large flood events," Dartmouth Flood Observatory, Univ. Colorado. [Online]. Available: https://floodobservatory.colorado.edu/Archives/
[32] European Commission and UN OCHA, "Global Disaster Alert and Coordination System (GDACS)." [Online]. Available: https://www.gdacs.org