A Bidirectional GRU Network with Temporal Attention for Flood Prediction Using Time-Series Sensor Data

Abderahman Yakoub Atoui
École supérieure en Sciences et Technologies de l'Informatique et du Numérique
RN 75, Amizour 06300, Bejaia, Algérie
a_atoui@estin.dz

Belkacem Messaoudi
École supérieure en Sciences et Technologies de l'Informatique et du Numérique
RN 75, Amizour 06300, Bejaia, Algérie
b_messaoudi@estin.dz

Youssef Elmir
Laboratoire LITAN
École supérieure en Sciences et Technologies de l'Informatique et du Numérique
RN 75, Amizour 06300, Bejaia, Algérie
elmir@estin.dz

Abbes Amira
Faculty of Science & Engineering (FSE)
University of Wolverhampton
Wolverhampton, United Kingdom
A.amira@wlv.ac.uk

***Abstract*—Predicting floods from meteorological time-series data is a challenging classification problem due to imbalanced classes and long-term dependencies. In this paper, we propose a Bidirectional Gated Recurrent Unit (BiGRU) network with temporal attention for binary daily flood prediction. This approach uses a lookback period of 60 days with 62 engineered features combining local precipitation, temperature, humidity, past flooding, and upstream catchment signals from the Cherrapunji headwater and Barak River sub-basin. We evaluated the proposed approach on 25 years (2000–2024) of daily data collected in Sylhet, Bangladesh. Our method achieves a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.9575 and an F1-score of 0.8291, significantly lowering the number of false alarms compared to previous recurrent baseline methods. The use of temporal attention improves interpretability by identifying the periods leading up to a flood event that most influence the prediction. These results confirm that sequential deep learning with multi-source catchment features is an effective approach to forecasting floods from environmental time-series data.**

***Index Terms*—Flood Prediction, BiGRU, Temporal Attention, Time-Series Forecasting, Deep Learning, Environmental Sensors, Disaster Management**

I. Introduction

Floods are among the most common and economically costly natural disasters, causing about one-third of all disaster-related losses worldwide [1]. Reliable flood warnings depend on the ability to model both local rainfall and upstream hydrological signals that can predict local floods 24–72 hours in advance. Physically-based hydrological models, currently used in practice, depend on extensive stream-gauge networks, which are rarely available [2]. Data-driven modelling provides an attractive alternative; yet most existing studies use only local meteorological station information, ignoring upstream catchment processes, which typically form the primary cause of downstream floods [3], [4]. Deep learning models, especially recurrent neural networks, combined with attention mechanisms [5], [6], offer novel ways of recognising patterns in multistep time series. Separately, large language models (LLMs) have demonstrated broad domain knowledge that could, in principle, be exploited for zero-shot hydrological reasoning, motivating a comparative evaluation.

This paper makes three contributions:

1) A feature-engineering pipeline combining 62 features across twelve categories: core meteorological, precipitation lags, rolling statistical aggregates, trend & anomaly indices, spell counters, humidity & temperature lags, the Standardised Precipitation Index (SPI-14/SPI-30), seasonal encodings, flood-history lags, upstream Cherrapunji catchment signals, Barak Valley catchment signals, and water-balance indicators.
2) A BiGRU architecture with temporal attention for sequential flood prediction using long-term temporal dependencies.
3) An experimental evaluation demonstrating the effectiveness of recurrent deep learning models for flood forecasting on imbalanced environmental datasets, including a deliberately constrained zero-shot benchmark against a large language model (Llama-3.3-70B) that contrasts sequence-based numerical modelling with single-snapshot textual reasoning, rather than characterising the maximum forecasting capability of LLMs.

The remainder of this paper is organised as follows. Section II reviews related work on deep learning for flood

979-8-3195-3958-8/26/$31.00 ©2026 IEEE

forecasting and imbalance handling. Section III describes the dataset, feature engineering pipeline, model architecture, and training protocol. Section IV presents and discusses the experimental results, including comparisons with baseline models and LLMs. Section V concludes the paper and outlines directions for future work.

II. Related Work

A. Deep Learning for Flood Forecasting

Recent research has shown that Long Short-Term Memory (LSTM) networks have superiority over conceptual rainfall-runoff models in 241 basins of the United States, and therefore deep recurrent networks have been established as a viable option in hydrology [4]. A review of machine learning techniques for flood forecasting found that neural networks and gradient-boosted trees were the best predictors, but noted the lack of consideration for upstream information in current datasets [3]. Selva Jeba et al. found that a GRU-based model outperformed LSTMs for predicting rainfall from raw meteorological time-series data, achieving lower Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) on an Indian dataset spanning 52 years [7]. However, this method used only raw meteorological data, without feature extraction, attention mechanisms, or consideration of class imbalance – problems addressed by this study.

Attention mechanisms in time-series have proven to increase interpretability of the models [6], where attributing higher importance to physically meaningful time steps (such as heavy rain days upstream) results in transparent models that can be used in real-world hydrology. Class imbalance is a common problem in flood datasets. While the Synthetic Minority Oversampling Technique (SMOTE) [8] is commonly applied, it could generate unrealistic synthetic sequences when interpolated in high-dimensional windows. Focal Loss [9] addresses imbalance at the loss level without modifying the training distribution, avoiding this pitfall.

To our knowledge, no study has benchmarked LLMs as zero-shot flood predictors against sequence-trained deep models on a balanced subsample of the same test set.

B. Operational Systems and the Case for Binary Classification

Established operational flood forecasting systems typically target one of three output types: (i) *river-stage regression*, predicting continuous water-level values at gauged cross-sections; (ii) *inundation mapping*, delineating spatially explicit flood extents using hydrodynamic models or Earth Observation (EO)-based classifiers; or (iii) *binary early warning*, issuing a flood / no-flood alert from meteorological precursors. While river-stage models such as those deployed by national meteorological agencies achieve high physical accuracy, they require dense stream-gauge networks and calibrated hydraulic models that are unavailable in many flood-prone developing regions [2]. Inundation mapping systems similarly depend on high-resolution digital elevation models (DEMs) and satellite revisit cycles that introduce latency incompatible with 24–72-hour advance warning requirements.

TABLE I
Comparison of Related Flood Forecasting Approaches

Study	Model	Output	Features	Imbal.	Attn.
Kratzert et al. [4]	LSTM	River stage	Meteoro-logical	No	No
Mosavi et al. [3]	Various ML	Flood risk	Meteoro-logical	Partial	No
Selva Jeba et al. [7]	RNN-GRU	Rainfall (reg.)	Raw weather	No	No
Song et al. [6]	LSTM+Attn	Water level	Hydro. sensors	No	Yes
Proposed	**BiGRU+ Attn**	**Binary alert**	**62 engi-neered**	**Focal Loss**	**Yes**

RNN: Recurrent Neural Network; ML: Machine Learning; Imbal.: imbalance handling; Attn.: Attention; reg.: regression.

Binary classification offers a practical and deployable alternative in data-scarce environments: it transforms the forecasting problem into a decision-support output directly actionable by civil protection authorities, without requiring continuous hydrological measurements. This design choice is consistent with the operational context of Sylhet, Bangladesh, where stream-gauge infrastructure is limited and timely binary alerts have demonstrated value for evacuation planning.

Table I summarises representative related work across these three paradigms and positions the proposed system within the landscape.

III. Methodology

A. Dataset

The dataset consists of daily meteorological measurements and binary flood/no-flood labels for the period 2000–2024 (9,132 days) in a flood-prone sub-region of South Asia, namely the Sylhet district in Bangladesh. Flood days account for 11.7% of all observations ($n = 1{,}068$), indicating significant class imbalance. The local meteorological measurements were sourced from the Visual Crossing Weather Query Builder [10], providing daily precipitation, minimum/maximum/mean temperature, and relative humidity. External upstream signals were retrieved from the Open-Meteo ERA5 (European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5) archive [11]: upstream precipitation from the Cherrapunji headwater region (25.27°N, 91.72°E) and precipitation from the Barak River sub-basin, both of which contribute to downstream flooding in Sylhet. The flood occurrence labels were derived from Bangladesh Flood Forecasting and Warning Centre (FFWC) flood bulletins.

B. Feature Engineering

In total, 62 features were constructed and categorised into twelve categories (see Table II). Local meteorological features come from Visual Crossing, while 13 upstream signals (Cherrapunji, Barak Valley, and water balance) were fetched from or derived using the Open-Meteo ERA5 archive. All features derived from past observations are shifted by at least one day (`shift(1)`) to prevent data leakage. Features were normalised to $[0, 1]$ using Min-Max scaling fitted exclusively on the training set.

TABLE II
FEATURE CATEGORIES USED BY THE PROPOSED MODEL

Category	N	Primary Role
Core meteorological	5	Raw daily weather signal
Precipitation lags	7	Short-term rainfall trajectory
Rolling statistical	12	Multi-scale moisture accumulation
Trend & anomaly	4	Rainfall acceleration & climatological context
Spell counters	2	Soil saturation proxy
Humidity & temp. lags	6	Atmospheric moisture state
SPI-14 / SPI-30	2	Anomalous precipitation detection
Seasonal encoding	6	Climatological flood risk window
Flood history lags	5	Temporal autocorrelation of floods
Upstream Cherrapunji	8	1–3 day catchment pulse
Barak Valley	2	Secondary upstream catchment
Water balance	3	Net moisture accumulation (precipitation − ET_0)
Total	**62**	—

SPI: Standardised Precipitation Index (defined above); ET_0: reference evapotranspiration. Upstream Cherrapunji, Barak Valley, and Water balance (8+2+3=13 features) are drawn from the ERA5 archive.

C. Model Architecture

The proposed model, illustrated in Fig. 1, consists of three stacked BiGRU layers followed by a temporal attention mechanism and a dense classification head.

The encoder processes an input sequence of $T = 60$ time steps with engineered environmental features. The three BiGRU layers contain 128, 96, and 64 hidden units per direction, respectively.

Dropout and recurrent dropout are applied between layers to reduce overfitting and improve generalisation.

The temporal attention layer learns the importance of each time step within the sequence. Attention scores are normalised using a softmax function to produce attention weights α_t.

The context vector is computed as

$$c = \sum_{t=1}^{T} \alpha_t h_t, \tag{1}$$

where h_t represents the hidden representation at time step t, and α_t denotes the corresponding attention weight.

Finally, the context vector is passed through three dense layers (128-64-32) with Rectified Linear Unit (ReLU) activation, Batch Normalisation, and Dropout before the final sigmoid output layer produces the flood probability.

D. Training Protocol

Five independent models are trained with different random seeds (42, 123, 7, 2024, 314); their probabilities are averaged to form the ensemble prediction. Optimisation uses the Adam (Adaptive Moment Estimation) optimiser. Table III summarises the key hyperparameters.

E. LLM Baseline Setup

We emphasise that general-purpose LLMs are not purpose-built spatiotemporal time-series forecasting models and are not proposed here as a competitive alternative to the recurrent architecture. The primary quantitative baselines for this study are the conventional models in Table IV (majority class and rolling-sum logistic regression). The LLM experiment is included as an auxiliary, exploratory benchmark for three specific reasons: (i) to examine whether zero-shot textual reasoning over raw daily weather values can support binary flood classification without task-specific training; (ii) to contrast single-snapshot textual reasoning with the sequence-trained temporal modelling used by the BiGRU; and (iii) to probe, rather than assume, whether the broad domain knowledge captured by LLMs transfers to a structured, sequential hydrological task. It is not intended to establish LLMs as state-of-the-art time-series forecasters, nor to substitute for comparison against dedicated sequence architectures.

Llama-3.3-70B [12] was queried via the Groq Application Programming Interface (API) as a zero-shot flood predictor. For each day in a balanced 120-day subsample (60 flood / 60 no-flood) drawn from the test set, a structured text prompt was constructed containing the date and the raw daily precipitation, temperature, and humidity values for that day, together with a description of the task, and the model was asked to return a binary flood / no-flood answer. This design deliberately contrasts single-snapshot textual reasoning over raw daily values with the BiGRU's access to a full 60-day, 62-feature engineered sequence. We explicitly note that this comparison is asymmetric in input context: the BiGRU is given the complete 60-day, 62-feature sequence, whereas the LLM receives only a single day's worth of three raw weather values with no engineered features (no lags, rolling sums, SPI, or flood history). The benchmark is therefore designed to evaluate zero-shot, snapshot-based textual reasoning over minimal raw inputs under this specific prompt design, not the maximum forecasting capability that LLMs could achieve under richer input features, temporal context, few-shot prompting, or fine-tuning; the results should be interpreted with this constraint in mind.

TABLE III
TRAINING HYPERPARAMETERS

Parameter	Value
Loss function	Focal Loss (γ=2.0, α=0.78)
Optimiser	Adam (η=10^{-3})
Learning-rate schedule	Cosine Annealing + Restarts (T_0=20, T_{mult}=2)
Class weight cap	10× minority class
Epochs (max)	100; EarlyStopping patience = 20
Batch size	64
Train / Val / Test	70% / 15% / 15% (chronological)
Decision threshold	Max-F1 on validation set (= 0.539)

IV. RESULTS AND DISCUSSION

A. Overall Test Set Performance

Table IV compares the BiGRU ensemble against a naive majority baseline, a logistic regression (LR) on rolling sums, and two earlier BiGRU development iterations. The majority-class and rolling-sum logistic regression models constitute the primary conventional baselines for this study; the BiGRU rows

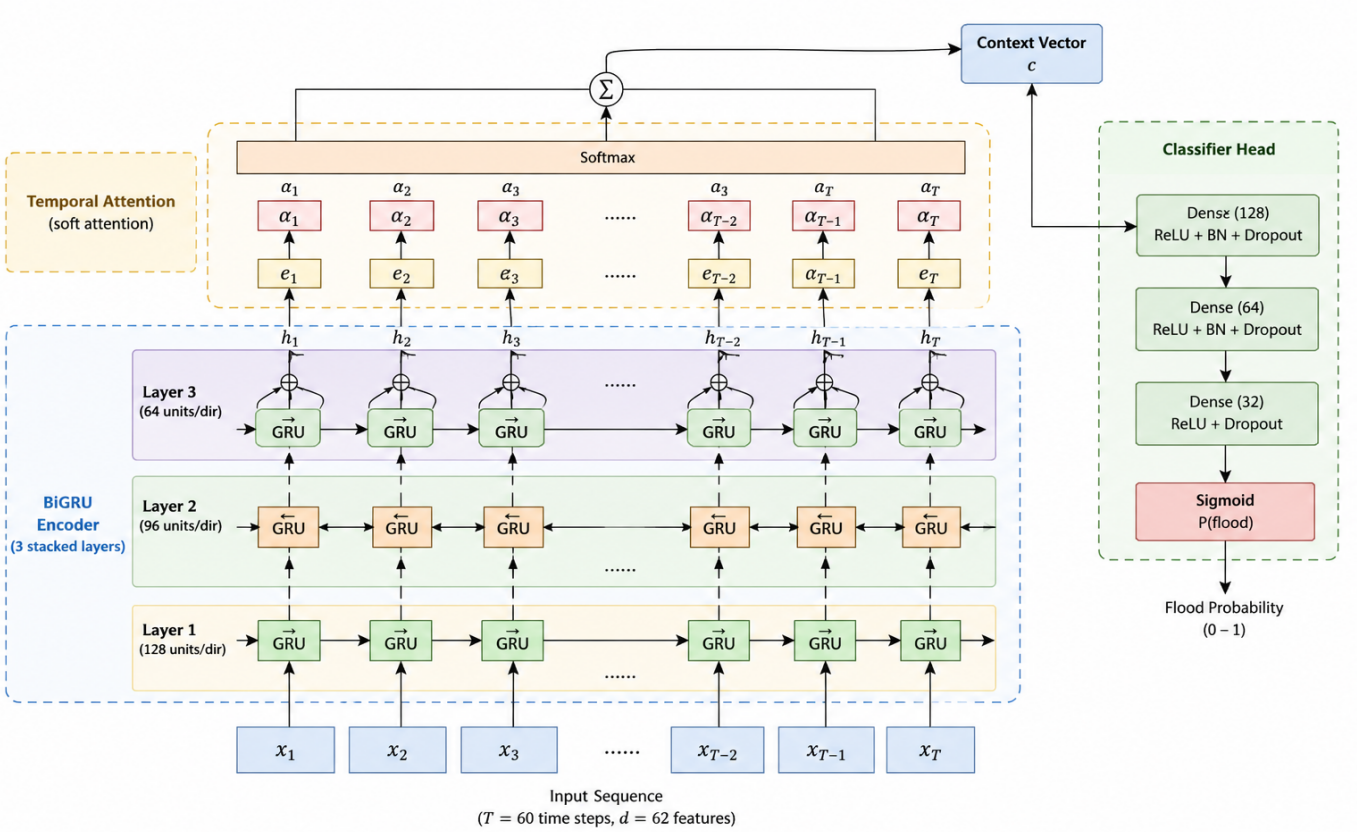

Fig. 1. Proposed BiGRU architecture with temporal attention for flood prediction using environmental time-series data.

TABLE IV
TEST SET RESULTS (1,361 DAYS: 161 FLOOD / 1,200 NO-FLOOD)

Model	AUC	PR-AUC	F1	Rec.	Prec.	Acc.
Majority class	0.500	—	0.000	0.000	—	0.882
Rolling-sum LR	0.857	0.512	0.410	0.540	0.330	0.880
BiGRU (no attention)	0.860	0.824	0.402	0.901	0.250	0.877
BiGRU + Attn. (limited features)	0.907	0.684	0.563	0.671	0.484	0.877
BiGRU + Attn. (proposed)	**0.958**	**0.834**	**0.829**	**0.814**	**0.845**	**0.960**

AUC: Area Under the ROC Curve (equivalently ROC-AUC); PR-AUC: Area Under the Precision–Recall Curve; F1: F1-score; Rec.: Recall; Prec.: Precision; Acc.: Accuracy; LR: Logistic Regression.

represent an internal ablation across the model's development history. The table reports the Area Under the ROC Curve (AUC, equivalently ROC-AUC), the Area Under the Precision-Recall Curve (PR-AUC), the F1-score, and Recall, Precision, and Accuracy for the flood class.

The proposed model reduces false alarms from 115 to 24—a **79% reduction**—and missed floods from 53 to 30 (a 43% reduction), while improving flood recall from 67.1% to 81.4%. The accuracy of 96.0% meaningfully exceeds the naive baseline of 88.2%, whereas the no-attention BiGRU baseline falls slightly below it.

Effect of removing SMOTE. Substituting SMOTE with Focal Loss and capped class weights, and correcting the temporal attention implementation, produced the first substantial gain (F1: +0.161, from 0.402 to 0.563; AUC: 0.860 to 0.907). Notably, the number of false alarms was unchanged at 115 between these two iterations, indicating that the precision problem was not attributable to SMOTE alone but also to the limited feature set available at this stage (local meteorology only, without upstream catchment signals). Recall declined from 0.901 to 0.671, reflecting Focal Loss's lower penalty on negative examples relative to SMOTE and yielding a markedly better precision–recall trade-off.

Effect of upstream features, SPI, and flood history. The largest single gain (F1: +0.266, false alarms: −91, missed floods: −23) occurred when the upstream Cherrapunji and Barak Valley signals, water-balance indicators, flood-history lags, and the SPI-14/SPI-30 indices were added, together with an increase in BiGRU capacity (96/64/32 → 128/96/64 units per direction) and ensemble size (3 → 5 members). SPI-14 and SPI-30 allow the model to distinguish anomalous rainfall from ordinary monsoon precipitation, while flood history lags capture the temporal clustering of flood episodes characteristic of multi-day inundation events. This indicates that the information carried by the upstream features was largely complementary to the local precipitation signal, rather than redundant with it.

Flood history lags. Features encoding recent flood occurrence (lag-1/2/3/7, streak) proved critical for precision: once

TABLE V
BIGRU (PROPOSED, OVERALL TEST-SET METRICS) VS. LLAMA-3.3-70B (ZERO-SHOT, 120 BALANCED TEST DAYS: 60 FLOOD / 60 NO-FLOOD)

Model	Prec.	Rec.	F1	AUC
BiGRU (proposed)	**0.845**	**0.814**	**0.829**	**0.958**
Llama-3.3-70B	0.000	0.000	0.000	0.500

a flood sequence begins, the model maintains high confidence without generating spurious predictions on isolated high-rain days.

Imbalance handling. Replacing SMOTE with Focal Loss and capped class weights ($10\times$) avoided the unrealistic synthetic sequences that SMOTE can introduce when applied to time-series windows. SMOTE flattened each 60-day window into a $60\times62 = 3{,}720$-dimensional vector and interpolated between flood sequences, producing synthetic temporal patterns with no real-world counterpart. The class-weight cap itself was tuned empirically: an earlier configuration using a cap of $14.1\times$ over-penalised false negatives and produced 115 false alarms; reducing the cap to $10\times$ in the proposed model gave a substantially better precision–recall balance (24 false positives, 30 false negatives).

B. Temporal Generalisation

Five-fold walk-forward validation was used to test whether the test-set result depends on the particular years spanned by the held-out period. In each fold, the model was trained on all data up to a given date and evaluated on the following block of approximately 550 days, using the same architecture and training settings as the main experiments. The proposed model achieves a mean AUC of 0.952 ± 0.010 and a mean F1-score of 0.817 ± 0.013 across the five folds (individual fold AUCs: 0.941, 0.964, 0.958, 0.943, 0.951; individual fold F1-scores: 0.801, 0.837, 0.822, 0.809, 0.818). This is closely consistent with the held-out test AUC of 0.9575, and the low standard deviation indicates that the model's discriminative ability does not depend on any particular evaluation period. As a point of comparison, applying the identical walk-forward procedure to the BiGRU + Attn. (limited features) variant—i.e., without the upstream catchment signals—yields a mean AUC of 0.893 ± 0.021, a lower mean and a higher variance consistent with greater reliance on local rainfall thresholding, which is itself more sensitive to antecedent soil moisture conditions not captured by that variant.

C. LLM Comparison

This subsection reports the auxiliary zero-shot LLM benchmark motivated in Section III; it supplements, and does not replace, the primary comparison against conventional forecasting baselines in Table IV. Table V reports aggregate Precision, Recall, F1-score, and AUC for Llama-3.3-70B on the balanced 120-day subsample, alongside the proposed BiGRU ensemble's overall test-set metrics as a point of reference.

Llama-3.3-70B classified none of the 120 examples as a flood, yielding 0 true positives and 60 false negatives out of 60 flood days, an F1-score of 0.000, and an AUC of 0.500 (equivalent to random-chance discrimination). This behaviour is consistent with the constrained nature of the prompt: given only a single day's raw precipitation, temperature, and humidity values, and without access to the 60-day engineered sequence available to the BiGRU, the model did not distinguish flood-prone conditions from ordinary rainy days.

This result underscores the value of sequential, feature-engineered temporal modelling for this task: under the tested zero-shot, snapshot-based prompting condition with minimal raw inputs, the signal lies not in any single day's raw weather values but in the accumulation and trend of precipitation over the preceding weeks. Given the input asymmetry described in Section III, this finding should be read as evidence for the benefit of sequence-based numerical modelling in this setting, rather than as a general claim that LLMs are intrinsically incapable of flood prediction; an LLM supplied with engineered features, longer temporal context, few-shot examples, or task-specific fine-tuning may perform differently, and we leave such comparisons to future work.

D. Attention Interpretability

For correctly predicted flood events with strong upstream precursors, the learned attention weights peak approximately 1–3 days before the flood day, with a secondary peak aligned with the 1-week rolling precipitation maximum, consistent with the known hydraulic travel time from the Cherrapunji catchment to the Surma–Kushiyara confluence. By contrast, missed flood events (false negatives) show attention weights distributed more uniformly across the 60-day window, with no clear upstream peak; inspection of the corresponding feature values indicates that upstream precipitation was close to seasonal average in the days preceding these events, suggesting they were driven by purely local factors not fully captured by the current feature set. False alarms show strong attention on upstream precipitation pulses that did not translate into flooding, plausibly due to lower antecedent soil moisture than the training distribution. This behaviour is consistent with the model having learned physically plausible temporal patterns rather than exploiting spurious statistical correlations, though the analysis is qualitative, based on inspection of representative cases rather than a formal statistical test across the full test set.

E. Contextualisation Against GeoAI Advances

Recent advances in geospatial artificial intelligence (GeoAI) for hydrology have introduced multimodal architectures that fuse satellite EO imagery, digital elevation models, and meteorological signals into unified representations. Geospatial foundation models pre-trained on large-scale EO datasets have further demonstrated transfer learning capabilities across diverse hydrological regimes. While these approaches represent the current frontier of flood modelling, they typically require dense sensor infrastructure, high-resolution satellite archives, and substantial computational resources that are often unavailable in data-scarce or low-resource regions such as the Sylhet district.

The proposed BiGRU system occupies a complementary and practically important role in this landscape. Operating exclusively on freely accessible meteorological time-series data—precipitation, temperature, and humidity—it delivers robust early-warning predictions (ROC-AUC = 0.9575, F1 = 0.8291) with minimal infrastructure requirements. Rather than competing with EO-based or foundation model pipelines, this work is best understood as a *meteorology-driven early warning component* that can serve as an upstream signal provider within more complex multimodal systems: the daily flood probability output can directly feed into geospatial pipelines as a lightweight, low-latency trigger, reducing the computational burden of activating expensive EO-based inference only when the meteorological risk threshold is exceeded.

F. Three-Zone Alert System

The operational output maps probabilities to three actionable zones, defined using two thresholds, T_{HIGH} and T_{LOW}: **HIGH** ($p \geq T_{\text{HIGH}}$) triggers emergency response; **MODERATE** ($T_{\text{LOW}} \leq p < T_{\text{HIGH}}$) prompts preparation and monitoring; **LOW** ($p < T_{\text{LOW}}$) corresponds to normal operations. The upper threshold was set at $T_{\text{HIGH}} = 0.539$, selected to maximise F1 on the validation set; the lower threshold was set at $T_{\text{LOW}} = T_{\text{HIGH}} - 0.12 = 0.419$ to provide an early-warning buffer.

V. Conclusion

In this paper, we propose a BiGRU model with temporal attention to predict floods using time-series data of environmental parameters spanning 25 years. Using 62 physically motivated engineered features—including SPI values, lagged flood history, rolling precipitation aggregates, seasonality encoding, and 13 upstream catchment signals from the Cherrapunji headwater, Barak Valley, and water-balance indicators—as well as replacing SMOTE with Focal Loss, we achieve ROC-AUC = 0.9575 and F1-score = 0.8291, improving the baseline AUC and F1-score by 0.098 and 0.427, respectively, while reducing false alarms by 79% and missed floods by 43% relative to the earlier BiGRU + attention variant without upstream features. Five-fold walk-forward validation confirms the temporal robustness of the proposed approach, with mean AUC $= 0.952 \pm 0.010$ and mean F1-score $= 0.817 \pm 0.013$ across folds. A deliberately constrained zero-shot benchmark against Llama-3.3-70B, in which the model was given only a single day's raw precipitation, temperature, and humidity values rather than the full 60-day engineered sequence used by the BiGRU, gives F1 $= 0.00$ under this prompt design. This result highlights the value of sequence-based numerical modelling for this task; it characterises zero-shot, snapshot-based textual reasoning over minimal raw inputs under the tested prompt design and should not be interpreted as evidence that LLMs are intrinsically incapable of flood prediction more broadly.

Limitations. Despite the strong results, several limitations should be acknowledged. First, the model is trained and evaluated on a single geographic location (Sylhet, Bangladesh), and its generalisation to other catchments with different climatological or hydrological regimes has not been verified. Second, the flood labels are derived from historical records that may themselves contain annotation noise or reporting gaps, which could affect both training and evaluation. Third, the current framework operates at daily temporal resolution, limiting its utility for flash-flood scenarios that develop within hours. Fourth, the LLM baseline reported in Section IV is a general-purpose model not designed or trained for spatiotemporal time-series forecasting; it serves only as an auxiliary probe of zero-shot textual reasoning, and this study does not benchmark the proposed BiGRU against other dedicated sequence architectures (e.g., Transformer-based or Temporal Convolutional Network (TCN)-based time-series forecasters) beyond the BiGRU ablations and conventional baselines in Table IV. Such comparisons remain an open direction for future evaluation. Finally, a direct soil-moisture measurement was originally planned via the ERA5 archive but was unavailable at daily resolution through the API; the water-balance category (precipitation minus reference evapotranspiration) used in its place is a physically motivated approximation rather than a direct measurement, meaning the model does not yet fully exploit true catchment saturation signals.

Future work will address hourly time-series forecasting, incorporate MODIS (Moderate Resolution Imaging Spectroradiometer) satellite imagery of flood extents, and automate daily predictions.

References

[1] UNDRR, "The human cost of disasters: An overview of the last 20 years (2000–2019)," United Nations Office for Disaster Risk Reduction, Geneva, 2020.

[2] M. Hasan, M. A. Islam, and M. S. Rahman, "Flood prediction using machine learning models: Literature review," *Water*, vol. 11, no. 7, 2019.

[3] A. Mosavi, P. Ozturk, and K. Chau, "Flood prediction using machine learning models: Literature review," *Water*, vol. 10, no. 11, p. 1536, 2018.

[4] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, "Rainfall–runoff modelling using long short-term memory (LSTM) networks," *Hydrology and Earth System Sciences*, vol. 22, no. 11, pp. 6005–6022, 2018.

[5] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] T. Song, F. Ding, J. Wu, H. Xu, and C. Peng, "Flash flood forecasting based on long short-term memory networks," *Water*, vol. 12, no. 1, p. 81, 2020.

[7] G. Selva Jeba, P. Chitra, and U. M. Rajasekaran, "Time-series analysis and flood prediction using a deep learning approach," in *Proc. IEEE Int. Conf. on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2022, pp. 139–142.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE ICCV*, 2017, pp. 2980–2988.

[10] Visual Crossing Corporation, "Weather Query Builder — Historical Weather Data," https://www.visualcrossing.com/weather-query-builder/, accessed May 2025.

[11] Open-Meteo, "Historical weather API (ERA5)," https://open-meteo.com, accessed May 2025.

[12] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.