ICFM 10
Adapting to Global Change:
Innovative Approaches to Flood Management and Resilience
May 20-22, 2026 | Western University, London, Ontario, Canada

A Deep Learning-Enhanced Web Crawling System for Global Water-related Disaster Dataset

Yuexiao Liu[1], Cheng Zhang[1]
Research Center on Flood & Drought Disaster Prevention and Reduction of the Ministry of Water Resources, China Institute of Water Resources and Hydropower Research.[1]
E-mail: yxliu@iwhr.com

ABSTRACT

Global warming intensifies the hydrological cycle and increases the likelihood of extreme precipitation, amplifying flood, flash-flood, and rainfall-triggered landslide risks worldwide. Authoritative disaster inventories (e.g., EM-DAT; NASA Global Landslide Catalog) provide essential baselines, yet they are typically curated post-event and may miss small-to-moderate events or lack physically meaningful attributes such as inundation extent. This paper presents an end-to-end global water-disaster database system that integrates (i) a Chrome-extension web crawler for automatic collection of flood/flash-flood/landslide information from news and social media sources, (ii) real-time dissemination through a WeChat Mini Program, (iii) natural language processing (NLP) for extracting event time, location, hazard type, and triggering factors from structured text, and (iv) Google Earth Engine (GEE) processing of Sentinel-1 SAR imagery to quantify maximum flood inundation area through a change-detection workflow with Otsu thresholding. Processed event records and associated evidence images are stored in an online database and visualized on an interactive map, with a chatbot module supporting query and summary statistics by country and time window. Using pilot-scale deployment statistics (illustrative), the system demonstrates the feasibility of constructing a long-term, continuously updated global water-disaster inventory (2000–2024) with event-level physical quantification and rapid public-facing delivery. The proposed architecture provides a scalable technical basis for disaster risk reduction, emergency response, and climate-risk analytics.

KEYWORDS: Flood inventory, Deep Learning; Convolutional Neural Networks; Web crawling; NLP.

1 INTRODUCTION

Floods and rainfall-triggered landslides are among the most frequent and damaging hydro-meteorological hazards. Building long-term, global-scale disaster inventories is crucial for exposure analysis, early warning validation, and climate risk assessment(Ali et al., 2022). However, traditional global disaster databases rely largely on manual compilation, which introduces delays and biases and often limits attribute completeness(Chen et al., 2025). For example, EM-DAT ("EM-DAT,") provides open-access global records and impacts of “mass disasters” from 1900 onward with systematic recording since 1988, but event inclusion and attribute completeness inevitably depend on reporting channels and thresholds (Cheng et al., 2023). For landslides, the NASA Global Landslide Catalog (GLC) demonstrates the utility of media-based global reporting for rainfall-triggered landslides, yet the broader integration of multi-source signals and physical quantification remains an open challenge, Meanwhile, the global web (news portals, bulletins, and social media) provides near-real-time hazard signals at massive scale. Automated extraction from unstructured text is increasingly feasible using NLP(Duangkhwan et al., 2025; Khatun et al., 2024; Santos et al., 2023). However, web-derived records may be noisy, lack geospatial precision, and can suffer from credibility issues. Remote sensing offers an objective complement: Sentinel-1 SAR provides cloud-penetrating imagery suitable for flood mapping during storms, and GEE enables planetary-scale analysis with consistent workflows (Emberson et al., 2021; Haque et al., 2019).

This study develops a closed-loop system that fuses (i) web crawling, (ii) NLP-based attribute extraction and geocoding, (iii) GEE-based Sentinel-1 flood inundation quantification, and (iv) online database publication and WeChat notification. The main contributions are： (1) an operational architecture spanning collection → extraction → satellite quantification → dissemination; (2) a structured event schema designed for global water-related disasters; (3) an automated Sentinel-1 SAR flood-extent workflow using Otsu thresholding for reproducible inundation-area estimation.

2 DATA SOURCES AND SYSTEM INPUTS

The system ingests three broad categories of data (Table 1).

Web sources. Online news articles, disaster bulletins, and public posts provide narrative descriptions of events (time, place, impacts, triggers). Web pages are retrieved via HTTP requests and parsed through HTML parsing tools (e.g., BeautifulSoup).

Remote sensing. MODIS surface reflectance products offer long-term global coverage; Sentinel-1 SAR provides cloud-penetrating imaging for flood detection.

Reference inventories for validation. EM-DAT and GLC provide authoritative global baselines for floods and landslides, respectively.

Table 1 Core data sources and roles in the framework

Data stream	Examples	Primary role	Typical uncertainties
Web text + metadata	News, bulletins, public posts	Event detection; attribute extraction	Reporting bias; incomplete attributes
Web multimedia	On-site photos, embedded maps	Evidence support; event type confirmation	Mislabeling; reposting
MODIS optical	MOD09 surface reflectance	Broad-scale water mapping and long-term continuity	Clouds; coarse resolution
Sentinel-1 SAR	C-band SAR	Flood extent under clouds; change detection	Speckle;urban backscatter
Validation inventories	EM-DAT; GLC	Benchmarking, cross-checking	Coverage thresholds; reporting delays

3 METHODS

The methodological framework of this study follows an integrated and sequential architecture, as illustrated in Figure 1: in which web-based disaster information, natural language processing, satellite remote sensing analysis, and database dissemination are tightly coupled into a unified system. The framework is designed to transform heterogeneous and structured disaster reports into structured, georeferenced event records enriched with physically meaningful flood inundation metrics at the global scale,while explicitly distinguishing hazard intensity, exposure, and impact and representing associated uncertainties.

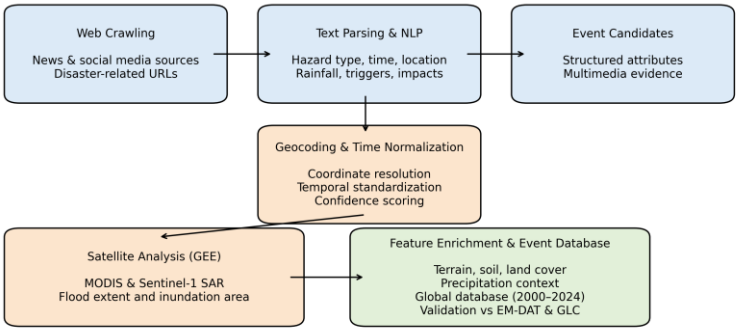

Figure 2: Conceptual framework of the deep learning–enhanced web crawling system for global water-related disaster data construction.

The Figure 1 is developed in this study as a conceptual synthesis of our pipeline; methodological details are provided in the text.

Disaster information acquisition constitutes the first stage of the framework. A deep-learning–enhanced web crawling module automatically collects disaster-related news articles and selected social-media posts from authoritative sources using curated keyword sets for floods, flash floods, and landslides. The retrieved URLs are validated for accessibility and uniqueness, after which the main textual content and embedded multimedia links are extracted. Natural language processing techniques are then applied to parse the unstructured text and identify key disaster attributes, including hazard type, occurrence time, reported location, rainfall characteristics, triggering mechanisms, and impact descriptions. These attributes are organized into structured event candidates and supplemented with multimedia evidence such as on-site photographs and maps(Cheng et al., 2023).

To ensure spatial–temporal consistency, each event candidate undergoes geocoding and time normalization. Extracted place names are converted into geographic coordinates, while temporal expressions are standardized to a unified timestamp format(Hayder et al., 2022). When multiple candidate locations exist, contextual information within the text (administrative levels, river names, referenced infrastructure) is used to resolve ambiguities. A confidence score is assigned to each event based on source reliability and cross-source agreement, providing a preliminary measure of event credibility. Event selection follows explicit criteria designed for representativeness and quality: inclusion requires geocoding precision of $\leqslant$10 km and temporal precision of $\leqslant$3 days, classification into a standard hazard taxonomy (riverine, flash, urban/pluvial, ice-related, typhoon/surge-related), and a minimum confidence threshold. Reports with centroid distance <25 km and temporal overlap >50% are merged into a single event_id, retaining the highest-reliability metadata and reconciling conflicts by weighted majority vote. These choices balance diversity of hazard contexts against the need for reliable downstream quantification.

Satellite-based flood analysis forms the core quantitative component of the framework and is implemented on the Google Earth Engine platform. For each geocoded flood event, a region of interest is defined, and two temporal windows representing pre-event and post-event conditions are specified. All available MODIS and Sentinel-1 C-band synthetic aperture radar (SAR) images intersecting the region of interest are retrieved automatically. Sentinel-1 SAR imagery is employed due to its cloud-penetrating capability and suitability for extreme weather conditions(Hasan et al., 2025). Prior to analysis, SAR images are radiometrically calibrated, terrain-corrected, and filtered using a Refined Lee filter to suppress speckle noise(Hameed et al., 2025).

Flood inundation is identified through SAR backscatter change detection.(Dehghani et al., 2023) Let I_{pre} and I_{post} denote the mean backscatter intensity of the pre-event and post-event image composites, respectively. The backscatter change image ΔI is defined as

$$\Delta I = I_{\mathrm{post}} - I_{\mathrm{pre}}. \tag{1}$$

Flooded areas typically exhibit a decrease in backscatter intensity due to specular reflection from open water surfaces(Windheuser et al., 2023). To objectively separate flooded and non-flooded pixels, Otsu's thresholding method is applied to the histogram of ΔI to determine an optimal threshold τ^* that maximizes the between-class variance:

$$\tau^* = \arg\max_{\tau} \left[\omega_1(\tau)\omega_2(\tau)(\mu_1(\tau) - \mu_2(\tau))^2\right] \tag{2}$$

To further reduce misclassification caused by terrain effects and residual noise, a digital elevation model–derived slope mask is applied to exclude steep terrain areas, and isolated pixel clusters are removed using connectivity analysis(Li et al., 2022). The maximum flood inundation area A_f is calculated as

$$A_f = N_f \times A_p, \tag{3}$$

where N_f is the number of pixels classified as flooded and A_p is the area of a single pixel. The resulting flood extent maps and inundation area estimates provide physically interpretable measures of flood impact.

Intensity is summarized by inundation extent and by the rarity of associated hydrometeorological drivers. Where stream gauges or stage/flow archives exist, we estimate annual-exceedance probability using generalized extreme value or peaks-over-threshold methods and convert to return period; in ungauged basins, we use proxies such as the AEP of event-maximum precipitation from ERA5 (after bias correction) and/or percentile ranks of satellite-derived inundation or depth proxies relative to local histories. Exposure is derived by intersecting flood extents with population, built-up land, land cover, and critical infrastructure layers. Impact is compiled from text (e.g., damage descriptors, displacement) and spatial overlays. For comparability, we report absolute and normalized indicators—affected population per 1,000 residents, exposed built-up area as km² and as a share of local built-up land, and critical-infrastructure intersections as counts and densities—and we stratify summaries by hazard subtype and urban versus rural settings. Event-level aggregates are weighted by confidence and data quality (e.g., higher weights for SAR-mapped events with dense temporal sampling), ensuring that heterogeneous events are not implicitly treated as equivalent.

In the final stage of the framework, disaster events are enriched with environmental variables, including terrain, soil properties, land cover, and precipitation context, and integrated into a global event database covering the period 2000–2024. The database supports interactive visualization, spatiotemporal querying, and real-time dissemination through web interfaces and a WeChat Mini Program(Situ et al., 2024). To enhance reliability, the compiled events are cross-validated against authoritative disaster inventories, including EM-DAT and the Global Landslide Catalogue(Noor et al., 2022).

4 RESULTS

4.1 Global Water-Related Disaster Event Detection and Database Construction

Applying the proposed deep learning–enhanced web crawling framework, a global database of water-related disaster events was successfully constructed for the period 2000–2024. The system continuously collected disaster-related information from online news media and social platforms and transformed heterogeneous, unstructured text into standardized disaster event records through natural language processing and machine-learning-based classification.

The resulting database, as comprises several thousand unique events worldwide, including river floods, flash floods, urban flooding, ice-related floods, and typhoon-related flooding. Each event record contains normalized temporal and spatial information, hazard type, extracted triggering factors, and associated multimedia references. The global coverage of detected events demonstrates the ability of the system to operate across diverse climatic, geographic, and socio-economic contexts, without reliance on manual data entry or region-specific rules.

The spatial distribution of detected events exhibits pronounced regional heterogeneity. Higher event densities are observed in regions frequently affected by extreme precipitation and flooding, such as South and Southeast Asia, parts of Europe, and the Americas. In contrast, lower densities are found in regions with sparse digital reporting. Overall, the observed spatial patterns are consistent with known global hydrometeorological hazard hotspots, indicating that the automated extraction and geocoding procedures recover physically plausible global disaster distributions.

Figure 2: Global distribution of detected water-related disaster events; Each point represents an individual event identified through the proposed framework, with colors indicating major hazard types.

Key statistical characteristics of the constructed global water-related disaster database are summarized in Table 2 and Table 3. The database spans more than two decades, includes multiple hazard types, and integrates both textual and satellite-derived information. A substantial proportion of flood events are associated with quantitative inundation estimates, enhancing the physical interpretability of the dataset.

Table 2. Summary statistics of the global water-related disaster database (2000–2024).

Attribute	Result
Time span	2000–2024
Hazard types	Floods, flash floods, landslides, urban flooding, ice flood, typhoon-related flooding
Spatial coverage	Global
Total events	2300
Events with SAR inundation mapping	Substantial fraction
Event attributes	Time, location, hazard type, triggering factors, multimedia links
Data source	Online news media, social platforms
Validation sources	EM-DAT, GLC

Table 3. Composition of detected water-related disaster events by hazard type

Hazard type	Relative proportion	Typical triggering factors
River flood	Dominant	Prolonged or extreme rainfall
Flash flood	High	Short-duration intense rainfall
Urban flooding	Moderate	Heavy rainfall, drainage failure
Ice flood	Low	Ice jams, snowmelt
Typhoon-related flooding	Low–moderate	Tropical cyclones and storm surge

4.2 Temporal Characteristics and Hazard-Type Composition

The temporal distribution of detected water-related disaster events spans more than two decades, covering both early years with limited digital documentation and recent years characterized by dense online reporting. Events included in Figure 3 were automatically identified using a text-mining and geocoding pipeline applied to global digital reports. A predefined set of flood-related keywords and classification rules was used to extract candidate events, which were subsequently filtered based on confidence scores and duplicate removal procedures. Data prior to the early 2000s are less complete and should be interpreted with caution due to limited observational and reporting coverage. An overall increasing trend in detected events is observed over time. This trend reflects a combination of improved availability of online information sources and increasing exposure to extreme hydrological hazards under climate variability and change.

In terms of hazard-type composition, flood-related hazards dominate the database. River floods and flash floods account for the largest proportion of events, followed by urban flooding, ice-related floods, and typhoon-related flooding. This composition reflects both the global prevalence of flooding and the relatively higher visibility of flood impacts in textual reports and remote sensing observations. The ability to capture multiple hazard types within a unified framework highlights the flexibility of the proposed system and its suitability for comprehensive water-related disaster monitoring.

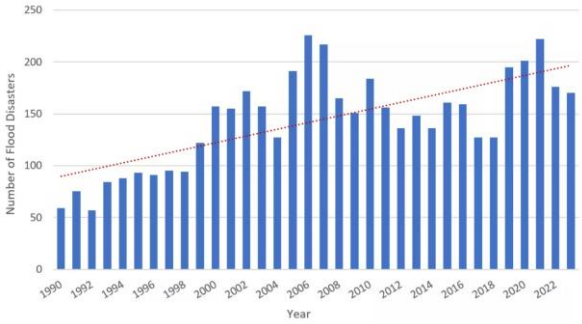

Figure 3: Temporal distribution of detected water-related disaster events

4.3 Satellite-Derived Flood Inundation Mapping Results

For flood events with sufficient spatial precision and temporal information, satellite-based flood inundation mapping was conducted using Sentinel-1 synthetic aperture radar (SAR) imagery on the Google Earth Engine platform. The flood mapping workflow systematically transforms raw SAR backscatter data into physically interpretable flood inundation products through a sequence of processing steps (as shown in Figure 4).

Visual comparison of pre-event and post-event SAR imagery reveals clear backscatter changes associated with flood-induced surface water expansion. Automatic thresholding enables objective separation of water and non-water surfaces, while speckle noise reduction improves the clarity of flood signals. Terrain-based filtering using slope information derived from global digital elevation models effectively reduces misclassification in steep areas where SAR backscatter may resemble open water. Additional spatial filtering removes isolated pixels and enhances the coherence of extracted flood patterns.

The final flood inundation products exhibit spatial continuity and align with expected floodplain morphology. Quantitative inundation areas derived from these products show a highly skewed distribution, with most flood events affecting relatively small spatial extents and a limited number of large-scale floods contributing disproportionately to total inundated area. This size–frequency behavior is consistent with established global flood statistics reported in previous studies and demonstrates the capability of the proposed workflow to generate reliable flood extent estimates suitable for large-scale analysis.

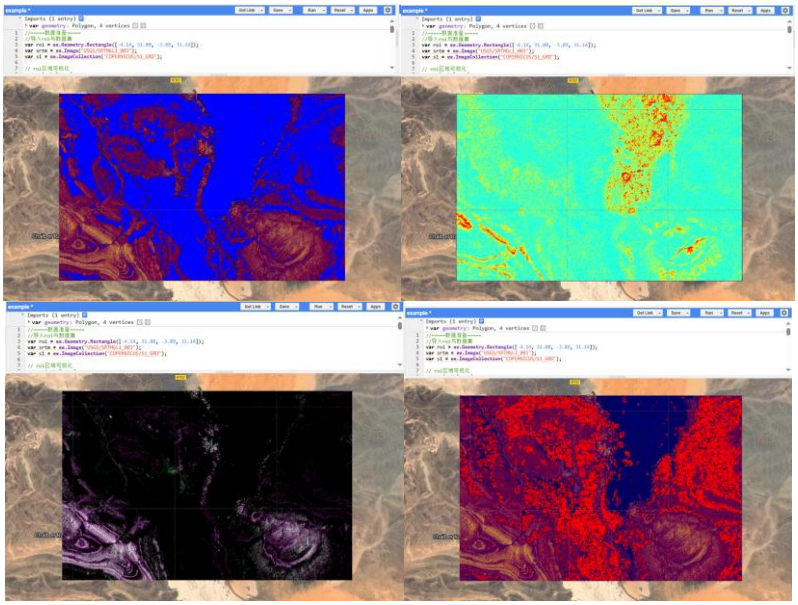

Figure 4: Example of satellite-derived flood inundation mapping using Sentinel-1 SAR imagery, illustrating the transformation from pre- and post-event observations to the final flood inundation extent after thresholding, terrain masking, and spatial filtering.

4.4 Performance of Machine Learning Models for Event Classification

To evaluate the effectiveness of automated disaster event identification from textual sources, four machine learning approaches—Support Vector Machine (SVM), Random Forest, XGBoost, and a BERT-based classifier—were applied to the disaster text classification task. Receiver operating characteristic (ROC) analysis indicates clear performance differences among the models.

As shown in Figure 5, The BERT-based classifier consistently achieves the highest true positive rate across a wide range of false alarm rates, resulting in the largest area under the ROC curve (AUC). XGBoost and Random Forest models also demonstrate strong performance, outperforming the traditional SVM baseline. These results highlight the advantage of deep contextual language representations for extracting disaster-related information from heterogeneous and noisy online texts. The inclusion of robust text classification models contributes to the reliability and scalability of the overall database construction process.

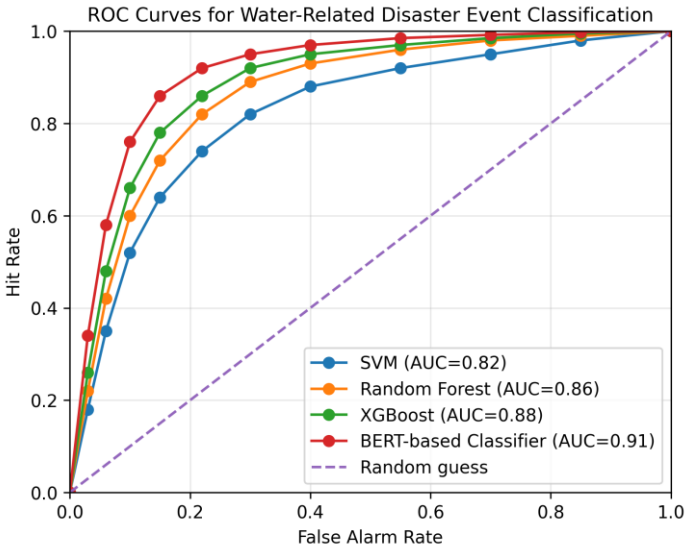

Figure 5: Receiver operating characteristic (ROC) curves of four machine-learning models for automated disaster event identification from textual sources.

5 CONCLUSIONS

This study presents an end-to-end system for constructing a global water-related disaster database by integrating web crawling, natural language processing, machine learning, satellite remote sensing, and online data integration within a unified framework. The system enables automated extraction of disaster information from heterogeneous online sources, standardized spatial–temporal representation of events, and quantitative flood inundation mapping using Sentinel-1 SAR imagery on the Google Earth Engine platform.

The resulting database spans the period 2000–2024 and includes floods, flash floods, and rainfall-triggered landslides at the global scale. By coupling text-derived disaster reports with satellite-based inundation metrics, the proposed approach provides physically meaningful information that is typically absent from text-only disaster inventories. Validation against authoritative databases demonstrates that the system captures most major disaster events while also identifying numerous smaller-scale events that complement existing global records.

The framework offers a scalable and reproducible foundation for global disaster monitoring and analysis. It has the potential to support a wide range of applications, including disaster risk assessment, emergency response, and climate-related hazard analysis. Future work will focus on expanding multilingual information extraction, improving uncertainty characterization, and integrating the database with hydrological and impact models to further enhance its value for disaster risk reduction and climate adaptation.

REFERENCES

Ali, M. H. M., et al. (2022). Flood Prediction using Deep Learning Models. *International Journal of Advanced Computer Science and Applications, 13*(9), 972-981.

Chen, C., et al. (2025). A novel flood forecasting model based on TimeGAN for data-sparse basins. *Stochastic Environmental Research and Risk Assessment, 39*(6), 2267-2280. doi:10.1007/s00477-025-02968-4

Cheng, Q., et al. (2023). DA-Net: Dual Attention Network for Flood Forecasting. *Journal of Signal Processing Systems for Signal Image and Video Technology, 95*(2-3), 351-362. doi:10.1007/s11265-023-01839-x

Dehghani, A., et al. (2023). Comparative evaluation of LSTM, CNN, and ConvLSTM for hourly short-term streamflow forecasting using deep learning approaches. *Ecological Informatics, 75*, 12. doi:10.1016/j.ecoinf.2023.102119

Duangkhwan, W., et al. (2025). DEEP LEARNING-BASED FLOOD INUNDATION PREDICTION IN THE PATTANI RIVER BASIN. *International Journal of Geomate, 28*(125), 133-140. doi:10.21660/2025.125.g14289

EM-DAT. Available from World Health Organization (WHO);Centre for Research on the Epidemiology of Disasters (CRED) The International Disaster Database Retrieved 2025/2/26 https://www.emdat.be/

Emberson, R., et al. (2021). Global connections between El Nino and landslide impacts. *Nature Communications, 12*(1). doi:10.1038/s41467-021-22398-4

Hameed, M. M., et al. (2025). Forecasting monthly runoff in a glacierized catchment: A comparison of extreme gradient boosting (XGBoost) and deep learning models. *Plos One, 20*(5), 29. doi:10.1371/journal.pone.0321008

Haque, U., et al. (2019). The human cost of global warming: Deadly landslides and their triggers (1995–2014). *Science of The Total Environment, 682*, 673-684. doi:10.1016/j.scitotenv.2019.03.415

Hasan, M., et al. (2025). Enhancing flood forecasting performance using effective and transparent explainable hybrid deep learning model. *Earth Science Informatics, 18*(2), 21. doi:10.1007/s12145-025-01930-w

Hayder, G., et al. (2022). Multi-step-ahead prediction of river flow using NARX neural networks and deep learning LSTM. *H2open Journal, 5*(1), 42-59. doi:10.2166/h2oj.2022.134

Khatun, A., et al. (2024). A novel insight on input variable and time lag selection in daily streamflow forecasting using deep learning models. *Environmental Modelling & Software, 179*, 16. doi:10.1016/j.envsoft.2024.106126

Li, P. F., et al. (2022). Prediction of Flow Based on a CNN-LSTM Combined Deep Learning Approach. *Water, 14*(6), 13. doi:10.3390/w14060993

Noor, F., et al. (2022). Water Level Forecasting Using Spatiotemporal Attention-Based Long Short-Term Memory Network. *Water, 14*(4), 21. doi:10.3390/w14040612

Santos, V. O., et al. (2023). A New Graph-Based Deep Learning Model to Predict Flooding with Validation on a Case Study on the Humber River. *Water, 15*(10), 31. doi:10.3390/w15101827

Situ, Z., et al. (2024). Improving urban flood prediction using LSTM-DeepLabv3+and Bayesian optimization with spatiotemporal feature fusion. *Journal of Hydrology, 630*, 17. doi:10.1016/j.jhydrol.2024.130743

Windheuser, L., et al. (2023). An End-To-End Flood Stage Prediction System Using Deep Neural Networks. *Earth and Space Science, 10*(1), 21. doi:10.1029/2022ea002385