

FloodEcho: Linking Forecast to Impact via LLM-Extracted Multi-Dimensional Flood Consequences from Media Reports

Table of contents

1 Introduction	3
2 FloodEcho Workflow	5
2.1 Module 1: Spatial-temporal triggering	5
2.2 Module 2: Targeted media evidence retrieval	6
2.3 Module3: LLM-driven Multi-dimensional Extraction	7
2.4 Link Verification Pathway	10
2.5 Implementation Consideration	13
3 FloodEcho System Demonstration	14
3.1 Event Spatial-Temporal Distribution	14
3.2 Case Study: Barcelona Event - 21 September 2025	17
4 Validation and Comparative Analysis	19
4.1 Cross Validation with Hanze Database and Groundsource	19
4.2 Manual Evaluation of 100 Selected Events	21
4.3 Evaluation of Link verification pathway	23
4.4 Evaluation of Event Impact Severity Assigned by FloodEcho	26
5 Discussions and Conclusions	27

1 Introduction

Floods remain among the most disruptive natural hazards for societies, infrastructure systems, and economies. Their consequences are not determined by hydrometeorological intensity alone: the same rainfall or river discharge can produce very different outcomes depending on exposure, drainage capacity, land use, infrastructure redundancy, warning uptake, and local emergency response. This distinction is becoming more important as climate change alters heavy-precipitation hazards while urbanization and development patterns continue to change the people and assets exposed to flooding (Climate Change 2022, 2023; Tellman et al., 2021; Winsemius et al., 2015). For flood early warning, the scientific and operational challenge is therefore not only to anticipate where hazardous rainfall or river levels may occur, but also to understand what those hazards do on the ground. This challenge has motivated a broader shift from hazard-based forecasting toward impact-based forecasting and warning. Traditional flood warning systems are often organized around physical drivers and thresholds, such as accumulated rainfall, river stage, or modelled discharge. These variables are essential for detecting hazardous conditions, but they do not directly encode the societal consequences of an event. Impact-based forecasting reframes the warning problem around expected consequences for people, assets, services, and response systems, aligning with the World Meteorological Organization's guidance that warnings should communicate not only the hazard but also its likely impacts (World Meteorological Organization, 2015).

Interviews and data-needs studies with weather, flood, and emergency-management personnel further show that impact-based warning systems are valued for improving understanding of potential impacts, but that their implementation is constrained by limited impact data and by the difficulty of verifying warnings against observed impacts (Harrison et al., 2022; Potter et al., 2021). Observed impact data are thus a load-bearing component of impact-based flood forecasting. Without systematic observations of consequences, forecast developers can evaluate hydrological timing or magnitude but have limited evidence for whether forecasts correctly anticipated transport disruption, economic interruption, safety consequences, or emergency response needs. This is particularly problematic for flash floods and localized pluvial events, whose impacts may be severe at neighborhood or sub-regional scales but short-lived and sparsely represented in conventional observation networks.

A forecast-verification dataset for impact-based services should therefore be temporally explicit, spatially localized, multidimensional, and traceable to the evidence from which each impact claim was derived. Existing disaster and flood databases provide important foundations, but they are not fully designed for this verification role. EM-DAT, for example, is a long-running and authoritative global database for disaster events, supporting cross-national and long-term disaster-risk analysis (Delforge et al., 2025). Recent work on geocoding EM-DAT climate-related events further illustrates the value of improving spatial event representation (Teber et al., 2025). Regionally focused resources such as the HANZE database provide detailed

historical flood-impact records for Europe and are valuable for long-term loss and exposure analysis (Paprotny et al., 2024). However, natural-hazard loss and damage databases differ in inclusion criteria, reporting thresholds, geocoding, and treatment of small events, which can affect their suitability for local verification tasks (Gall et al., 2009; Panwar & Sen, 2019). Similarly, operational and research systems for flood impact forecasting demonstrate how exposure and vulnerability information can be combined with flood hazard estimates (Lang-Ritter et al., 2022), while emerging news-derived resources further show the potential of media-based flood-event databases (Mayo et al., 2026).

The remaining gap is not the absence of disaster data altogether; it is the lack of a scalable, evidence-traceable record of what actually happened, where, when, and across which impact dimensions. Media reports offer a complementary source of such observations. Journalists, public authorities, emergency responders, and citizens routinely document road closures, railway suspensions, evacuations, casualties, damaged buildings, business interruption, school closures, and emergency interventions during flood events. Prior work has shown that social media streams can be used to detect flood events at large scale, including a global database derived from multilingual Twitter data (de Bruijn et al., 2019), and that social-media information can support rapid flood inundation mapping and damage assessment (Fohringer et al., 2015; Kryvasheyeu et al., 2016). Broader crisis-informatics research has also demonstrated the value of processing social media messages during emergencies and across crisis types (Alam et al., 2018; Imran et al., 2015; Olteanu et al., 2015; Sakaki et al., 2010; Saroj & Pal, 2020; Starbird et al., 2010). Yet these sources are not directly usable as validation data. Reports are unstructured, multilingual, unevenly distributed across regions, duplicated across outlets, and influenced by media attention, population density, and source availability. A news-rich urban event can be overrepresented, while a rural event with comparable hydrometeorological severity may leave a much weaker digital trace. These biases make blind web crawling or simple keyword counting unsuitable for constructing reliable impact evidence.

Large language models provide a practical opportunity to transform unstructured narratives into structured impact records, but only if the extraction process is constrained. Recent language models have demonstrated strong few-shot and instruction-following capabilities, including the ability to summarize, classify, and reason across heterogeneous text (Brown et al., 2020; OpenAI et al., 2023). Retrieval and passage-ranking methods further support workflows in which generation or extraction is conditioned on external evidence rather than model parameters alone (Karpukhin et al., 2020; Lewis et al., 2020). At the same time, the hallucination risks documented for natural-language generation mean that unsupported inference cannot be treated as evidence (Ji et al., 2023). For flood impact verification, LLM outputs must therefore be grounded in retrieved sources, restricted to information explicitly present in those sources, and stored with source-level traceability. Moreover, the retrieval step itself should be guided by independent hydrometeorological evidence rather than by media availability alone, so that the resulting database is less dominated by reporting intensity and duplicate stories.

This paper aims to demonstrate that a physically triggered, LLM-driven extraction of flood impacts from media can produce validation datasets superior to existing databases for impact-based forecast validation. Section 2 introduces the Floodscan system architecture, detailing the forecast-triggered mechanism, media retrieval pipeline, and LLM extraction methods. Section 3 describes the data ontology and multi-dimensional impact taxonomy. Section 4 presents validation results through comparison with existing databases and detailed case studies. Section 5 discusses limitations and future work. Section 6 concludes with implications for forecast verification and disaster management.

2 FloodEcho Workflow

FloodEcho is organized into three modules as shown in Figure 2.1. First, a spatial-temporal trigger based on hydrometeorological records is applied to search for potential events on the identified locations and on the target dates (Module 1). Then, a **DeepSearch** process is initiated to obtain the information available online about the occurred socioeconomic impacts. This process comprises two modules: the first aims to retrieving the sources (Module 2) and the second (Module 3) aims to extract the specific details of the impact through the use of large language models (LLMs). The process has been running on a daily basis across European regions.

The final list of identified events with the retrieved socioeconomic impacts are marked as identified events. A severity score is assigned to these events, and the impacts are categorized into four classes: transport, economy, safety and emergency response.

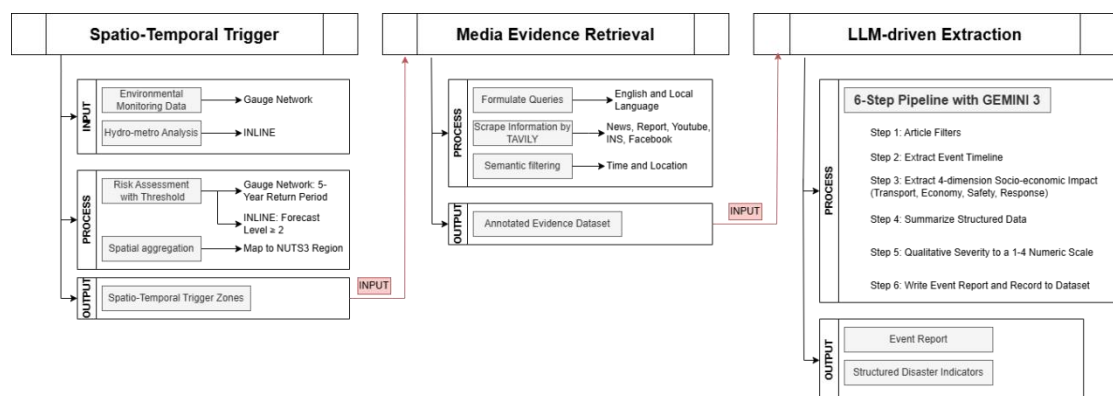


Figure 2.1: Workflow designed for FloodEcho.

2.1 Module 1: Spatial-temporal triggering

Instead of massively searching for information on flash flood impacts in each European region on a daily base, FloodEcho uses a spatial and temporal trigger. This module pre-identifies regions where potential events may have occurred on the analysis date using hydrometeorological information (i.e., observations and simulations).

The current version of FloodEcho integrates two sources of hydrometeorological information to trigger the search for potential impacts, the region is activated when

these products reach a certain threshold:

- The daily accumulations recorded at the EFAS rain gauge networks (composed of more than 20000 rain gauges across Europe). The region are identified when the 24-h rain gauge accumulations exceed the estimated local 5-y return period. That is, if a rain gauge in a region exceeds this 5-year return period, that region is identified for the given date.
- The daily summary of the INLINE flash flood impact analysis (over the river network, with a resolution of 1 arcmin). INLINE flash flood impact product adopts the TAMIR framework (Copernicus Emergency Management Service, 2024), which combines flash flood hazard with European-wide spatial exposure data. The flash flood hazard is generated by blending hourly ensemble radar nowcasts and ensemble NWP precipitation forecasts, which are accumulated on the river network and compared against climatological thresholds (Park et al., 2019). The exposure data combines population and critical infrastructure to estimate where potential impacts may be greatest. The analysis used as the trigger for this study is the INLINE flash flood impact at a 0-hour leadtime, which captures the actual impact conditions at the time of observation. The potential event is identified based on the daily maximum impact level on the river network. Specifically, a region is pre-identified as potentially affected by floods if more than 10% of its river network reaches an INLINE impact level 2 (i.e., the 90th percentile of the regional impact distribution exceeds level 2) on the analysed date.

Once any of these datasets exceeds its specific threshold in a given region, it activates the subsequent two modules to confirm that flash floods have affected the targeted region on the analysed date and to retrieve the details of their impacts. Any region and date meeting these criteria is classified as a potential event.

2.2 Module 2: Targeted media evidence retrieval

For a given search date and in those regions where potential events have been identified by Module 1 (as described in Section 2.1), this module performs a **deep search** to establish a collection of online sources (URLs). The actual extraction of socioeconomic impacts from this content is delivered to Module 3. To build this evidence-based link dataset regarding the occurred impacts, FloodEcho integrates a search engine that scrapes web content from official sources, news outlets, and social media platforms.

The search engine used here is Tavily (Tavily, 2026), which specifically designed for Large Language Models (LLMs). Unlike traditional search engines built for human browsing, Tavily is optimized to efficiently retrieve web content **tailored** for AI system. Official government disaster announcements, mainstream news reports, local social media channels and social media platforms (including Facebook, Instagram, X, TikTok and YouTube) have been considered as useful sources and recorded to the platform.

To implement this retrieval process, FloodEcho utilizes the Tavily Search API. The API is invoked using structured search queries for each potential event (i.e., the specific region and date identified in Module 1) with targeted flood related keywords. The

information sources are retrieved by imposing two primary conditions: they must mention specific locations and dates (or times) that closely coincide with the potential event; and they must include specific thematic keywords, such as "rainfall" and "flood".

To ensure comprehensive coverage, this search and matching process is conducted in both English and the local language of the affected region. This yields two distinct lists of retrieved URLs—the top 20 links from the English search and the top 20 from the local language search—resulting in an initial pool of up to 40 potentially relevant links per event.

However, the raw web results vary significantly in reliability. Once this initial link pool is established, FloodEcho ranks the combined list based on the degree of matching with the region and date of the potential event. A link verification pathway (comprising Stage 1 and Stage 2) has been applied in this study to maximise the suppression of prevalent noise in open web data; the specifics of this verification are detailed in Section 2.4.

Links that successfully pass this pathway are compiled into an "Annotated Evidence Dataset", which serves as a curated repository of highly relevant, verified web sources. Finally, to maintain data quality and processing efficiency, only the top 10 highest-ranked links from this Annotated Evidence Dataset are forwarded to Module 3.

2.3 Module3: LLM-driven Multi-dimensional Extraction

The purpose of this step is to extract the details of flood impacts from the links resulting in module 2 and classify them into four classes: Transport (e.g., road closures, bridge damage), Human Safety (e.g., casualties, evacuations), Economy (e.g., industrial disruption, infrastructure damage), and Emergency Response (e.g., warning levels, rescue deployments), and to assess the severity of the event in general and in each class. This is achieved with the help of a Large Language Model (LLM).

In FloodEcho, Google Gemini-3-Flash-Preview is used via structured prompt design to ensure consistent, standardized data extraction. The LLM is used to extract two types of information: (1) spatio-temporal attributes (i.e., when and where the reported flood took place) to cross-verify consistency against the objective target event defined in Module 1 (this belongs to the link verification pathway Stage 3 detailed in Section 2.4); and (2) detailed observations of the actual socioeconomic impacts to assess the severity of the identified event. From this data, those links that show discrepancies with the objective date and location (i.e., those resulting from Module 1), or that have no content in terms of the occurred impacts, are filtered out and replaced by the next link in the list.

To comprehensively evaluate the consequences of flood events, this study categorises disaster impacts into four classes, Transport, Human Safety, Economy and Emergency Response, to construct a multidimensional dataset of impacts. This classification is conceptually grounded in the disaster risk reduction framework of the Sendai Framework (UNDRR, 2015) and operational impact-based warning guidelines from the World Meteorological Organization (WMO, 2015; 2021), while being

specifically suited for online media streams. Emergency response has been added to track the operational actions for the disaster management.

Specifically, as shown in Table 2.1, Transport focuses on functional mobility and network disruptions (e.g., road blockages, railway cancellations, and structural bridge collapses), serving as the earliest and most frequently reported spatial proxy for surface inundation in both urban and rural environments (Pregnotato et al., 2017). Human Safety captures direct, irreversible human consequences (e.g., casualties, trapped individuals, and precautionary or forced evacuations), representing the paramount priority in hazard assessment (Jonkman et al., 2008). Economy covers broader direct asset damage and secondary systemic disruptions, including industrial stoppages, agricultural losses, and critical utility failures (e.g., power grids, water treatment plants, and telecommunications) (Merz et al., 2010).

While traditional impact assessments often focus on the physical damage, we incorporate Emergency Response (e.g., elevation of civil protection alert levels, deployment of military/firefighting rescue assets, and barrier installations), which captures the emergency management in real-time and serves as a robust and reliable indicator of the event's on-the-ground severity. By capturing direct impacts on people and infrastructure (Transport, Human Safety), indirect economic losses (Economy), and institutional actions (Emergency Response), this four-class system maintains clear boundaries for LLM extraction while providing a comprehensive basis for forecast verification.

Table 2.1: Four-class flood impact categorization used for LLM-based media extraction

Class	Definition	Related phenomena	Prompt cue terms
Transport	Mobility & transport infrastructure impacts	Closures, bridge damage, network paralysis	road closures, traffic disruptions, bridge damage, highway closures
Human Safety	Human & residential harm	Casualties, evacuations, housing damage	casualties, injuries, deaths, evacuations, house damage
Economy	Financial & productive consequences	Losses, industry/business interruption, agricultural damage	economic loss, damage estimates, business impact, agricultural damage
Emergency Response	Official mobilisation & response urgency	Alerts, rescue, emergency deployment	emergency measures, alerts issued, rescue operations

In FloodEcho, an LLM-based extraction procedure is applied to automatically identify and extract specific impact information from unstructured media sources. The four impact classes are defined through operational definitions in the extraction prompt, under which the model performs multi-label assignment. Each class is associated with structured fields in the expected output format, so the result is

constrained by a reproducible extraction schema rather than free-form summarisation. Classification therefore operates on consequence facts: a single article may contribute to several classes simultaneously, while an unmentioned class is left null to indicate insufficient information.

The prompt also provides cue terms to guide extraction within each class—for example, road closures and bridge damage for Transport; evacuations, and house damage for Human Safety; economic loss and business impact for Economy; casualties; and emergency measures, alerts, and rescue operations for Emergency Response. These cues orient the model toward relevant evidence. Within each class, the model prefers quantitative statements when available and otherwise retains concise qualitative descriptions. The LLM extraction process is therefore a prompt-specified, guided multi-label extraction over validated media evidence, combining soft cue terms with operational class definitions.

Finally, the system has been trained to translate the semantic descriptions retrieved with the LLM into numerical scores assessing the impact level for each impact class: into a four level scale: 1-low, 2-moderate, 3-high and 4- severe.

Within each impact class, a prompt-specified ordinal scale has been applied to the LLM. These criteria combine quantitative thresholds, when available, with qualitative anchors: Transport severity depends on whether closures affect secondary roads or critical arteries; Human Safety on the scale of casualties and evacuations; Economy on the density and duration of productive disruption; and Emergency Response on alert level and the extent of institutional mobilisation (Table 2.2).

The model prefers numerical evidence when present and otherwise infers severity from concise descriptive language against the same rubric. If a specific impact class is not mentioned in the source text, it is strictly recorded as a null value, preventing unreported data from artificially lowering the assessment. The resulting labels are then mapped deterministically to integer scores from 1 to 4, yielding a reproducible category-level severity assessment. The average score from the four classes is calculated and saved as the overall impact level.

Table 2.2: Class-specific severity score for LLM impact assessment (Levels 1 – 4)

Impact class	1-Low	2-Moderate	3-High	4-Severe
Transport	Minor traffic jams, surface water, cautionary speed limits; no closures	Secondary roads closed; delays, but main network still functions	Regional railway suspended; bridge damage; 1–2 main (national) highways blocked < 24 h	Total paralysis: critical artery severed; airport/port closed; systemic rail collapse; damage needing weeks to repair

Human Safty	No injuries, deaths, or evacuations	Minor injuries; precautionary evacuation < 100 people	Confirmed fatalities 1–10; significant injuries; evacuation 100–1000	Mass casualties > 10 deaths; mass evacuation > 1000; towns cut off/isolated; missing persons
Economy	Resident inconvenience only (puddles, jams); no substantial work stoppage	Low/medium-density areas affected, or short-term blockage in high-density areas	Economic hubs / industrial zones hit, or widespread infrastructure damage	Core tier-1 cities or national economic hubs paralyzed
Emergency Response	Routine monitoring; no special alerts	Yellow/Orange alert; local fire/police handling	Red alert; regional forces deployed; specialised rescue (helicopters/boats)	State of emergency/catastrophe; military deployed; international aid requested

Additionally, the system includes a "Global Assessment" field to independently assess the event's severity level. Global assessment is a macro-level judgement produced in the same extraction step as the four impact classes. It analyzes the situation by combining all information together to assess whether the regional systems remained manageable or were overwhelmed during the event.

The ordinal scale is defined in the prompt as: Level 1 (Manageable)—severe weather that local resources can handle with urban functions largely intact; Level 2 (Strained)—noticeable disruption requiring non-routine measures while essential systems still hold; Level 3 (Overwhelmed)—breakdown of essential systems such as transport, power, or logistics, such that normal functioning depends on external support; and Level 4 (Collapsed)—systemic failure of historical magnitude with recovery measured in months or years.

In summary, the final output generated by FloodEcho for each identified event is a structured record containing two main parts: (1) concise text descriptions in bullet points detailing the specific consequences observed in each of the four classes and global assessment; (2) corresponding numerical severity scores (from Level 1 to Level 4, or null if unmentioned) for each class.

2.4 Link Verification Pathway

Open web data often contains significant noise, such as outdated events, incorrect locations, or weather forecasts which can be mistakenly interpreted as actual impacts.

To filter out this noise while minimising the computational costs of Large Language Models (LLMs), FloodEcho employs a Three-Stage Link Verification pathway consisting of (1) rule-based pre-filtering, (2) page-level verification, and (3) LLM-driven semantic validation.

Table 2.3: Overview of the three-stage link verification pathway

Stage	Host Module	Purpose	Applied Method	Max remaining links
Upstream collection	Module 2	Multi-source links collection	Tavily	20
Stage 1: Rule-based pre-filtering	Module 2	Remove obvious noise with time, location and topic	Rule-based method	15 (media ≤ 3)
Stage 2: Page-level verification	Module 2	Enforce publication time check	Newspaper3k	10
Stage 3: LLM-driven semantic validation	Module 3	Final relevance check and rank links with relevance score	Prompt	10

2.4.1 Stage 1: Rule-based Pre-filtering

In the first step, candidate links collected by Tavily are filtered out based on their title, URL, and summary using a rule-based method to avoid spatio, temporal and topical mismatch, and reduce the workload for the subsequent compute-intensive text scraping and semantic understanding phases in Module 2 (section 2.2).

This process serves three main goals: First, for quota control and deduplication, the system limits the maximum number of social media links to three to prioritize authoritative journalism, and uses exact URL matching and normalised title similarity to avoid cross-channel redundancy. Second, temporal constraints are applied by extracting URL publication dates from metadata to verify if the link falls within the time window between two days before the searched date and up to seven days after. If the date cannot be extracted, the link is provisionally passed to the stage 2 for further checking. Third, the system conducts keyword filtering. Links are strictly required to contain flood-specific disaster keywords translated into the local language of the affected region (e.g., terms corresponding to "flash flood," "inundation," or "overflow"). Concurrently, a predefined exclusion list (negative dictionary) immediately discards search results containing non-impact indicators, such as routine weather forecasts, early warnings without reported consequences, or retrospective climatic summaries (e.g., "weather report," "forecast," "historical statistics," or "annual review").

Ultimately, this stage retains only those links that pass the time matching, disaster term verification, and absence of negative keywords. A maximum of 15 candidate links from each search in Module 2 are processed to the following stages

2.4.2 Stage 2: Page-Level Verification

Given that metadata publication dates provided by search engines are often missing, delayed or erroneous, the main objective of this step is to perform a more stricter temporal verification using the actual content of each link. This stage is also executed in Module 2.

The system uses the Python library `newspaper3k` (Ou-Yang, 2018) to download each candidate URL concurrently and parse the HTML page. Beyond the search-engine title and snippet retained from Stage 1, it extracts three additional fields from the page itself: (i) the main article body text, (ii) a page-level publication timestamp when available, and (iii) the lead image when present. The scraped body text (maximum 3,500 characters) replaces the short search snippet whenever it is longer, providing richer narrative context for subsequent extraction.

The publication time is recovered in a fixed priority order. First, `newspaper3k`'s built-in parser is used. If that fails, the raw HTML is scanned for embedded publication tags (for example Open Graph `article:published_time` and Schema.org `datePublished`). If those are also absent, day-level date patterns in the URL path (for example `/2025/10/07/`) are used as a last resort. When the publication time is obtained, the link is kept only if it falls inside the event window between two days before the searched date and up to seven days after, otherwise it is rejected.

If no publication date can be recovered, the link is marked as lacking a reference date and passed to the following step. Furthermore, if the scraped text mentions past disaster years but lacks the current target event year, the system flags it as historical noise and discards it.

The successfully scraped full texts provide richer context for the downstream LLM. A maximum of 10 verified links are retained for the final LLM processing stage.

2.4.3 Stage 3: LLM-driven Semantic Validation

Stage 3 is executed in Module 3, where the LLM performs a comprehensive semantic evaluation on the remaining links to construct a curated Evidence Dataset. Beyond basic topical relevance, the LLM determines whether each report constitutes usable, ground-truth evidence of occurred flood impacts for the target event.

The model cross-verifies date and location clues across the URL, title, metadata, and full text against the target event defined in Module 1. Crucially, the LLM distinguishes between "observed consequences" and "prospective meteorological information": links reporting solely weather alerts, preparedness measures, routine weather descriptions, or same-storm impacts outside the targeted territorial unit are strictly discarded. Conversely, if an article lacks an explicit publication timestamp but details highly specific, localized disaster consequences, the temporal constraint can be adaptively relaxed to preserve valuable ground-truth evidence.

To achieve these judgements, the LLM is guided by a structured validation prompt that forces an explicit, multi-label decision for each link reached this validation stage. For each link, the model must return controlled attributes covering (i) whether the content is a published report, (ii) the nature of the reported information (e.g., `observed_impact`, `official_confirmed`, `forecast`, `warning`, `preparedness`, or routine `observed_weather`), (iii) geographic scope relative to the target territorial unit, and (iv)

temporal scope relative to the Module 1 event window. The model accepts a link as usable evidence only when the report describes already occurred or officially confirmed flood consequences in the target area and within the same event episode.

To rank the evidence link, the system calculates a final relevance score S in $[0, 1]$ using a linear weighting of four-dimensional features:

$$S = 0.32 \cdot S_{\text{time}} + 0.28 \cdot S_{\text{source}} + 0.25 \cdot S_{\text{location}} + 0.15 \cdot S_{\text{content}}$$

Here, S_{time} represents temporal evidence strength, S_{source} is source authoritativeness, S_{location} measures geographic matching granularity, and S_{content} evaluates content fingerprint strength. The weights reflect a strict hierarchy. Time holds the highest weight, prioritising structured metadata. Source evaluates institutional credibility (e.g., official agencies over user-generated content). Location measures geographic matching granularity. Content quantifies the specificity of disaster details (e.g., casualty numbers), acting as a fallback for items with ambiguous dates.

To eliminate severe false positives, hard thresholds are applied: links are immediately rejected. If the temporal score equal to zero, which means it is a historical event or it outside the event time window, the final score will be set to zero also, which means this link will be delated. Besides, if the score for location is zero, which indicate the geographical irrelevant, the final score should be less than 0.3. Ultimately, this score ranks the sources in descending order, ensuring that higher-scoring, more accurate links carry more weight in the final impact extraction.

Through this three-stage pathway, FloodEcho progressively removes outdated reports, geographically mismatched pages, forecast- or warning-only items, and other non-impact noise. The links retained at the end of the pathway are therefore both temporally and geographically consistent with the target rainfall–flood event and contain usable descriptions of already observed or officially confirmed impacts. This cleaned Evidence Dataset then provides a reliable input for the subsequent impact extraction.

2.5 Implementation Consideration

2.5.1 Regional Resolution

The results of FloodEcho are aggregated and analyzed at the NUTS-3 level for most European countries, and at the NUTS-2 level for the United Kingdom and Germany, as NUTS-3 units in these two countries are too small for regional-scale media extraction.

In the current demonstration of FloodEcho, a few countries (such as Italy) are excluded due to limited radar quality or coverage gaps, which could otherwise bias the event retrieval process. The monitored geographical domain and regional administrative units are shown in [Figure 2.2](#).

2.5.2 Near Real-time Implementation

All events stored in FloodEcho have a daily temporal resolution: observational and forecast inputs are ingested on a day-by-day basis for each identified territorial unit.

Because consecutive wet days within the same region typically represent a single sustained flood episode rather than separate events, the operational pipeline applies a 1-day temporal clustering rule: records in the same region on consecutive calendar dates are merged into a single multi-day episode, while a gap of at least one day initiates a new group. A short protection window relative to the latest available data delivery keeps incomplete episodes open until subsequent daily updates confirm whether the event has ended.

After the identification of potential events, FloodEcho triggers a media deep search for events pending impact assessment. A 7-day search window is used to accommodate the typical lag between physical flooding and news or social reporting, as well as multi-day follow-up coverage of the same episode. Within this near-real-time implementation—daily input, 1-day episode clustering, and 7-day delayed deep search—the system links events detection to qualitative impact evidence. **Figure 2.2** presents the temporal distribution of FloodEcho events from January 2025 to February 2026, illustrating the density of detected episodes under this operational configuration.

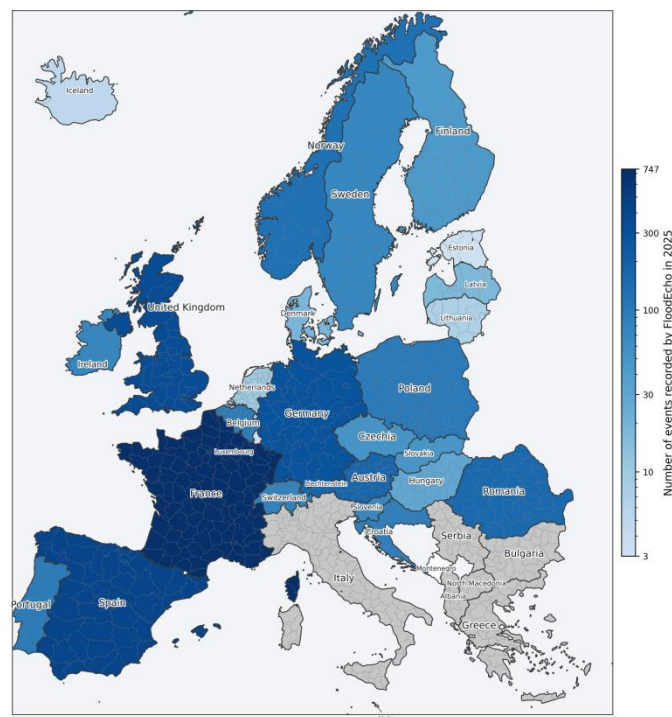


Figure 2.2: Spatial distribution of flood events recorded by FloodEcho across Europe in 2025. (Monitored administrative units (NUTS-3 for most countries, and NUTS-2 for Germany and the United Kingdom) are colored using a logarithmic scale, where darker shades of blue indicate a higher number of recorded events (ranging from 3 to 747). Non-monitored regions (e.g., Italy and parts of southeastern Europe) are shown in grey due to limited radar coverage or data quality gaps.)

3 FloodEcho System Demonstration

3.1 Event Spatial-Temporal Distribution

To demonstrate the operational performance of the FloodEcho system, a one-year

continuous demonstration was conducted from January 1 to December 31, 2025. During this period, the system registered 3,271 potential flood events based on hydrometeorology observations and estimates. After deep search processing, FloodEcho resulted in 509 Events causing actual socio-economic impacts (as explained in section 2).

Figure 3.1 shows the sources triggering the identification of potential flood events. Across all potential events, the INLINE estimated flash flood impact served as the primary driver compared to rain gauge observations, with 10% of total events triggered simultaneously by both datasets. However, this distribution shifts notably when isolating the 509 confirmed impact events. While the INLINE flash flood impact estimates remain the most frequent source for events with flash flood impacts, their share decreased by 10%. Conversely, events triggered exclusively by gauge observations accounted for around 40% of the detected events, and those triggered by both sources increased from 10.7% to 16.2%. This shift indicates that while the INLINE flash flood impact estimates reports provide high sensitivity for initial anomaly detection, the situations detected from rain gauge observations or validated where both sources show a clear signal are more likely to result in socioeconomic impacts.

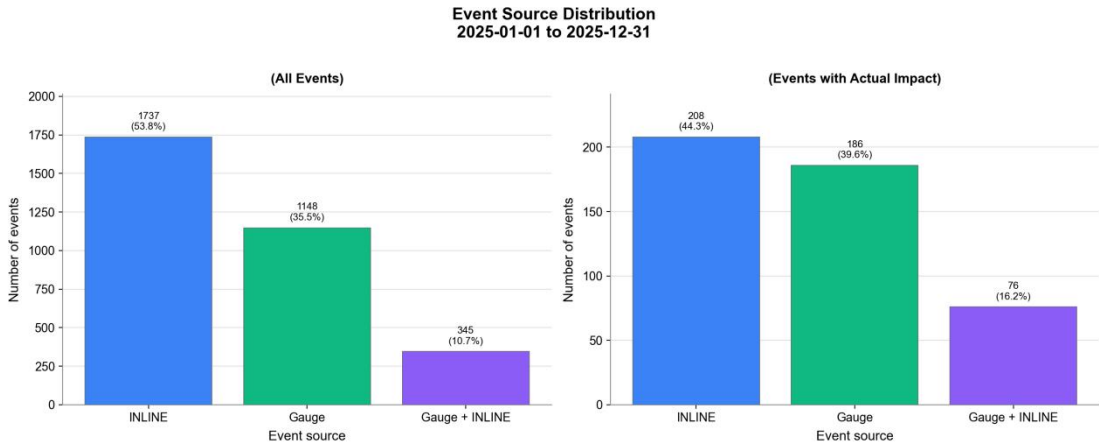


Figure 3.1: Event Source Distribution during the FloodEcho Demonstration. **The figure on the left presents the event**

Spatially, the 509 confirmed impact events were widely distributed across Europe but show distinct regional clustering (Figure 4.2). Country-level aggregation reveals that Spain is the state with the largest number of events, experiencing more than double the number of impact events compared to any other country, followed by France, Germany and Romania (Fig. 4.2b). Regarding event severity, the impact level of the confirmed events was medium and high in most cases. FloodEcho Level 3 impacts were the most frequent, representing 52.1% of the dataset, followed by Level 2 impacts at 37.1%. Notably, FloodEcho also retained lower-severity cases, with minor Level 1 events accounting for 5.7%, close to the 5.1% share of extreme Level 4 events.

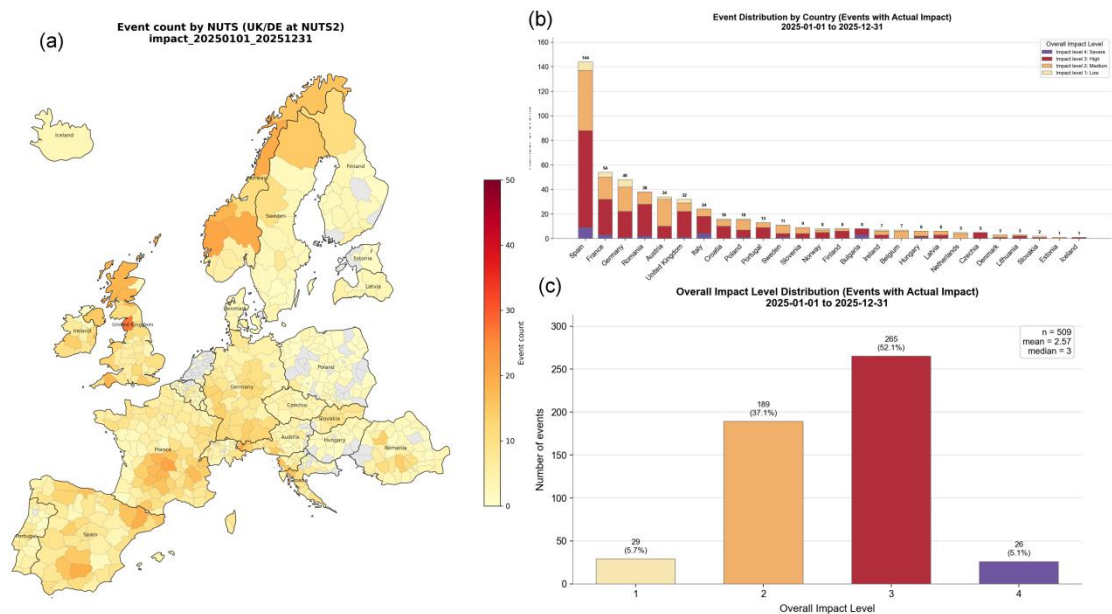


Figure 3.2: Spatial and impact level distribution of FloodEcho outputs for year 2025. (a) Distribution of the number of events with real impact during 2025 by FloodEcho. (b) Number of events by country and impact level within each country. (c) Overall impact level distribution within the events with actual impact.

The temporal distribution of flood events throughout 2025 exhibited pronounced seasonal variability, as shown in Figure 4.3. The total number of potential events peaked in the summer and autumn months, with the highest counts in July, September, and November. February records the lowest monthly total, with 68 potential events. When isolating events with identified socioeconomic impacts, July remains the most active month with 107 confirmed events.

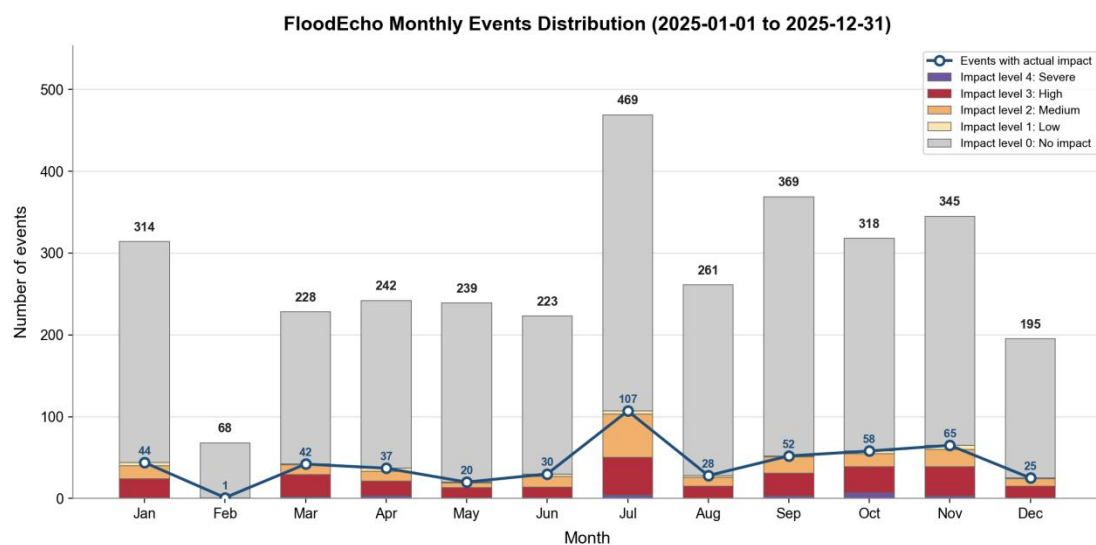


Figure 3.3: Monthly distribution of all events during 2025 by FloodEcho. The number on top of each bar indicate the total number of potential events in each month and the number in the middle show the number of identified events.

3.2 Case Study: Barcelona Event - 21 September 2025

To illustrate the results obtained with FloodEcho, a representative case study is provided for the rainfall event that affected the region of Barcelona (Spain) on 21 September 2025. The Barcelona event was pre-identified in FloodEcho through both two triggering source: the INLINE flash flood impact estimates and rain gauge observations. During this event, the maximum 24-hour rainfall accumulation reached 129.3 mm, closing to 10-year return period threshold. The INLINE flash flood impact estimates assigned the event a Level 4 signal in five subcatchments. Figure 4.4 shows the surrounding rain gauge observations together with the corresponding INLINE flash flood estimates for this event.

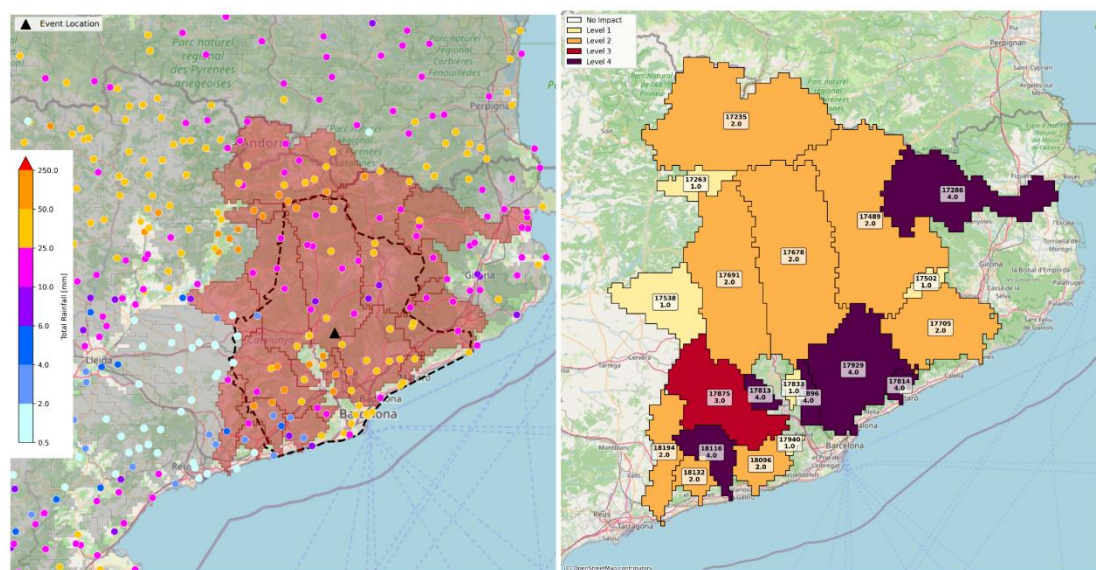


Figure 3.4: (Left) Rain Gauge distribution around the Barcelona area and daily rainfall accumulation during the event. (Right) Maximum INLINE flash flood estimates around Barcelona during the event.

Figure 3.5 presents how the event appears in the platform after the system searches for real-world consequences. The upper panel (3.5a) reports detailed information for this event, including the affected period, location, trigger, and the final result after searching for ground consequences. The event pop-up (Figure 3.5b) summarises the four extracted impact dimensions: transport, economy, safety and emergency response.

The table also shows the evidence sources for each impact type and the impact level assigned to each of them. For the Barcelona event, the impact on transportation reached the highest level, while the impact on people's safety reached only level 2. The remaining were assigned level 3.

The detailed impact list will be presented when an impact block is selected (Figure

3.5c). In the transport dimension, the system records multiple highway disruptions caused by flooding and rainfall-triggered landslides, as well as temporary disruption at the airport. These transport interruptions also contributed to the economy dimension through logistics delays and flight cancellations. In the safety dimension, no injuries or deaths were reported across the extracted sources, supporting the lower level 2 classification. The emergency response dimension records 200 emergency calls and public warnings advising people to avoid outdoor activity. Overall, FloodEcho classified the Barcelona event as level 3. The extracted evidence indicates a rainfall event with high accumulation with landslides and substantial transport consequences.

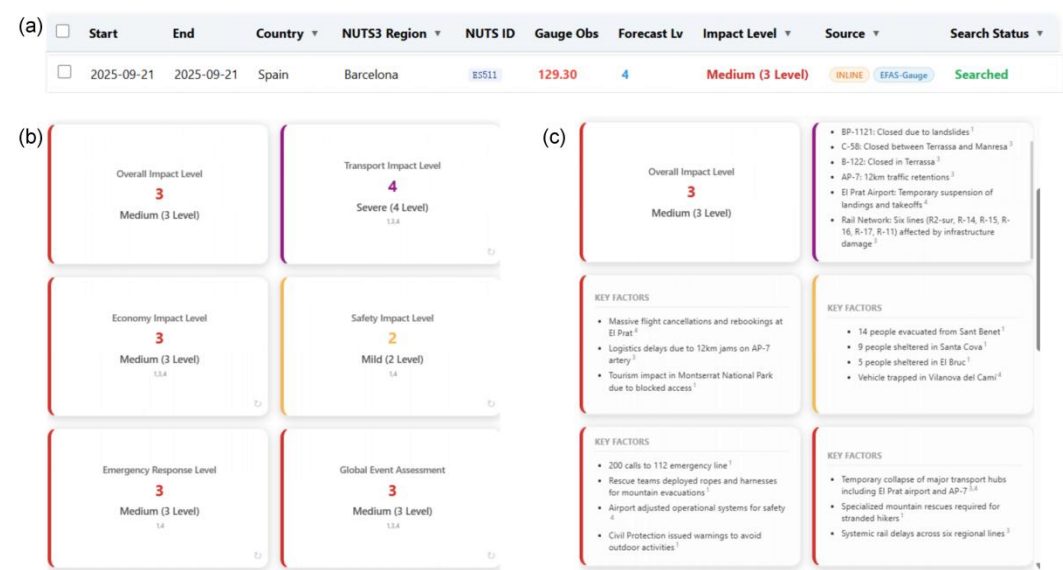


Figure 3.5: FloodEcho record for the Barcelona Event on 21 September 2025.

All validated sources used to extract the detailed impact information for this event were retained in the FloodEcho platform. For the Barcelona event, six useful sources were saved, with confidence scores ranging from 0.97 to 0.63. Four sources were news websites, and two were Facebook posts; both English- and Spanish-language links were presented in the final evidence set (Figure 4.6).

The link verification pathway (**as presented in**) for this case also illustrates the function of the Three-stage Cascading Link Verification mechanism. The six final evidence links were selected from 40 Tavily search results. After rule-based pre-filtering, 15 links remained. Page-level verification then removed six links with mismatched or unreliable publication dates, and the final LLM-driven semantic validation removed three additional links that contained forecasts or weather reporting information rather than evidence of observed impacts.

Manual review of all 40 candidate links confirmed that the six retained links were valid for impact extraction. The review also identified two false negatives, which means the link could be used for impact extraction but was removed by the link verification pathway: both valid links were removed during page-level verification because their publication dates were uncertain. Although these two valid links were incorrectly

excluded, they did not compromise the integrity of the final case classification. This illustrates that the filtering process effectively maintains data reliability despite occasional minor information loss.

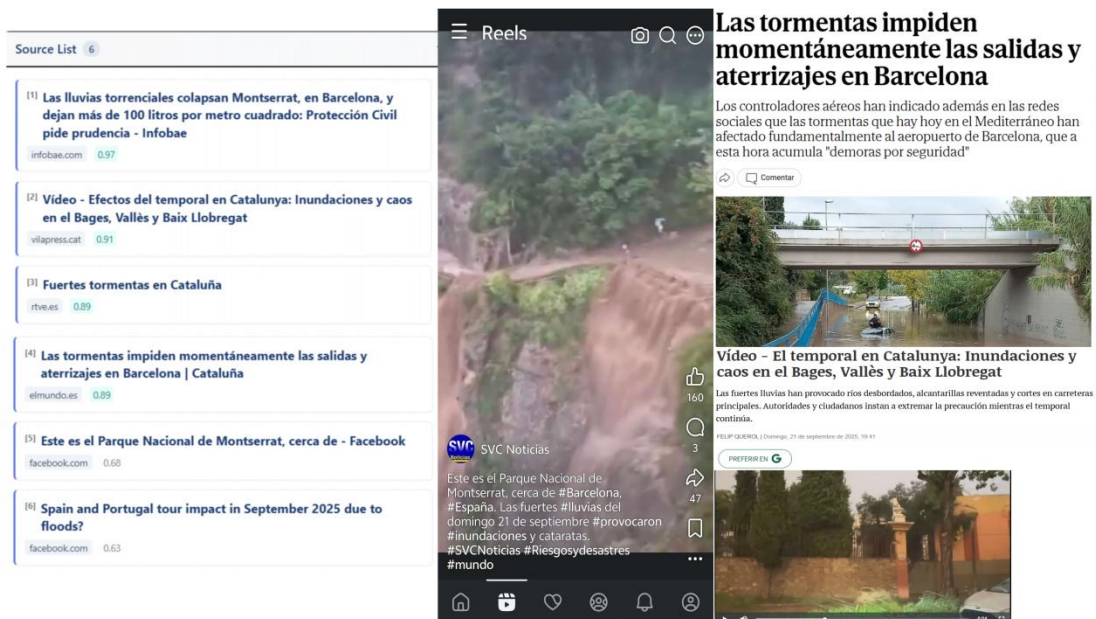


Figure 3.6: Links saved by FloodEcho for the Barcelona Event

All events saved in the FloodEcho have daily resolution, and the result will be stored and analysed at Nuts3 level for most of european country and at Nuts2 level for United Kindom and Germany, while nuts3 area for these two country is too small. In the demonstration of FloodEcho, some country haven't been monitored because of the bad radar quality such as Italy, which may have effect of the event retrieve.

4 Validation and Comparative Analysis

To assess both the methodological robustness of the FloodEcho workflow and the reliability of its extracted flood records, this section presents a comprehensive evaluation across three key aspects: a comparative benchmark against two established databases (HANZE and Groundsource), the link verification pathway and the precision of LLM scoring. A benchmark set of 100 flood events from 2025 was manually reviewed to serve as the ground truth for this evaluation.

4.1 Cross Validation with Hanze Database and Groundsource

To evaluate the performance and operational coverage of FloodEcho, we benchmarked its event catalog against two established flash flood databases: HANZE and Groundsource. The HANZE database provides historical assessments of natural hazards in Europe by systematically compiling disaster records from official yearbooks, administrative reports, and scientific literature, with a primary focus on extreme events. In contrast, Groundsource was developed as a global, AI-driven event repository covering more than 150 countries; it automates the ingestion of millions of multilingual news reports using LLMs and specializes in resolving localized flood events

by generating spatio-temporal boundary polygons. FloodEcho, designed as an operational impact verification database, which can be compared together with these two database.

In order to enable a direct comparison across all three databases, all recorded events were standardized to the spatio-temporal resolution of FloodEcho. For Groundsource, the analysis was restricted to events with an affected area exceeding 10 km². Figure 4.1 illustrates the spatial distribution of all events recorded in 2025 across Europe at this unified scale. Groundsource records the highest event density, followed by FloodEcho, while HANZE exhibits the sparsest coverage.

In 2025, following domain and threshold filtering, the three datasets display marked differences in scale and coverage. Groundsource cataloged the largest number of events (2,714), FloodEcho recorded 450 verified real-impact events, while HANZE recorded only 33 events. Temporally, both FloodEcho and Groundsource exhibit an activity peak in July with a secondary concentration in autumn, whereas HANZE records are more temporally dispersed with a slight increase in spring. Spatially, FloodEcho events are predominantly concentrated in Spain, followed by France, Germany, Romania, and Austria. Groundsource is led by France and Spain, followed by Poland, the UK, and Romania. HANZE primarily covers Spain, France, Romania, Croatia, and Italy. In general, Groundsource serves as a high-density monitoring layer, FloodEcho captures fewer but impact-verified events, and HANZE provides a small, highly selective set of major events.

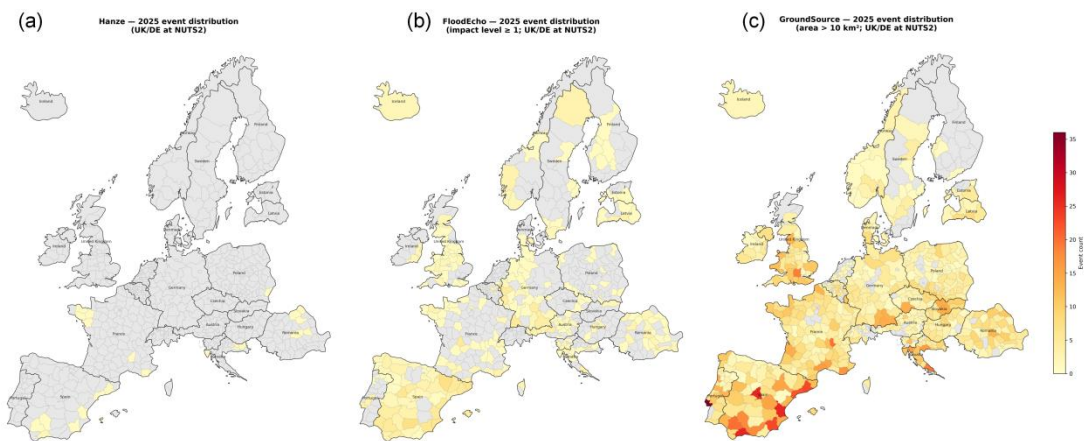


Figure 4.1: Spatial distribution of flood events recorded by three database across Europe in 2025. (a) presents the events recorded by Hanze database. (b) presents the events recorded by FloodEcho and (c) presents the events recorded by Groundsource database.

To systematically assess the detection efficacy of FloodEcho and evaluate cross-source consistency, a comparative classification was conducted on a total of 5,203 event instances across the three sources. As shown in Table 4.1, the records were categorized into 11 distinct groups based on detection combinations across the sources. Events jointly identified by all three sources account for a small fraction of the total (Group 1, 8 events, 0.15%). Co-detection rates were higher between pairs of sources: 197 events were captured by both FloodEcho and Groundsource (Group 6,

3.79%), and 14 events were shared by HANZE and Groundsource (Group 3, 0.27%). However, no events were shared exclusively between HANZE and FloodEcho (Groups 2 and 10, 0.00%).

In contrast, the majority of detections were reported by a single source, particularly Groundsource. Group 7 (Groundsource only) comprises 2,012 events (38.67%), whereas FloodEcho uniquely identified 240 events (Group 5, 4.61%). Furthermore, cases identified by FloodEcho as potential events but without confirmed impact account for more than half of the total records; these predominantly fall into Group 8 (detected by Groundsource, but categorized as zero/no-impact by FloodEcho; 473 events, 9.09%) and Group 11 (potential events with no recorded impact across all sources; 2,249 events, 43.23%). To verify the physical authenticity of these events and evaluate false-alarm dynamics, a representative stratified sample of 100 events across these groups was manually evaluated (Table 4.1), as detailed in Section 4.2.

Table 4.1: Cross-comparison and validation of flood event detection across multiple data sources

Group	Hanze	Flood Echo	Ground source	Event Count	Distribution	Sampled (n)	True Event	False Alarm
1	✓	✓	✓	8	0.15%	5	5	0
2	✓	✓	×	0	0.00%	0	0	0
3	✓	×	✓	14	0.27%	5	5	0
4	✓	×	×	9	0.17%	5	5	0
5	×	✓	×	240	4.61%	25	14	11
6	×	✓	✓	197	3.79%	20	20	0
7	×	×	✓	2012	38.67%	15	2	13
8	×	–	✓	473	9.09%	19	2	17
9	✓	–	✓	1	0.02%	1	1	0
10	✓	–	×	0	0.00%	0	0	0
11	×	–	×	2249	43.23%	5	1	4
Total				5203	100.00%	100	55	45

4.2 Manual Evaluation of 100 Selected Events

A representative sample of 100 events was manually verified against web records and local official news sources, to evaluate detection capability across different grouping combination outlined in Table 4.1. Through this ground-truth validation, 55 cases were confirmed as real flood events, while the remaining 45 were classified as non-events.

Across all subsets where an event was co-detected by two or more sources (Groups 1, 3, 6, and 9), manual validation yielded a 100.0% verification rate. Notably, when FloodEcho coincided with Groundsource (Group 6), all 20 sampled instances were validated as true flood events without any false alarms. Within Group 5 (events captured solely by FloodEcho), 14 out of 25 sampled instances were confirmed as true flood occurrences, corresponding to a true positive rate of 56.0%. Conversely,

candidate events recorded only by Groundsource (Group 7) or instances where Groundsource detections coincided with no recorded impact in FloodEcho (Group 8) exhibited markedly lower true positive rates, with false alarms dominating both subsets.

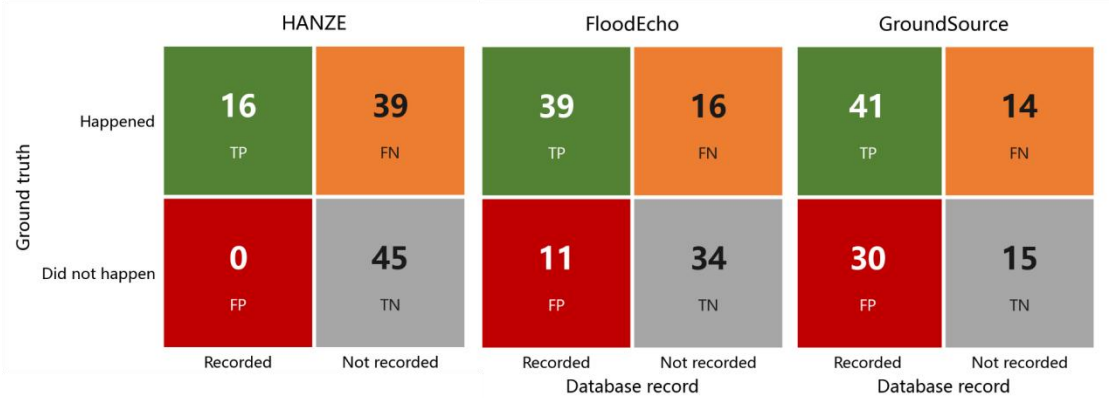


Figure 4.2: Confusion matrices of HANZE, FloodEcho, and Groundsource evaluated on the 100 validated flood event samples. Green denotes true positives (TP, correctly recorded actual events); orange presents false negatives (FN, missed actual events); red indicates false positives (FP, falsely recorded non-events); while grey shows true negatives (TN, correctly identified non-events).

Figure 4.2 presents the comparative confusion matrices for HANZE, FloodEcho, and Groundsource evaluated against the 100 ground-truth samples. The vertical axis represents the manually verified occurrence, while the horizontal axis indicates whether the event was cataloged by the respective database. The four quadrants correspond to True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN). The HANZE database successfully identified 16 true events with zero false positives, correctly rejecting 45 non-events; however, it suffered from severe under-detection, missing 39 actual events. Groundsource achieved high recall by capturing 41 true events, but this came at the expense of substantial noise, incurring 30 false alarms alongside 15 true negatives. In comparison, FloodEcho captured 39 true events with 16 misses, while incurring only 11 false alarms and correctly discarding 34 non-events.

To quantify detection performance, three standard verification metrics were calculated (Table 4.2): Probability of Detection (POD), False Alarm Ratio (FAR), and the Critical Success Index (CSI). HANZE achieved an ideal FAR of 0.00, confirming that every recorded event was really happened. However, its low detection rate (POD) decrease its overall skill, yielding a CSI of only 0.29. Conversely, Groundsource attained the highest POD of 0.75, but its vulnerability to web noise resulted in a high FAR of 0.42, which lowered its CSI to 0.48. FloodEcho achieved an optimal balance between sensitivity and precision, delivering a robust POD of 0.71 while suppressing the FAR to 0.22. Consequently, FloodEcho secured the highest overall performance, with a CSI of 0.59.

Table 4.2: Evaluation metrics of Hanze, FloodEcho, and Groundsource evaluated on the validation samples

Evaluation Metrics	Hanze	FloodEcho	Groundsource
--------------------	-------	-----------	--------------

POD	TP/ (TP+FN)	0.29	0.71	0.75
FAR	FP/(TP+FP)	0	0.22	0.42
CSI	TP/(TP+FP+FN)	0.29	0.59	0.48

These comparisons highlight the distinct operational profiles of each database: HANZE functions as a conservative benchmark with high reliability but severe blind spots; Groundsource provides broad event coverage but suffers from high contamination by unverified non-events; and FloodEcho substantially expands the volume of reliable flood detection while effectively filtering out raw internet noise. Overall, FloodEcho demonstrates superior detection efficacy, outperforming both HANZE and Groundsource through its practical advantage in balancing comprehensive event discovery with a strict false-positive control.

4.3 Evaluation of Link verification pathway

To assess the filter capability of the link verification pathway, all online resources retrieved by FloodEcho for the 100 benchmark events (amounting to 3,094 candidate URL), were cross-checked against manual assessments.

Figure 4.3 tracks the process of these 3,094 candidate links through the link verification pathway described in Section 2.4, comparing the automated FloodEcho workflow against the manual review process. The candidate URLs were progressively evaluated across rule-based pre-filtering (Stage 1), page-level content verification (Stage 2), and LLM-driven semantic validation (Stage 3). In the manual validation, Stages 1 and 2 were combined into a unified manual inspection step, as both stages share the common objective of confirming whether the event timestamp, geographical location, and topical relevance match the target incident.

As demonstrated in Figure 4.3(a), the initial two stages eliminated the overwhelming majority of candidate noise. FloodEcho retained 965 links (31.2%) after Stage 1 and 565 links (18.3%) after Stage 2, whereas the stricter manual review retained 293 links (9.5%) across the corresponding checks. In the final stage (LLM semantic validation), FloodEcho narrowed the candidate pool down to 184 surviving links (5.9%), which matches the human-verified benchmark of 187 confirmed links (6.0%).

Figure 4.3(b) details the distribution of rejection causes identified during the manual evaluate of the 2,907 rejected links. Temporal mismatch constituted the most substantial reason, with non-matching event timing accounting for 1,012 cases (34.8%) and analyses of historical floods contributing an additional 604 cases (20.8%). Geographical mismatch represented another major factor (444 cases, 15.3%), followed by unrelated non-flood news reporting (305 cases, 10.5%), prospective weather forecasts (207 cases, 7.1%), pages lacking detailed impact descriptions (188 cases, 6.4%), and routine meteorological bulletins (147 cases, 5.1%).

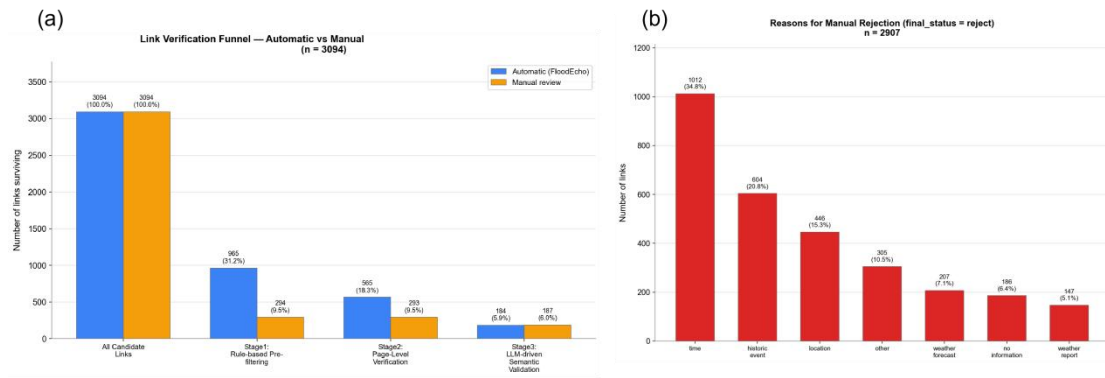


Figure 4.3: Funnel comparison of candidate link verification pathway and root cause analysis of manual rejections: (a) Funnel comparing the link verification pathway designed in FloodEcho (in blue) and manual review (in yellow); (b) Distribution of reasons for manual link rejections (n = 2907).

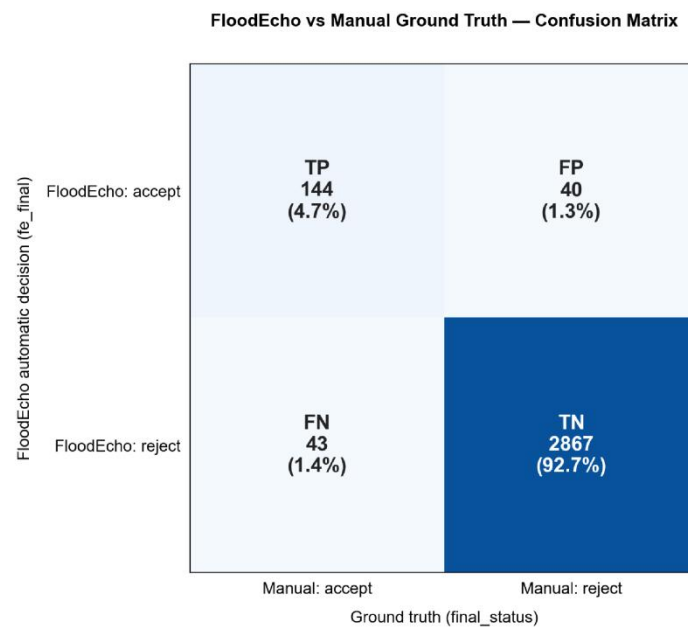


Figure 4.4: Confusion matrices of the evaluation of link verification pathway designed in FloodEcho.

The overall classification performance of the link verification pathway is summarized in the contingency matrix shown in Figure 4.4. The vast majority of candidate URLs fell into the True Negative category (2,867 links, 92.7%), confirming that FloodEcho successfully eliminated most of the link original from the searching engine that were unrelated to the target events. True Positives comprised the second-largest share (144 links, 4.7%), denoting URLs that reported the target events and were successfully retained by the automated workflow. However, there are still some False Positive (40 links, 1.3%) and False Negative (43 links, 1.4%) remained as noise after the link verification. The result confirm that FloodEcho reliably isolates irrelevant internet sources, and catches most of the useful articles.

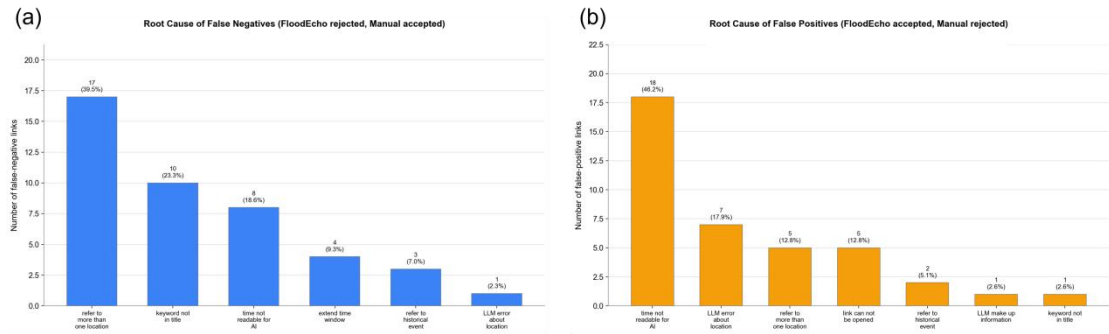


Figure 4.5: Root cause analysis of classification noises between FloodEcho and manual validation: (a) False negatives (FloodEcho rejected, manual accepted); (b) False positives (FloodEcho accepted, manual rejected).

To understand the drivers of these classification errors, a root-cause analysis was conducted on both false negatives and false positives (Figure 4.5). Among the 43 false negatives (Figure 4.5a), the dominant cause was reports from articles referencing multiple geographic locations (17 cases, 39.5%). This was followed by the absence of flood keywords in the title (10 cases, 23.3%), ates or times unreadable to the algorithm (8 cases, 18.6%), extended time windows (4 cases, 9.3%), references to historical flood analogues (3 cases, 7.0%), and localized extraction errors by the LLM (1 case, 2.3%).

In contrast, among the 39 false positives shown in Figure 4.5(b), unreadable temporal information was the dominant error factor (18 cases, 46.2%), with other errors from incorrect location extraction by the LLM (7 cases, 17.9%), multi-location uncertainty (5 cases, 12.8%), broken or inaccessible URLs (5 cases, 12.8%), historical event references (2 cases, 5.1%), and single instances of LLM hallucination and missing title keywords (1 case each, 2.6%).

This distribution explains the primary causes behind both types of errors. False negatives mainly occur when articles report on multiple regions in the same link, causing the system to reject them under strict filtering rules. In contrast, false positives mostly arise from relative date references (such as "yesterday" or "last week") in the article text, which can mislead the model about the actual event time. Therefore, future work should focus on separating multiple locations within a single report and improving the extraction of relative dates.

In Stage 3 of the verification pathway, the LLM assigns a composite relevance score across four dimensions: temporal evidence, source reliability, geographic matching, and content usefulness. To validate these automated ratings, the relevance score have also been calculated with the same evaluation standard during the manual evaluation within the 567 candidate links.

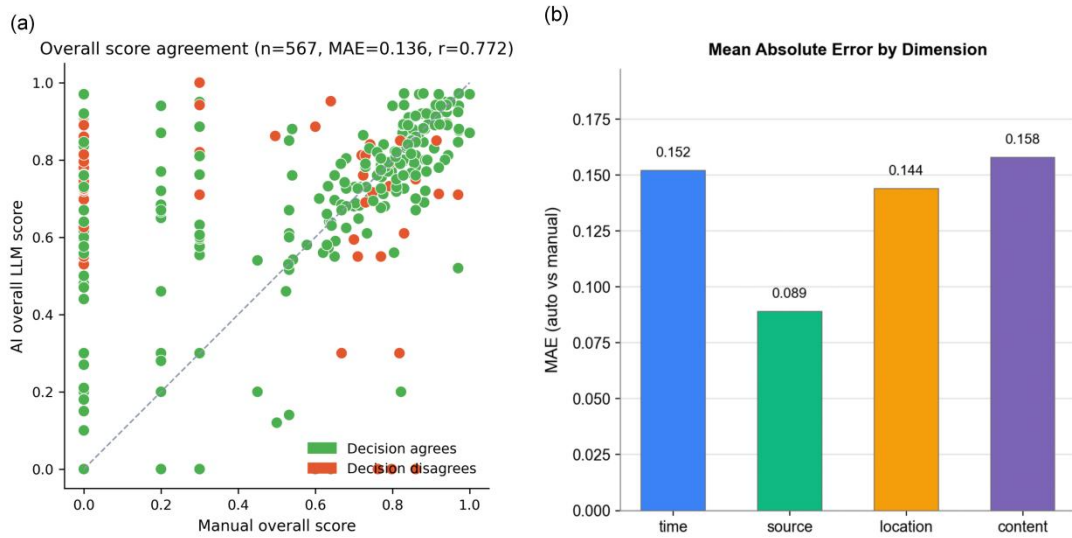


Figure 4.6: Agreement evaluation between LLM-assigned relevance scores and manual assessments in Stage 3: (a) Scatter plot of overall relevance score concordance; (b) Mean Absolute Error (MAE) across the four evaluation dimensions.

As shown in the scatter plot in Figure 4.6(a), the overall relevance scores cluster tightly along the 1:1 reference line, demonstrating high linear correlation and a low overall error. The majority of evaluated instances align on final classification decisions (green markers), with orange markers shows the final decision disagree with two approach.

Figure 4.6(b) further explain the scoring assigned across the four sub-dimensions: source authoritativeness yielded the highest agreement with an MAE of only 0.089, geographic matching and temporal evidence had moderate errors of 0.144 and 0.152, while content usefulness exhibited the largest discrepancy with an MAE of 0.158. These results show that the LLM can reliably replace human reviewers in this scoring task. The source dimension has the lowest error because news website domains and official channels are clear and easy for the model to recognize. In contrast, the time and location dimensions show slightly higher errors, mainly because articles often use relative dates or complex place names that are harder to pin down. The content dimension has the highest error because judging whether an article's content is truly useful involves personal subjective judgment.

In summary, the link verification workflow effectively filters out the majority of irrelevant search results while retaining reliable reports of real flood impacts. This screening process provides a solid foundation for the subsequent event severity evaluation.

4.4 Evaluation of Event Impact Severity Assigned by FloodEcho

As described in the methodology, flood impacts extracted from verified news reports were categorized into four classes: Transport, Human Safety, Economy, and Emergency Response. In addition, an Overall impact and a Global impact score were assigned to each event. To assess the accuracy of these severity ratings, the 55 real

flood events identified in the manual validation sample were independently evaluated and scored on an integer scale from 0 to 4. Figure 4.7 compares the automated scores from FloodEcho with the manual ratings. Across all six categories, the data points cluster closely along the 1:1 reference line, indicating consistent agreement between FloodEcho and manual scoring.

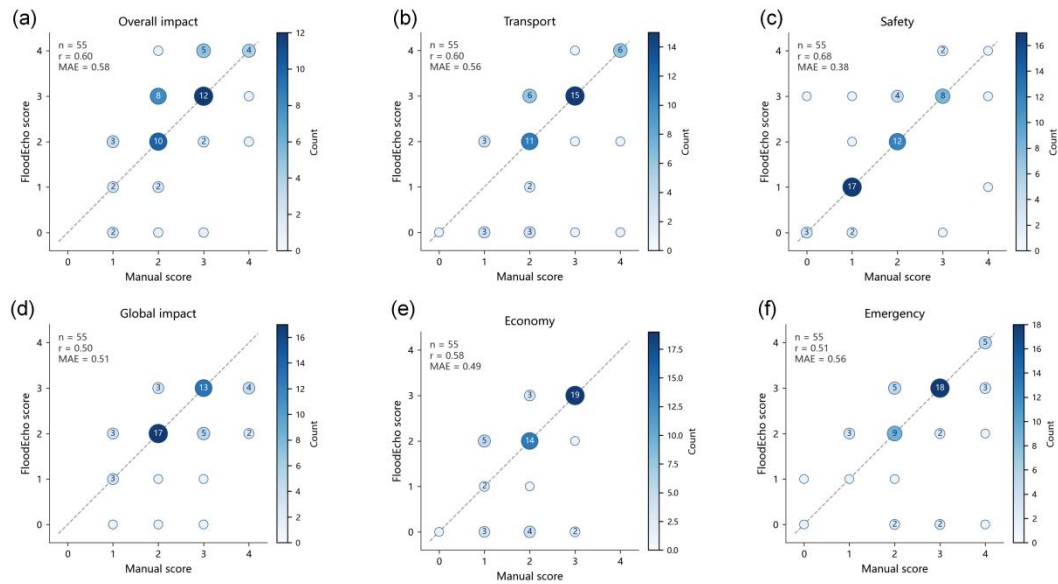


Figure 4.7: Comparison of flood impact severity scores assigned by FloodEcho against manual evaluation across 55 verified flood events: (a) Overall impact, (b) Transport, (c) Safety, (d) Global impact, (e) Economy, and (f) Emergency. The size and color depth of each circle indicate the number of events with the same score combination.

Among the four classes, Safety showed the highest agreement, with the highest correlation coefficient (0.68) and the lowest error (0.38), with most cases concentrated at severity scores of 1 and 2. Transport and Economy also showed good agreement, with events mainly distributed across moderate-to-high severity levels. Emergency response and Global impact showed slightly lower correlations, while the composite Overall impact score maintained steady consistency.

These differences reflect the nature of the information reported in the news. The model performs best on Safety because casualties and injuries are typically reported as concrete numbers. In contrast, emergency operations and overall damage are often described using qualitative descriptions rather than specific figures, leading to slightly larger discrepancies between automated and manual ratings. Nevertheless, the mean absolute error remained below 0.60 across all categories, and correlation coefficients ranged between 0.50 and 0.68. These results demonstrate that FloodEcho can reliably quantify flood impact severity across different sectors automatically.

5 Discussions and Conclusions

This paper presents the operational workflow designed for FloodEcho, an automated database that systematically collects impactful flood events from open-

source online media to evaluate flash flood impact forecasts. The system comprises three modules: a hydrometeorological trigger to identify potential events, a targeted search module to retrieve online media evidence, and an LLM-based framework to extract detailed impact information and impact severity. Through this approach, FloodEcho establishes a structured framework for transforming unstructured open-source information into structured flood impact database.

The system was demonstrated over a one-year period across Europe. Based on an expert manual evaluation of the candidate links retrieved by the system, results indicate that the link verification pathway effectively filters out more than 80% of links irrelevant to the target events. Despite this high filtering efficiency, classification errors persist in the form of false positives (FP) and false negatives (FN).

False positives mostly arise from the misinterpretation of web link timestamps. The system sometimes fails to extract event publication dates, a limitation particularly pronounced when processing Facebook posts. However, because Facebook represents one of the most active sources of reporting where local flash flood incidents are documented in user posts, excluding this platform would substantially degrade overall event recall. Consequently, Facebook is retained as a critical data source while acknowledging and accounting for this temporal uncertainty.

Conversely, false negatives are mainly driven by spatial extraction uncertainties. First, online articles often mention broad administrative levels (e.g., regional or provincial names) without specifying the affected municipality, failing to match the official name of NUTS-3 level. Second, when an article reports impacts across multiple distinct locations, the strict criteria in the prompt frequently trigger a conservative rejection, resulting in missed detections of valid flash flood impacts.

To overcome location mismatches, the spatial evaluation could incorporate Local Administrative Units (LAU) and cross-referencing meteorological warnings or hydrological basin boundaries to infer the most probable impact zone. Furthermore, to prevent valid multi-region reports from being rejected, the extraction prompt could be restructured into a multi-location parsing routine. By interpreting news text into location-specific passages before evaluation, impacts across distinct locations can be recorded and scored independently without triggering a global rejection.

To further improve link verification accuracy, two other practical methods can be considered. First, a dedicated judgment model such as JEV (Deußner et al., 2026) can be applied to evaluate whether a candidate link is related to the target flood event. Instead of relying on keyword matches or strict rules, it evaluates the overall context of the page to determine whether the reported damage actually corresponds to the target event, helping to reduce both false negatives and false positives.

Second, an AI agent framework could be introduced to automate and refine the verification workflow. Unlike the link verification pathway with fixed rules and prompts, an agent can review its own output, identify limitations, and dynamically adjust its evaluation path. For instance, if an error or an uncertain date is detected, it can modify the extraction prompt or adjust the parsing script to re-evaluate the source. This self-correction loop allows the system to adapt to complex web sources without requiring

manual rule engineering, directly improving verification accuracy and reducing false alarms.

A comparative evaluation was conducted on a representative sample of 100 events selected from the entire database, each manually reviewed and validated through expert evaluation to establish ground truth. When benchmarked against the HANZE and Groundsource databases, FloodEcho achieves a good balance between false alarms and missed detections, securing an overall Critical Success Index (CSI) of 0.59. Although the recall of FloodEcho ($POD = 0.71$) is slightly lower than that of Groundsource (0.75), FloodEcho significantly reduces spurious detections, exhibiting a False Discovery Rate (FDR) of 0.22 compared to 0.42 for Groundsource.

In the original evaluation of Groundsource, Mayo et al. (2026) noted that unconstrained web extraction remains vulnerable to considerable uncertainties in event identification and spatiotemporal boundary delineation, which directly increases false positives. Similarly, Andreozzi et al. (2026) investigated the reliability of LLM-based disaster event extraction from online text and highlighted Groundsource as an example, warning that unverified text extraction frequently produces unreliable disaster records.

The primary reason for FloodEcho's lower false discovery rate is the guidance provided by the hydrometeorological trigger. Rather than widely scanning online sources, FloodEcho searches for flood impacts only when a physical event trigger is detected. Constraining the process to a targeted spatiotemporal window effectively filters out unrelated news, fabricated reports, and historical reports before media extraction begins. Furthermore, this physically guided retrieval substantially reduces the risk of LLM hallucinations. In this way, the event trigger filters out irrelevant news early, allowing the model to focus only on real flood events.

A promising strategy to enhance event recall is to diversify the initial triggers used to detect potential events. To capture localized flash floods more comprehensively, multiple hydrometeorological signals can be integrated into the system. For instance, these could include real-time streamflow observations from gauge networks, operational early warning notifications (such as those issued by the European Flood Awareness System, EFAS), and distributed hydrological model outputs. Incorporating these varied hydrological feeds would allow FloodEcho to capture a wider range of potential flood occurrences.

At the same time, relying on specific hydrometeorological triggers inherently defines the operational scope of the system. Currently, FloodEcho is primarily configured to capture flash floods. It misses other flood types that follow different physical mechanisms, such as snowmelt floods, coastal storm, and localized pluvial or urban drainage events. This limitation can be addressed by introducing dedicated triggers tailored to these hazard regimes. For example, incorporating positive temperature anomalies alongside snow water equivalent metrics would allow the detection of snowmelt floods; integrating tide gauge records and coastal surge forecasts could capture marine inundation; and coupling high-resolution radar rain rates with urban drainage thresholds could initiate searches for localized pluvial events.

Expanding triggers in this manner will allow the system to capture more types of flood hazards while maintaining its low false discovery rate.

Beyond physical hydrometeorological drivers, civil protection and emergency management datasets offer valuable direct signals of flood impacts. Incorporating operational response records could substantially broaden the event discovery window. Relevant sources include the volume of calls logged by emergency dispatch centers (e.g., 112 calls), formal activation logs of municipal civil protection protocols, and post-event incident summaries. Coupling multi-source hydrometeorological observations with operational emergency records would enable FloodEcho to trigger targeted web searches for small-scale, convective events that cause significant localized damage.

Beyond serving as a disaster repository, FloodEcho provides practical value across several operational domains. For civil protection agencies and emergency managers, the automated extraction of localized impact details and severity ratings delivers post-disaster situational reports with minimal delay. Emergency authorities can utilize these records to rapidly map affected infrastructure, evaluate regional vulnerability, and guide long-term risk mitigation and resource allocation across flood-prone municipalities.

Furthermore, the resulting database provides an objective ground-truth benchmark for evaluating and calibrating impact-based flood forecasts. By documenting ground-level consequences across transport, infrastructure, and human safety with precise spatial attribution, FloodEcho enables true impact-based verification, helping forecasters assess whether issued warnings corresponded to actual real-world impact. By accumulating standardized impact records linked to physical observations, FloodEcho can help understand the relationships between physical hazard levels and resulting social impact.

In summary, FloodEcho provides an automated and reliable methodology for extracting verified flood impact data from unstructured online sources. By coupling physical hydrometeorological triggers with a multi-stage link verification pathway and LLM-based severity scoring, the workflow effectively limits false alarms while preserving real impactful events. Beyond recording historical floods and supporting emergency planning, this database provides a reliable benchmark for evaluating flood impact forecasts. As future work, the flash flood impacts recorded by FloodEcho will be applied to evaluate operational flash flood impact forecasts generated within the INLINE project framework.