

When LLMs Become Scientific Instruments: Groundsource Flash Flood Research Report

Published: October 6, 2026 at 11:51 AM Author: Richard Djarbeng 20 min read

A deep dive into Google's 2.6-million-event flood dataset: what the data shows, what claims hold up, and why the methodology matters more than the dataset itself.

Resources: [Enriched Dataset](#) | [Full Interactive Article](#) | [Original on Zenodo](#)

What is Groundsource?

In February 2026, Google Research released **Groundsource**, an open-access global dataset of 2.6 million historical flood events extracted from news articles using Gemini models. Google published the dataset on [Zenodo](#) alongside a preprint on [EarthArXiv](#) (Mayo, Zlydenko, Bootbool, Nearing, Cohen, et al.).

Google used Gemini to scan **5 million news articles across 80+ languages**, generating **2.6 million geo-tagged flood events** spanning 150+ countries. This archive serves as the ground-truth training data behind Google's operational flash flood forecasting system.

The largest existing global disaster databases, such as GDACS, held roughly 10,000 flood events, mostly major disasters. Groundsource records 2.6 million localized events, showing that language models can turn unstructured news archives into structured scientific datasets.

In [our initial March breakdown on Google's flash flood system](#), we covered the broader operational announcement—how Google deployed AI to forecast urban flash floods 24 hours in advance. This follow-up report takes a different angle: rather than evaluating the forecasting system from press announcements, we independently downloaded the raw 667 MB Groundsource Parquet archive to stress-test the underlying data, evaluate geographic representation, and verify whether the published claims hold up under empirical inspection.

What the data actually shows

The dataset is distributed as a single 667 MB Parquet file containing **2,646,302 flood events**. Each record contains a UUID, polygon boundary (stored as WKB geometry), area in km², start date, and end date.

Key numbers

Metric	Value
Total events	2,646,302
Null values	0
Duplicates	0
Date range	2000-01-01 to 2026-02-03
Median area	2.05 km ²
Mean area	142.0 km ²
Peak year	2024 (402,012 events)

What is missing

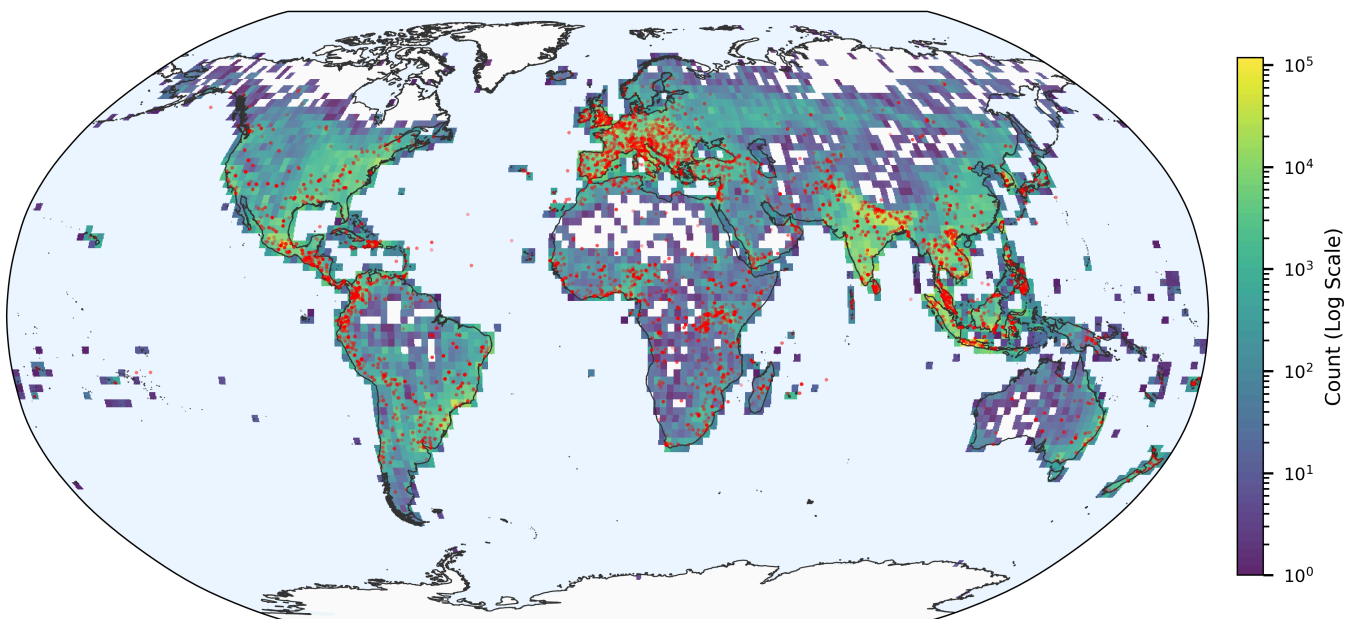
There is no country column, no source article URL, no language tag, no confidence score, and no casualty or severity rating. The dataset includes only geometry, start date, end date, and area.

Geographic distribution

To understand where these events actually cluster across the globe, we decoded the 2.6 million WKB geometries into latitude and longitude centroids. The table below breaks down event counts by world region:

Region	Events	Share
Europe	590,603	22.3%
Southeast Asia	488,885	18.5%
South Asia	484,418	18.3%
North America	412,254	15.6%
South America	248,652	9.4%
East Asia	179,846	6.8%
Africa	111,053	4.2%
Other	131,591	4.9%

Mapping these coordinates globally highlights major regional imbalances. In **Figure 2**, the blue-to-yellow heatmap shows the density of Groundsource events aggregated across global grid cells, while red markers indicate reference disaster events recorded by GDACS:

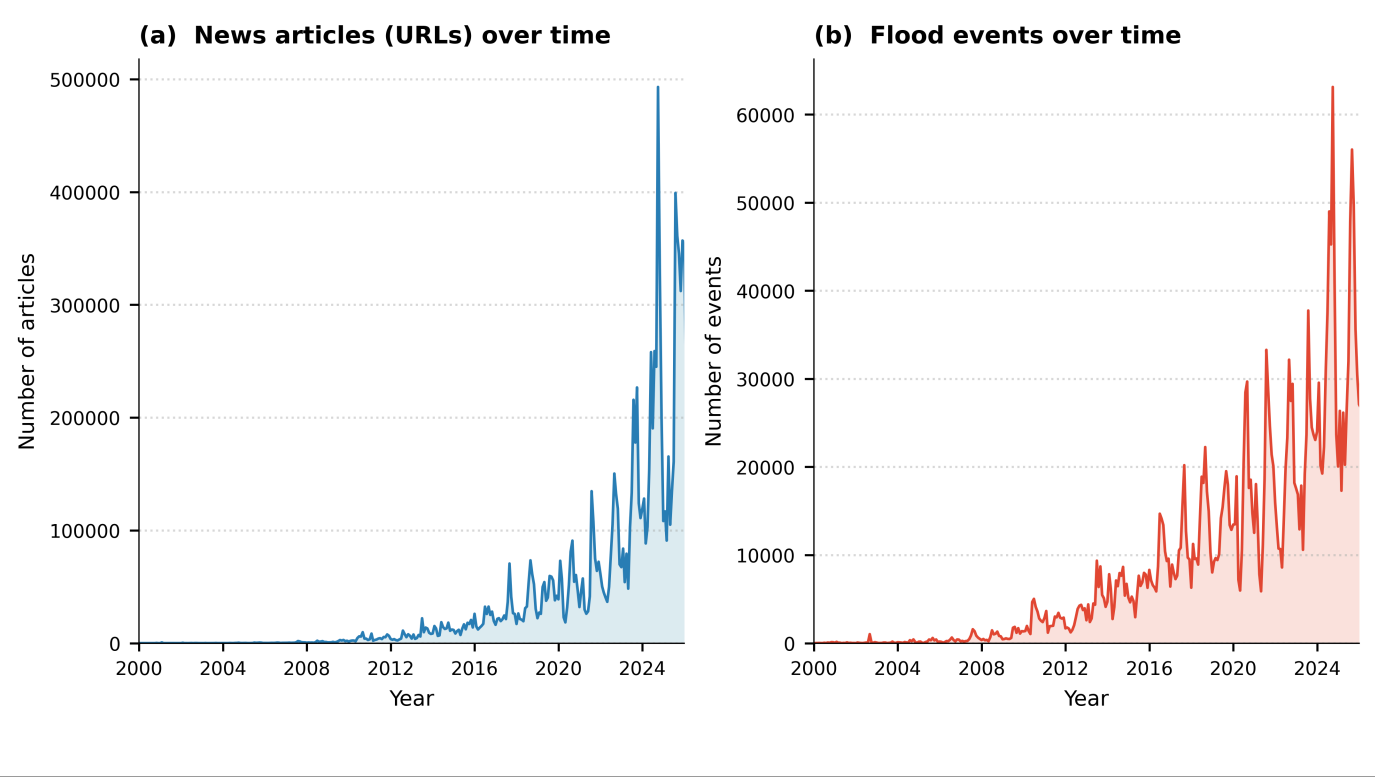


Temporal growth

A breakdown of event volume across the 26-year timeline reveals a heavy skew toward recent years:

Period	Events	Share
2000-2009	40,581	1.5%
2010-2019	876,630	33.1%
2020-2026	1,729,091	65.3%

Over 65% of all records come from the last six years. To see whether this reflects real-world climate trends or changes in digital data collection, **Figure 1** compares raw news article intake against extracted flood events over time. As shown below, event growth closely tracks Google's ingested web crawl volume rather than an actual 40-fold spike in worldwide flooding:



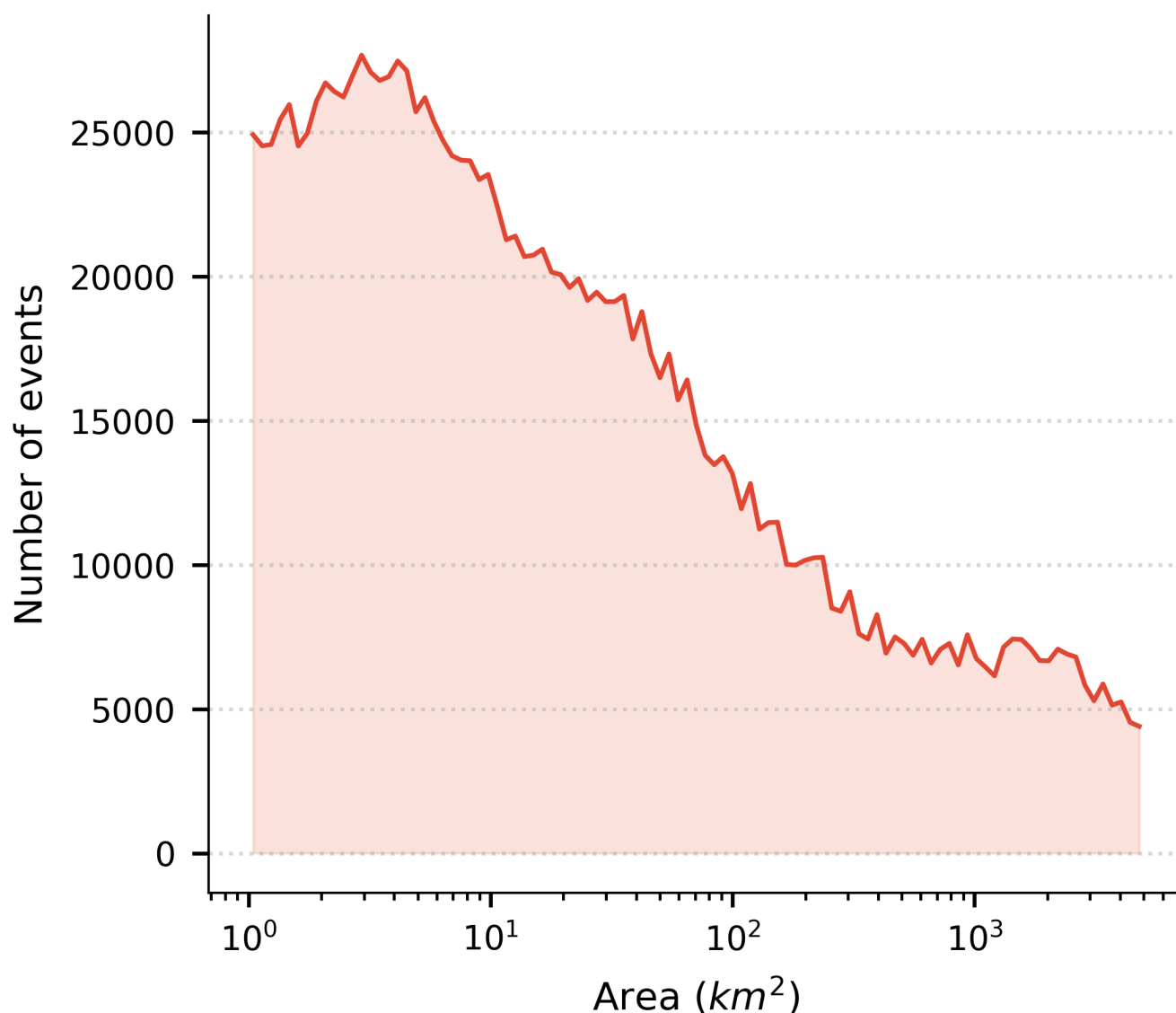
Data topology: entity polygons vs. regional disasters

When looking at disaster data, most people expect each record to represent an entire storm system—like "Hurricane Helene" or "2024 European floods." In Groundsource, however, the **median event area is just 2.05 km²** (about the size of Monaco or New York's Central Park), while the mean is 142.0 km².

This sharp contrast happens because Groundsource is **entity-based rather than storm-based**. The pipeline does not aggregate rainstorms into broad regional events; instead, it extracts specific localized entities named in news stories—individual neighborhood

boundaries, suburban street networks, or local municipal districts.

To illustrate how concentrated these footprints are, **Figure 3** shows the frequency distribution of event sizes on a logarithmic scale. The heavy skew toward the lower end reveals that most extracted records are hyper-local—meaning instead of marking an entire province or watershed as underwater, each record pinpoints a tiny, self-contained pocket like an underpass, a single residential block, or a small town square:



The authors explain this design explicitly:

"A single, large-scale real-world flood event may be represented by multiple entries within the Groundsource dataset. This occurs when an extensive flood inundates multiple distinct geographic entities (e.g., specific neighborhoods, towns, and districts), all of which are independently annotated by the LLM extraction process."

Key implications of this structure:

- **82% of events cover less than 50 km²** (median: 2.05 km², mean: 142 km²).
- When a severe cyclone strikes a region and floods 40 towns, 80 streets, and 12 districts, Groundsource records over 100 distinct polygon rows. Each is tied to a specific administrative boundary or a buffered point (0.001°0.001°).
- **Impact on hydrological modeling:** In gridded hydrological training (such as 0.05°0.05° ERA5-Land or 0.1°0.1° IMERG), tight footprints prevent positive flood labels from bleeding across dry terrain. But researchers studying macroeconomic disaster loss cannot treat individual rows as separate storms.

Ingestion and extraction pipeline

The Parquet file does not include source URLs, but the EarthArXiv preprint describes the five-stage pipeline:

```
[9.5M Candidate URLs]
  ↓ Google Web Crawler + WebRef (Topicality Score ≥ 0.6)
[7.5M Accessible Articles]
  ↓ Read Aloud User-Agent (80 Languages) + Cloud Translation to English
[5.0M Verified Flood Reports]
  ↓ Gemini 3 Flash Prompt (Gate Filtering: Precision 75%, Recall 90%)
[2.65M Spatiotemporal Events]
  ↓ Google Maps Geocoding API + Spatial Aggregation + Geometric Thresholds
```

1. **Entity filtering (9.5M URLs):** Google harvested news articles published since 2000 mentioning floods. The WebRef entity system assigned a topicality score between 0 and 1 for "flood". Articles scoring at least 0.6 were kept (9.5 million URLs).
2. **Text cleaning and language coverage (7.5M articles):** The Google Read Aloud user agent extracted article text and publication dates across 80 supported languages. Sites blocking bots or using unsupported languages were dropped, leaving 7.5 million accessible articles.
3. **Translation and entity candidates:** Non-English articles were translated into English with the Google Cloud Translation API. WebRef extracted named places from both original and translated text to build candidate Machine Identifiers (MIDs).
4. **Gemini 3 Flash extraction (5.0M articles):** Gemini 3 Flash evaluated each article using a structured prompt to confirm ongoing or past floods while discarding forecasts, insurance notices, or regional summaries. Gemini identified roughly 5.0 million articles as actual flood reports.

5. **Geocoding and spatial filtering (2.6M final events):** Locations were resolved to Google Maps polygons (or 0.001°0.001° buffered points). Consecutive daily reports for identical locations were merged. Events exceeding 5,000 km²5,000 km², with diameters over 500 km500 km, continuous durations over 7 days, or dates before 2000 were filtered out, leaving **2,646,302 finalized records**.
-

Claim verification

To evaluate whether the headline figures match reality, we audited the key claims made in the EarthArXiv paper against the downloaded Zenodo Parquet dataset and external disaster databases. Here is what held up, what required nuance, and where the discrepancies lie:

Confirmed: 2.6 million geo-tagged events

The dataset contains 2,646,302 events with valid WKB polygon geometries and dates. There are no null values and no duplicate records.

Confirmed: 5 million articles across 80 languages

The pipeline processed 7.5 million articles across 80 languages supported by Read Aloud. Gemini 3 Flash classified 5.0 million of them as real flood events, which geocoding and geometric filtering condensed into 2.65 million spatiotemporal polygons.

Confirmed with context: GDACS comparison

The 260x scale increase is real, but reflects a different unit of analysis. GDACS tracks roughly 10,000 major humanitarian disasters requiring international response (events affecting 100 or more people or triggering external aid). Groundsource captures localized street, town, and municipal flooding.

Confirmed: Africa coverage gap

Africa represents 4.2% of events in the dataset despite having roughly 17% of the world's population. This underrepresentation stems from uneven digital news infrastructure and search engine indexing.

Accuracy and failure modes

Google ran a manual audit of 400 randomly selected entries evaluated by human raters against the source news text:

- **Strict precision (60% $\pm$ 5% at 95% CI):** Matched both the location polygon and start and end dates in the article ("Accurate").
- **Practical usability (82% combined):** An additional 22% had minor discrepancies ("Approximate" or "Partial"), such as selecting a surrounding municipal district instead of a specific village, or a one-day offset from phrases like "over the weekend".
- **Hard error rate (18% "Wrong"):** Unusable due to metaphorical uses ("a flood of complaints"), ambiguous place names, or hallucinated dates.

Main failure modes

1. **Ambiguous place names:** Common names without regional qualifiers were misrouted during geocoding (for example, "Kherbari", which exists in multiple Indian states).
2. **Updated publication timestamps:** When an outlet updated an article days after an event, the system calculated relative dates ("last Tuesday") from the update time, shifting the event date by a week.
3. **Temporal guessing:** Vague phrasing like "last September" sometimes led Gemini to guess a specific date (such as September 1) instead of dropping the record.

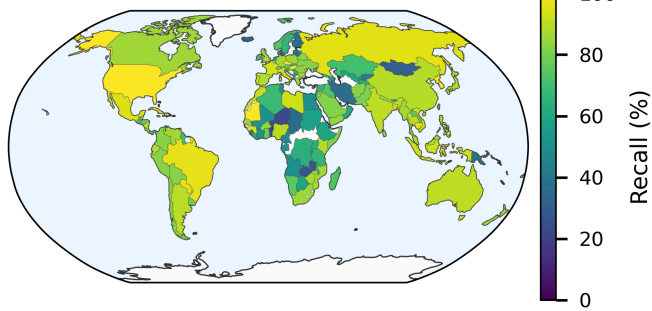
"The actual errors in the dataset are likely not independent and identically distributed (i.i.d.). For instance, the geocoding system may exhibit variable accuracy with location names across different languages, potentially concentrating spatial errors in specific regions or countries." — Mayo et al.

External benchmark validation: GDACS and DFO audits

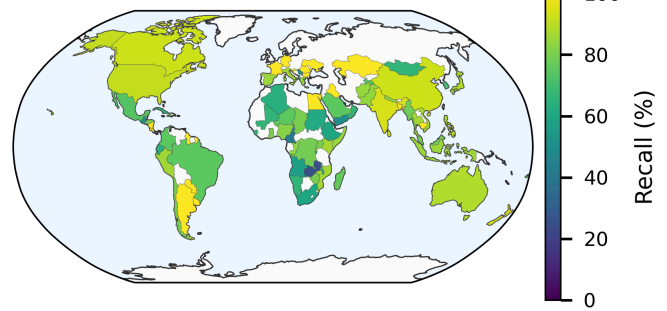
To measure recall against external records, the authors evaluated spatiotemporal overlap against **6,537 GDACS events (2017 to 2026)** and **3,875 Dartmouth Flood Observatory (DFO) satellite-derived events (2000 to 2023)**.

Figure 4 plots country-by-country recall against these two independent catalogs. Panels (a) and (b) show the percentage of recorded disasters that Groundsource successfully captured in GDACS and DFO, while panels (c) and (d) provide the raw count of baseline disasters recorded per country:

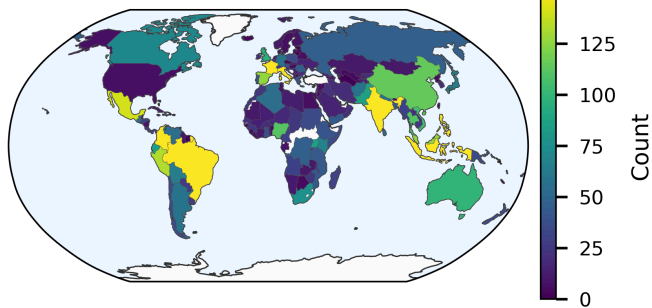
(a) GDACS recall per country



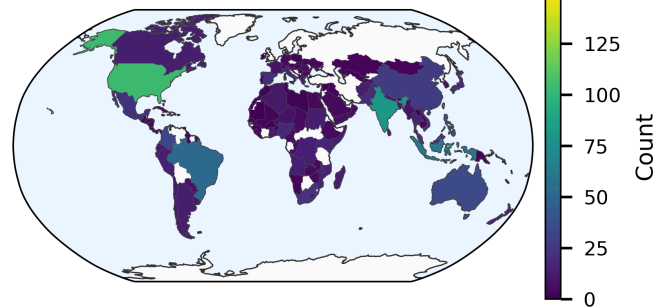
(b) DFO recall per country



(c) GDACS event counts

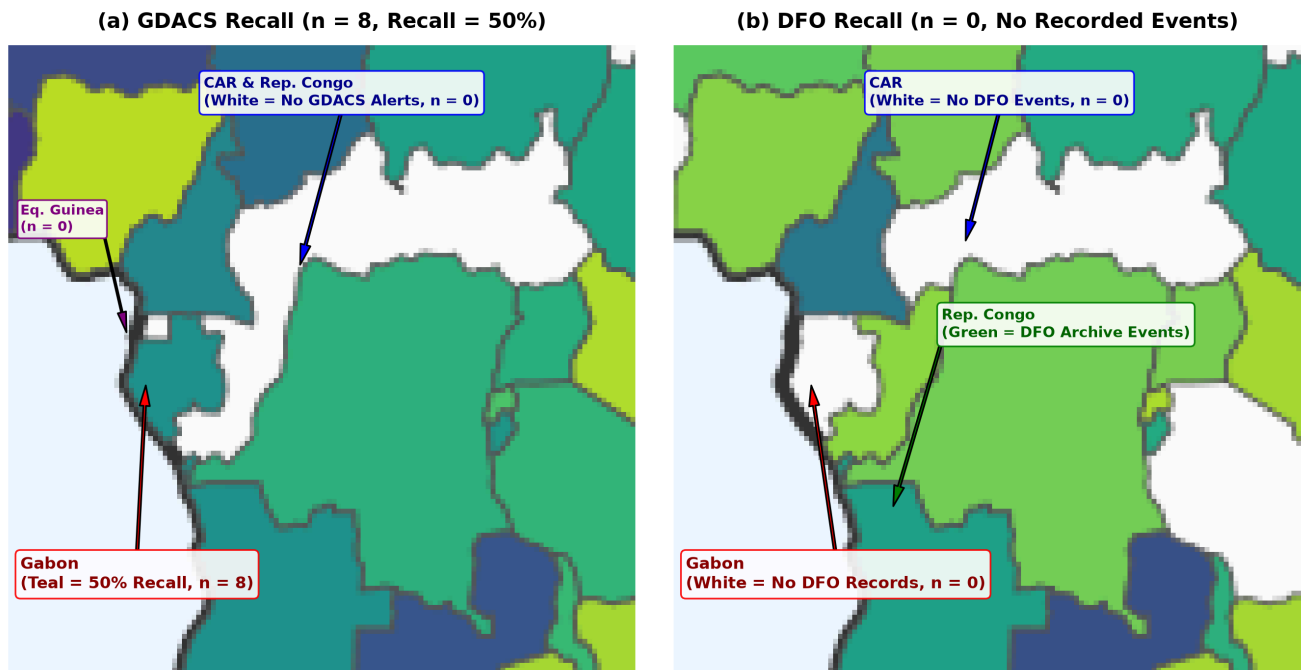


(d) DFO events counts



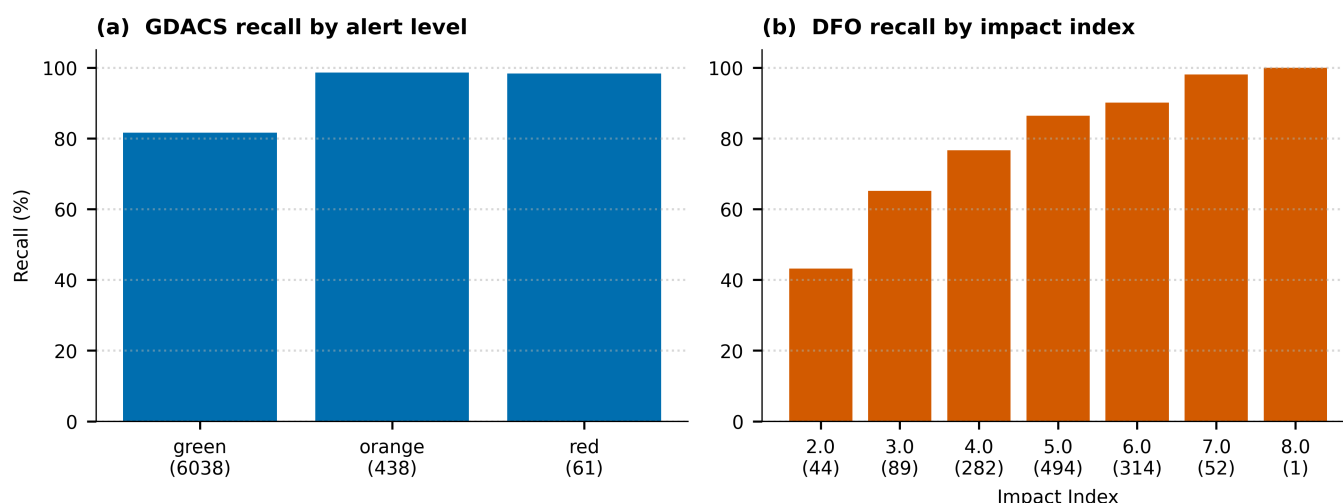
- **Annual recall vs. GDACS:** From 2020 to 2025, Groundsource captured **81% to 86% of all GDACS events globally** (reaching 90.1% in 2017, 94.2% in 2018, and 93.6% in 2019).
- **Annual recall vs. DFO:** Recall grew from 13.7% in 2000 to **93.6% in 2019**, following the growth of digital news publishing.
- **Regional disparities:** Recall exceeds 96% in the United States and 79% to 89% in the Philippines and Malaysia. It drops in areas with lower digital media presence or unsupported indigenous languages (39% in Papua New Guinea, 50% in Gabon).

Reading note on Figure 4 (white vs. shaded countries): In these maps, unshaded (white) countries represent **zero reference events in the baseline archive** ($n = 0$), meaning recall is undefined, rather than 0% model extraction performance. For example, Gabon appears unshaded in panel (b) because the DFO archive contains zero recorded flood events for Gabon ($n = 0$). In panel (a) GDACS, where $n = 8$ reference events exist, Gabon is shaded teal at **50% recall** (capturing 4 of 8 events). Conversely, neighboring inland countries like the Republic of the Congo and Central African Republic appear white in GDACS due to an absence of baseline disaster records.



Scaling with event severity

When floods cause widespread damage, news coverage multiplies, making extraction far more reliable. Figure 5 plots Groundsource's retrieval rates broken down by disaster severity across both reference archives:



As the bar charts show, model recall climbs steadily as an event's humanitarian or physical footprint grows:

- **GDACS alert categories (panel a):** Recall sits at **82%** for routine, locally managed floods ($n = 6,038$ green alerts), but rises to **99%** for major humanitarian crises ($n = 438$ orange alerts and $n = 61$ red alerts).
- **DFO Flood Impact Index (panel b):** Recall ranges from 43% to 65% for minor local floods (Impact Index 2 to 3), but exceeds **90%** for widespread or destructive events (Impact Index > 6).

In short, while localized street flooding can slip through without coverage, major flood disasters almost never escape news extraction.

Hydrological context: why flash floods needed Groundsource

The gap Groundsource addresses becomes clear when comparing **riverine flooding** with **flash flooding**:

1. **Riverine floods (tracked by stream gauges):** Hydrological machine learning models, including Google's global model ([Nearing et al., Nature 2024](#)) and state-space architectures like [RiverMamba \(Shams Eddin et al., 2025\)](#), predict river flow through established river basins (HydroATLAS, Caravan dataset) where physical gauges provide steady measurements.
2. **Flash and pluvial floods (the data void):** Flash floods are triggered by fast, intense rainfall over small ungauged drainage basins, city streets, and dry ravines. Physical gauges rarely exist in these locations.

3. **The human reporting network:** Cloud cover blocks optical satellites during rainstorms. Groundsource uses local news reports as a proxy observation network for flash flood events.
-

Forecasting with historical records

If the dataset is a static archive of past news, how does it warn about a flood happening tomorrow?

Groundsource provides training data, not live inputs. The forecasting model paired these 2.6 million historical events with local weather conditions (rainfall, soil moisture, runoff) at the time of each event to learn physical response patterns. For daily forecasts, the trained model evaluates live weather feeds (ECMWF, NASA, NOAA) to calculate flood probabilities:

TRAINING: Groundsource labels + Historical weather → Train model

OPERATIONAL: Live weather feeds → Frozen model → "Flash flood likely here tomorrow"

The model does not need daily news ingestion to generate forecasts, just as an image classifier does not need daily retraining to identify new photos.

The Africa coverage gap and potential fixes

Africa accounts for **4.2% of Groundsource events** despite having **roughly 17% of the global population**. Several factors contribute to this disparity:

1. **Fewer online news outlets** indexed by global aggregators, alongside heavy use of local radio broadcasts that web crawlers do not capture.
2. **Language limits:** Africa has more than 2,000 languages, but the ingestion pipeline relied on the 80 languages supported by Google's Read Aloud agent.
3. **Urban bias:** Floods in remote rural areas rarely generate written news stories.

Ways to address the gap

Current work in remote sensing and machine learning offers practical ways to fill these gaps:

1. **Synthetic Aperture Radar (SAR):** Optical satellites cannot see through storm clouds, but SAR penetrates cloud cover day and night. Multi-temporal SAR datasets like [Kuro Siwo \(Alberti et al., 2024\)](#) supply 33 billion m² of flood inundation masks across 43 disasters, offering ground truth independent of news reporting.
 2. **Synthetic data generation:** As shown by [SAGDA \(2025\)](#) for agricultural data in Africa, physics-informed synthetic generators can simulate extreme hydrological events in data-sparse regions.
 3. **Cross-regional transfer learning:** Models like [RiverMamba \(2025\)](#) demonstrate that spatial representations pretrained on global reanalysis can transfer predictive skill to ungauged basins in the Global South.
 4. **Multimodal surveillance methods:** In epidemiological surveillance, [Epidemic IE \(2024\)](#) and the WHO Disease Outbreak News Knowledge Graph ([eKG, Scientific Data 2025](#)) show that combining structured ontologies with multi-LLM ensembles extracts outbreak events (F_1 scores up to 0.954) from sparse reporting. [DengueNet \(2024\)](#) similarly pairs satellite imagery with sparse health reports in developing regions.
 5. **Support for regional languages:** Extending entity extraction to African languages (such as Swahili, Hausa, Amharic, and Yoruba) and transcribing local radio broadcasts using speech-to-text models.
-

Broader applications of the methodology

The core lesson from Groundsource reaches beyond hydrology: **language models can transform unstructured text into structured scientific ground truth at global scale.**

Other potential domains

Domain	Feasibility	Why
Disease outbreaks	Very high	Frequent public reporting; demonstrated by ProMED and WHO eKG ($F_1 \approx 0.954$)
Conflict and displacement	High	ACLED validation; extensive real-time journalistic coverage
Pollution events	Medium	Sudden chemical spills work well; ongoing air and water metrics require sensor measurements (such as AirPhyNet)
Wildfires	Medium	Satellite thermal data is already strong; text adds context on evacuations and ignition causes
Mining hazards	Medium	Dam failures are sudden and reported; underground leaks are chronic and rarely covered
Drought and crop failure	Lower	Gradual onset over months; lacks distinct start and end dates in news reports

This approach works best for **sudden, discrete events** that can be paired with **continuous physical Earth observation data**.

Tutorial: The Enriched Dataset

We've published an enriched version of Groundsource with decoded lat/lon centroids, duration calculations, and regional metadata:

```
from datasets import load_dataset

ds = load_dataset("rdjarbeng/groundsource-enriched")
df = ds['train'].to_pandas()

# Columns: uuid, area_km2, start_date, end_date,
#          longitude, latitude, year, month, duration_days, region

# Quantify Africa gap
africa = df[df['region'] == 'Africa']
print(f"Africa: {len(africa):,} events ({100*len(africa)/len(df):.1f}%)")

# Monthly time series by region
monthly = df.groupby(['year', 'region']).size().unstack(fill_value=0)
print(monthly.tail(5))
```

Resources

-  [Enriched Dataset on Hugging Face](#) — Decoded coordinates, geographic regions, durations
-  [Full Interactive Space](#) — Complete interactive exploration with maps
-  [Original Dataset on Zenodo](#) — CC-BY 4.0
-  [EarthArXiv Preprint \(Mayo et al., 2026\)](#) — *Groundsource: A Dataset of Flood Events from News*
-  [Google Blog](#)
-  [Google Research Blog](#)

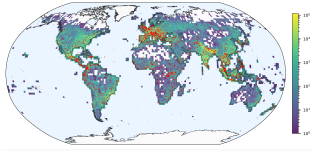
Key Scientific References

- **Riverine AI Benchmark:** Nearing et al. (2024). *Global prediction of extreme floods in ungauged watersheds*. [Nature 627, 559–563](#).
- **State-Space Hydrology:** Shams Eddin et al. (2025). *RiverMamba: A State Space Model for Global River Discharge and Flood Forecasting*. [arXiv:2505.22535](#).
- **All-Weather SAR Flood Mapping:** Alberti et al. (2024). *Kuro Siwo: 33 billion m² under the water*. [DOI:10.52202/079017-1204](#).
- **Epidemic Text Mining:** *Epidemic Information Extraction for Event-Based Surveillance Using Large Language Models*. [DOI:10.1007/978-981-97-4581-4 17](#).
- **WHO Outbreak Knowledge Graph:** *An epidemiological knowledge graph extracted from WHO Disease Outbreak News (eKG)*. [Scientific Data \(2025\)](#).
- **Satellite Health Equity:** *DengueNet: Dengue Prediction using Spatiotemporal Satellite Imagery for Resource-Limited Countries*. [arXiv:2401.11114](#).
- **Synthetic African Data:** *SAGDA: Open-Source Synthetic Agriculture Data for Africa*. [arXiv:2506.13123](#).
- **Physics-Guided Neural Networks:** *AirPhyNet: Harnessing Physics-Guided Neural Networks for Air Quality Prediction*. [arXiv:2402.03784](#).

The original Groundsource dataset is by Google Research, licensed CC-BY 4.0. Independent analysis and enriched dataset by [rdjarbeng](#).

Related Concepts & Posts

The following cards provide further definitions of concepts discussed above, and possible links to related posts.

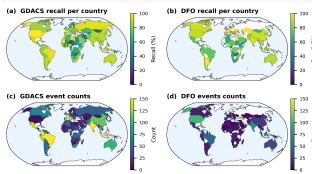


Groundsource Enriched Dataset

Dataset

Hugging Face

Explore Dataset

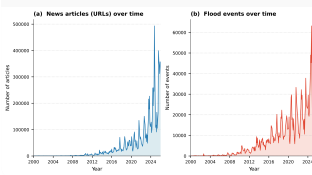


Interactive Analysis & Maps

Demo

Interactive

Open Space



Google EarthArXiv Preprint

Paper

Research

Read Preprint