Scientific and Social Research, 2026, Volume 8, Issue 8
http://ojs.bbwpublisher.com/index.php/ssr
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Understanding of Floods by Analyzing Open-Source Data

Yulu Yan*

The University of New South Wales, Kensington 2052, NSW, Australia

****Author to whom correspondence should be addressed.***

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: This study investigates the reliability of Google Alerts as an open-source information source for flood emergency response. The study aims to determine whether reports about floods retrieved through Google Alerts provide consistent, accurate, and spatially relevant information. The study uses quantitative content analysis and a spatial and temporal research design. Fifty flood events are selected as the reference dataset. Numerical analysis is conducted to compare the number and timing of reports across these events. Natural language processing and named entity recognition are used to identify location information, including place names, rivers, roads, suburbs, and infrastructure. The extracted locations are standardized and compared with reference information for each event to assess content consistency, geographic accuracy, and reporting completeness. The results indicate variation in the number of reports retrieved across flood events. Earlier flood events generally produce fewer relevant reports than more recent events, suggesting that the availability of historical information is limited. Search outcomes are also influenced by keyword selection, the level of location detail, and the time at which the search is conducted. Despite these limitations, Google Alerts can identify specific towns and local infrastructure within affected areas. The reports also contain useful information on flood impacts, economic losses, disaster assessment, and recovery activities. These findings suggest that Google Alerts can provide a broad record of the social and economic consequences of floods, although its usefulness varies across events and time periods.

Keywords: Open-source data; Named entity recognition; Google Alerts

Online publication: August 31, 2026

1. Introduction

1.1. Background on the use of open-source data in flood monitoring

Floods are among the most frequent and destructive natural hazards, creating a need for timely information during flood events [1]. Flood forecasting also depends on high-quality data from diverse observational sources [2]. Public reports published through social media and news websites can provide early spatiotemporal information that helps emergency agencies identify affected areas [3]. In the Zhengzhou "7.20"

flood, for example, locations identified from online information were strongly consistent with satellite imagery [4]. Nevertheless, open-source data may contain noise, inconsistent information, and ambiguous place names [5]. Its availability therefore does not guarantee spatial completeness, accuracy, or timeliness.

Unlike a single social media platform, Google Alerts retrieves information from multiple online sources. This may provide local and contextual flood information, but its performance across historical and event-specific searches has not been assessed. It remains unclear whether Google Alerts retrieves sufficient information across events, represents recorded flood locations, or provides information at operationally useful times.

1.2. Literature review

1.2.1. Current techniques in information extraction

Natural Language Processing (NLP) converts unstructured text into information that can be systematically analyzed, although its performance depends on the quality and structure of the input [6]. Named Entity Recognition (NER), a core NLP task, identifies entities such as people, organizations, and locations [4]. In flood reports, it can structure references to towns, suburbs, regions, and rivers for spatial analysis.

NER has developed from dictionary-based and rule-based techniques to machine-based and deep-learning approaches. Dictionary-based methods perform poorly when terms are ambiguous or highly variable, while rule-based systems depend on handcrafted linguistic patterns and may not generalize to informal, contextually variable text [7–9]. Machine-learning models can identify entities using contextual cues, but require large annotated datasets [10–11]. Deep-learning methods such as CNNs and RNNs have outperformed traditional models such as SVMs in text-classification tasks [5]. The present study therefore applies spaCy's pretrained NER model to extract geographic entities from Google Alerts reports.

1.2.2. Current progress in disaster information extraction

Online disaster content is frequently unstructured, informal, or irrelevant, making data quality important to NLP performance [6]. Nevertheless, automated methods have successfully extracted useful disaster information. Peng and Zhang (2024) combined social media and remote-sensing data to assess flood levels and obtain spatial and temporal observations [4]. Islam, Rabbi and Hossain (2024) used CNNs to classify disaster-related social media images, while Sufi and Khalil (2022) incorporated sentiment and location information into automated disaster detection [12–13]. Kumar and Singh (2024) also extracted geographic references from bilingual disaster-related social media text [14]. Collectively, these studies show that computational methods can transform online content into structured disaster information. Building on this foundation, the present study uses NER-extracted locations and publication dates to assess the spatial and temporal consistency of Google Alerts information.

1.2.3. Research gap

Previous studies have primarily developed extraction and detection methods or examined individual social media platforms and disaster cases. Less attention has been given to the reliability of information retrieved by an aggregated, multi-source service across multiple flood events. In particular, Google Alerts has not been systematically evaluated in terms of retrieval volume, correspondence with reference flood locations, and publication timing relative to recorded event periods. This study addresses this gap by distinguishing numerical availability from spatial and temporal consistency, thereby assessing whether Google Alerts can

supplement flood monitoring and where its limitations restrict its use in real-time emergency response.

1.2.4. Research aim

This study evaluates the reliability of Google Alerts flood information through numerical availability, spatial consistency, and temporal consistency, using flood events from the Emergency Events Database (EM-DAT) [15]. Numerical availability examines the presence and volume of retrieved information without treating quantity as evidence of accuracy. Spatial consistency compares locations extracted using Named Entity Recognition with EM-DAT reference locations, while temporal consistency compares validated publication dates with recorded flood periods [15]. Together, these dimensions determine whether Google Alerts can supplement flood monitoring and situational understanding and identify the limitations of its use for real-time emergency response.

2. Methodology

2.1. Research design

This study adopted a quantitative design to evaluate the reliability of Google Alerts as a source of flood information. Reliability was assessed through three dimensions: numerical availability, spatial consistency, and temporal consistency. EM-DAT provided the reference flood events, affected locations, and event dates, while Google Alerts provided the open-source information [15]. The research process involved selecting reference events, collecting Google Alerts results, extracting geographic entities, and comparing the retrieved information with EM-DAT spatially and temporally [15].

2.2. Selection of reference flood events

The reference dataset consisted of 50 flood events recorded in New South Wales by EM-DAT from 2000 onwards [15]. The events were arranged chronologically and assigned Event I Ds from 1 to 50, allowing information to be linked consistently across the EM-DAT records, Google Alerts results, and subsequent analyses [15]. For each event, the recorded affected locations were used as the spatial reference, while the start and end dates were used as the temporal reference.

2.3. Google Alerts data collection

A separate Google Alerts search was conducted for each reference event. Each search query combined the affected location, the term "flood," and the month and year of the event to distinguish it from other floods. Retrieved results were assigned to the corresponding Event ID for comparison with the appropriate EM-DAT record [15]. All results were retained, including those that appeared irrelevant, because they also reflected the performance of Google Alerts as an information source.

2.4. Numerical availability assessment

Numerical availability was assessed by examining the number and distribution of Google Alerts results across the 50 events. Descriptive statistics, including the total, mean, median, minimum, and maximum, were calculated to determine whether information was available for each event and how retrieval volume varied between events.

2.5. Natural language processing and named entity recognition

Google Alerts results primarily contained unstructured natural-language text. To support spatial comparison, geographic information was converted into structured data using Named Entity Recognition (NER). A pretrained spaCy NER model was applied to identify geographic entities, including settlements, administrative areas, and geographic features. The study applied an existing model rather than developing or training a new one. Extracted entities were cleaned and standardized to reconcile different expressions of the same location. For example, "NSW" was standardized as "New South Wales." Repeated location expressions were then consolidated to produce the location dataset used in the spatial assessment.

2.6. Spatial consistency assessment

Spatial consistency was assessed by comparing locations extracted from Google Alerts with EM-DAT reference locations within the corresponding Event ID [15]. The analysis involved location matching, calculation of event coverage, and classification of event-level consistency.

2.6.1 Location Match Classification

Each EM-DAT reference location was classified as an Exact Match, Partial Match, or No Match [15]. An Exact Match indicated direct correspondence between the standardized locations. A Partial Match indicated an identifiable geographic relationship without an exact match in name or spatial extent. A No Match indicated that no corresponding relationship could be identified. These categories measured correspondence at the individual-location level.

2.6.2. Event Coverage

Spatial coverage was calculated separately for each event. Exact and Partial Matches were treated as reference locations represented in the Google Alerts data. Event Coverage was calculated as:

$$\text{Event Coverage} = (\text{Exact Matches} + \text{Partial Matches}) / \text{Total EM-DAT Reference Locations} \times 100\ \%$$

The mean Event Coverage was also calculated to describe spatial coverage across the full sample.

2.6.3. Spatial consistency classification

Events with coverage of at least 75 percent were classified as High consistency, those with coverage from 40 percent to below 75 percent as Medium consistency, and those below 40 percent as Low consistency. These categories described overall performance at the event level, whereas Exact, Partial, and No Match referred to individual locations.

2.7. Temporal consistency assessment

Temporal consistency was assessed by comparing the publication date of each Google Alerts result with the corresponding EM-DAT start and end dates [15]. Candidate publication dates were extracted from available webpages, checked, and standardized to address differences in format and date order. Dates that could be established with sufficient confidence were linked to the relevant Event ID and classified as Before Event, During Event, or Post Event.

Results without a reliably identifiable publication date were classified as Unknown and retained in the dataset. An Unknown classification did not necessarily indicate that the original webpage contained no date; rather, it indicated that the date could not be determined confidently through the extraction and validation procedure. Percentages for the Before, During, and Post categories were calculated using only results with reliable dates, while Unknown results were reported separately.

3. Results

3.1. Numerical availability

A total of 913 Google Alerts results were retrieved for the 50 flood events. The mean number of results was 18.3 per event, while the median was 20. The number of results ranged from 1 to 26 per event.

The distribution of results indicates that Google Alerts returned information for all 50 flood events, although the amount of information differed between events. Most events were concentrated within a relatively narrow range, with 88 percent of the events returning between 11 and 20 results. A smaller number of events produced considerably fewer results, while the highest number of results for a single event was 26. **Figure 1** presents the number of Google Alerts results retrieved for each of the 50 flood events. The distribution shows that most events produced result counts close to the overall mean, while several events at the beginning of the dataset returned substantially fewer results. Event 50 produced the highest number of results.

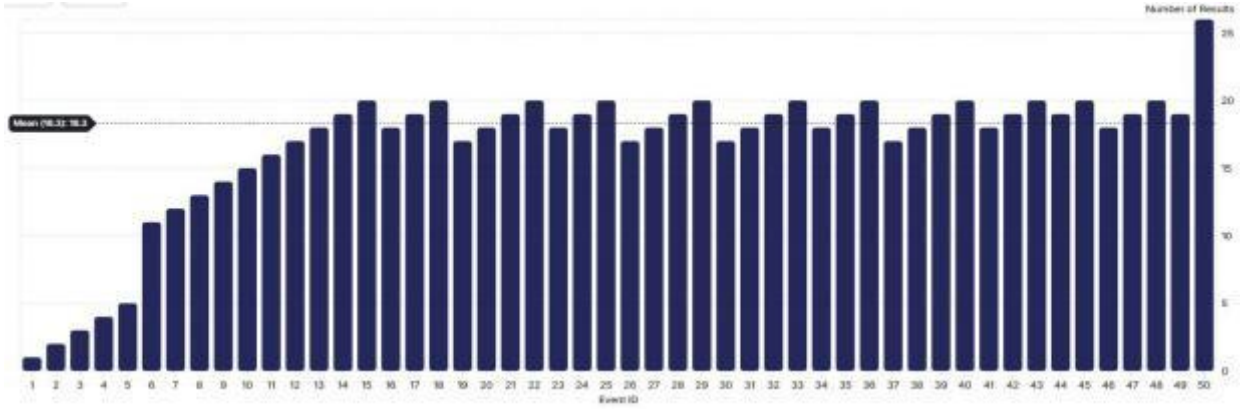


Figure 1. Number of Google Alerts results

3.2. Spatial consistency

3.2.1. Geographic information extracted from Google Alerts

The Named Entity Recognition process identified 1,243 location mentions from the Google Alerts reports. After location processing and consolidation of repeated geographic expressions, 421 unique locations were identified. These extracted locations were used to assess the spatial consistency of Google Alerts against the EM-DAT reference dataset [15]. Across the 50 flood events, EM-DAT contained 376 reference locations for spatial comparison [15].

3.2.2. Location match results

Location match results are shown in **Figure 2**. Therefore, 126 of the 376 reference locations had either an

exact or partial correspondence with geographic information extracted from Google Alerts. The location results show that Google Alerts reports contained geographic information associated with the selected flood events, but a substantial proportion of the reference locations were not represented in the extracted location dataset.



Figure 2. Location match results

3.2.3. Event spatial coverage

Location matching was subsequently aggregated at the event level. The result is shown in **Figure 3**. For each flood event, spatial coverage represented the proportion of EM-DAT reference locations that received either an Exact Match or a Partial Match in the Google Alerts data [15]. The mean reference location coverage across the 50 events was 49.80 percent. This indicates that, on average, approximately half of the reference locations associated with an event were represented either exactly or partially in the geographic information extracted from Google Alerts. 13 of the 50 events were classified as High consistency, with reference location coverage of at least 75 percent. A further 15 events were classified as Medium consistency, with coverage from 40 percent to below 75 percent. The remaining 22 events were classified as Low consistency, with coverage below 40 percent. In percentage terms, High, Medium, and Low consistency represented 26 percent, 30 percent, and 44 percent of the 50 flood events, respectively. The largest group was therefore the Low consistency category. Taken together, the location and event results show substantial variation in spatial representation. Although some flood events achieved high reference location coverage, almost half of the events were classified as having low spatial consistency.

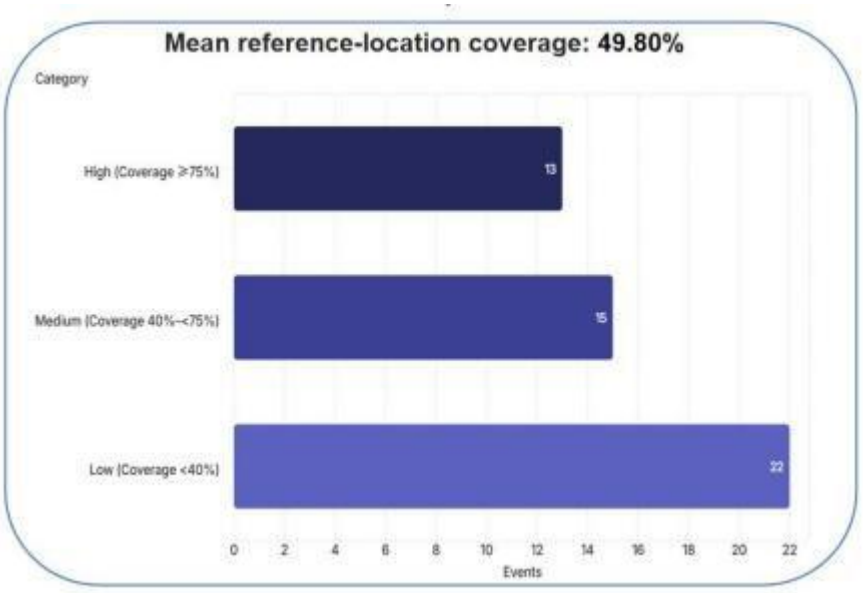


Figure 3. Event spatial coverage results

3.3. Temporal consistency

Temporal consistency was assessed across all 913 Google Alerts results. Candidate publication dates were initially extracted from 343 results, representing 37.6 percent of the complete dataset. Following further validation, 155 results contained publication dates that were considered sufficiently reliable for temporal classification, accounting for 17.0 percent of all results. The remaining 758 results could not be reliably classified according to publication timing. This included results for which no candidate date was identified and candidate dates that did not satisfy the validation process. The relatively small number of reliable publication dates was mainly related to differences in webpage structure and date presentation. The retrieved results came from different websites, where publication dates were stored and displayed in different formats. Some pages contained several dates, such as publication, update, or event dates, making it difficult to identify the actual publication date consistently. Differences in date format and date order also created ambiguity during automated extraction. Candidate dates that could not be interpreted with sufficient confidence were therefore excluded from the final temporal classification.

The 155 results with reliable publication dates were compared with the corresponding EM-DAT flood start and end dates and classified as Before, During, or Post [15]. Among these results, 7.1 percent were published before the recorded flood period, 18.1 percent were published during the flood period, and 74.8 percent were published after the flood period. The Post category was therefore dominant, accounting for approximately three quarters of the temporally classified results, while less than one fifth were published during the active flood period. These percentages refer only to the 155 results with reliable publication dates and not to the complete set of 913 Google Alerts results.

Overall, the temporal assessment reveals two main patterns. First, reliable publication dates were available for only a relatively small proportion of the retrieved information. Second, among the results that could be temporally classified, publication was concentrated predominantly after the recorded flood period. These findings indicate a clear difference between the availability of flood-related information through Google Alerts and the timing of that information relative to the corresponding flood event.

4. Discussion and conclusion

The findings of this study suggest that the reliability of Google Alerts depends on how and when the information is used. Google Alerts is most reliable when it is used as a complementary source to provide additional local and contextual information about flood events. Its ability to identify towns, suburbs, rivers, and infrastructure can provide useful geographic detail, while online reports can also contribute information about flood impacts, losses, and recovery. This information may complement broader records provided by official disaster databases and support a more detailed understanding of an event.

However, Google Alerts is less reliable when complete spatial coverage or timely information during an active flood is required. The information retrieved is not equally representative across all flood events, and officially recorded affected locations are not consistently reflected in the search results. The temporal findings also indicate that Google Alerts is more suitable for understanding event impacts and subsequent developments than for continuous monitoring during the active stage of a flood. Its usefulness may also depend on the search terms used and the time at which the search is conducted.

These findings highlight an important distinction between information availability and information reliability. The ability to retrieve flood-related information does not necessarily mean that the information provides complete spatial representation or is available at the time it is most needed.

Overall, Google Alerts can be considered conditionally reliable for flood information. I t is most useful when local and contextual information is required and when its results can be compared with authoritative sources. I t should not be relied upon as an independent emergency monitoring system when complete geographic coverage or timely operational information is essential. For emergency management, Google Alerts is therefore better positioned as a supplementary source that can extend, rather than replace, official disaster information.

5. Limitations

Several limitations should be considered when interpreting the findings of this study. First, the data were collected retrospectively rather than during the original flood events, so the retrieved results may not fully represent the information available at the time of each event. Second, Google Alerts uses search and ranking mechanisms that are not transparent, and the results may be influenced by keyword selection and the date of data collection. Third, publication dates could not be reliably identified for many results, which limited the scope of the temporal assessment.

Disclosure statement

The author declares no conflict of interest.

References

[1] Nearing G, Cohen D, Dube V, et al., 2024, Global Prediction of Extreme Floods in Ungauged Watersheds. Nature, 627(8004): 559–563. https://doi.org/10.1038/s41586024071451

[2] Das J, Manikanta V, Nikhil Teja K, et al., 2022, Two Decades of Ensemble Flood Forecasting: A State of the Art on Past Developments, Present Applications and Future Opportunities. Hydrological Sciences Journal, 67(3):

477–493. https://doi.org/10.1080/02626667.2021.2023157

[3] Rodriguez KM, Ofori SK, Bayliss LC, et al., 2020, Social Media Use in Emergency Response to Natural Disasters: A Systematic Review with a Public Health Perspective. Disaster Medicine and Public Health Preparedness, 14(1): 139–149. https://doi.org/10. 1017/dmp.2020.3

[4] Peng J, Zhang J, 2024, Spatiotemporal Assessment of Urban Flooding Hazard Using Social Media: A Case Study of Zhengzhou "7·20". Environmental Modelling & Software, 2024(176): 106021.

[5] Szczepanek R, 2023, A Deep Learning Model of Spatial Distance and Named Entity Recognition (SD NER) for Flood Mark Text Classification. Water, 15(6): 1197.

[6] Tounsi A, Temimi M, 2023, A Systematic Review of Natural Language Processing Applications for Hydrometeorological Hazards Assessment. Natural Hazards, 116(3): 28192870.

[7] Sasaki Y, Tsuruoka Y, McNaught J, et al., 2008, How to Make the Most of NE Dictionaries in Statistical NER. BMC Bioinformatics, 9(Suppl 11): S5.

[8] Li J, Sun A, Han J, et al., 2020, A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1): 50–70.

[9] Seow WL, Chaturvedi L, Hogarth A, et al., 2025, A Review of Named Entity Recognition: From Learning Methods to Modelling Paradigms and Tasks. Artificial Intelligence Review, 58(10): 315.

[10] Lai K, Porter JR, Amodeo M, et al., 2022, A Natural Language Processing Approach to Understanding Context in the Extraction and Geocoding of Historical Floods, Storms, and Adaptation Measures. Information Processing & Management, 59(1): 102735.

[11] Lawson N, Eustice K, Perkowitz M, et al., 2010, Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk. In Proceedings of the NAACLHLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk (pp. 71–79).

[12] Islam MA, Rabbi F, Hossain NU, 2024, Performance Evaluation of NLP and CNN Models for Disaster Detection Using Social Media Data. Social Network Analysis and Mining, 14(1): 213. https://doi.org/10.1007/s13278-024-01374-y

[13] Sufi FK, Khalil L, 2022, Automated Disaster Monitoring From Social Media Posts Using AI-Based Location Intelligence and Sentiment Analysis. IEEE Transactions on Computational Social Systems, 11(4): 1–11.

[14] Kumar A, Singh JP, 2024, Deep Neural Networks for Location Reference Identification from Bilingual Disaster-Related Tweets. IEEE Transactions on Computational Social Systems, 2024(11): 880–891. https://doi.org/10.1109/TCSS.2022.3213702

[15] UCLouvain/CRED, 2026, Emergency Events Database (EM-DAT). Brussels, Belgium. https://www.emdat.be/

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.