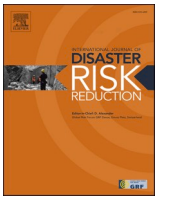
International Journal of Disaster Risk Reduction 74 (2022) 102908

Contents lists available at ScienceDirect

International Journal of Disaster Risk Reduction

journal homepage: www.elsevier.com/locate/ijdrr



Social sensing of flood impacts in India: A case study of Kerala 2018

James C. Young [a,*], Rudy Arthur [a,1], Michelle Spruce [a,1], Hywel T.P. Williams [a,b,1]

[a] Computer Science, University of Exeter, Innovation Centre, North Park Road, Exeter, EX4 4RN, UK
[b] Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, UK

ARTICLE INFO

Keywords:
Natural hazard
Flooding
Social sensing
Social media
Telegram
Twitter

ABSTRACT

Flooding is a major hazard that is responsible for substantial damage and risks to human health worldwide. The 2018 flood event in Kerala, India, killed 433 people and displaced more than 1 million people from their homes. Accurate and timely information can help mitigate the impacts of flooding through better preparedness (e.g. forecasting of flood impacts) and situational awareness (e.g. more effective civil response and relief). However, good information on flood impacts is difficult to source; governmental records are often slow and costly to produce, while insurance claim data is commercially sensitive and does not exist for many vulnerable populations. Here we explore "social sensing" – the systematic collection and analysis of social media data to observe real-world events – as a method to locate and characterise the impacts (social, economic and other) of the 2018 Kerala Floods. Data is collected from two social media platforms, Telegram and Twitter, as well as a citizen-produced relief coordination web application, Kerala Rescue, and a government flood damage database, Rebuild Kerala. After careful filtering to retain only flood-related social media posts, content is analysed to map the extent of flood impacts and to identify different kinds of impact (e.g. requests for help, reports of medical or other issues). Maps of flood impacts derived from Telegram and Twitter both show substantial agreement with Kerala Rescue and the damage reports from Rebuild Kerala. Social media content also detects similar kinds of impact to those reported through the more structured Kerala Rescue application. Overall, the results suggest that social sensing can be an effective source of flood impact information that produces outputs in broad agreement with government sources. Furthermore, social sensing information can be produced in near real-time, whereas government records take several months to produce. This suggests that social sensing may be a useful data source to guide decisions around flood relief and emergency response.

1. Introduction

Flooding is a major global hazard. Floods are expected to increase in frequency and severity as a consequence of climate change, as well as urbanisation and land use change, especially in the developing world [1,2]. In this paper, we will examine the Kerala flood event in 2018. The flood is generally agreed to have peaked on 16th August 2018, but heavy rains before this date caused floods and landslides in hilly areas, while floodwaters did not recede until many days later [3]. A major relief effort was launched, with more than

* Corresponding author.
E-mail address: jcy204@exeter.ac.uk (J.C. Young).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ijdrr.2022.102908
Received 14 October 2021; Received in revised form 2 February 2022; Accepted 10 March 2022
Available online 21 March 2022

3000 camps opened to temporarily house displaced people and substantial financial commitments by state and national governments [4]. These floods caused an estimated $4.4 billion in economic damage, including the destruction of 17,000 homes with an additional 217,000 damaged [5], and resulted in the deaths of over 430 people [5]. In addition to their direct impacts, the floods also had a number of subsequent impacts, including disease outbreaks, [6,7] and landslides [8].

Climate change and land use changes look set to dramatically increase the number of extreme flood events in India [9,10]. As the Indian population urbanises, with 52% of the population expected to live in cities by 2050 [11], urban flood planning and resilience will become more important [12], although impacts in 2018 were widespread in both rural and urban areas. The 2018 Kerala floods have been studied extensively and variously linked to climate change [13], land use change, such as deforestation [14], mining [15], and poor reservoir management [3]. These studies are part of a growing body of literature on major floods in India, with prior events in Uttarakhand (2013), Srinagar (2014), Chennai (2015) and Gujarat (2017). Ray et al. [16] emphasise the urgent need for improved planning, response and resilience measures in Indian flood management.

Good information is crucial for all of these goals. Modelling and forecasting flood events is an important task but a key part of informing risk and forecast models is post-crisis data acquisition so that models and plans can be evaluated and improved. A number of data sources have been investigated in this regard. Avashia and Garg [17] used a combination of satellite data (Landsat) and newspaper data to study the effects of land use change on urban flooding over a period from 1991 to 2017. Gupta and Nikam [18] used 35 automatic weather stations to monitor rainfall in Mumbai and provide live information to emergency responders and the public.[2] Publicly available satellite data has also been used to survey the flood extent [19,20].

In the wake of the Kerala floods, some have pointed to the failure of local and state governments to respond appropriately and made the case for a more community-focused disaster response plan [21]. Especially in the wake of the Chennai floods in 2015, academics have started to recognise the importance of social media as a means for communities to share vital information, request help and communicate during flood events [22,23]. Social sensing is a way for the concerns and impacts of floods on communities to be collected and heard by the government at mass scale. Across the world, National Meteorological and Hydrological Services (NMHS) are increasingly moving towards impact-based forecasting [24,25], with forecasts including details of expected impacts to people and property as well as the expected weather conditions. Validation and evaluation of these forecasts requires reliable impact data which provides details of how and where people and property were affected. Information on social and economic impacts in the wake of an extreme weather event is typically not as easily accessible as traditional meteorological observations, not widely dispersed, not in the correct format and is often only available with a significant time lag after a weather event has occurred [26]. Having information on impacts more readily available close to the time of a weather event occurring would help to improve situational awareness about where and how impacts of a weather event are being experienced.

Sources of impact data include news reports, social media, citizen data collated by government or official agencies, insurance data and eye-witness accounts. While citizen data collected by government or official agencies are usually of good quality, they can be slow and costly to gather and are not consistently produced for all events. Social sensing is an approach developed in recent years to utilise unsolicited social media data to detect and analyse real-world events. It has been applied in a variety of contexts, such as earthquakes [27], floods [28,29], heatwaves [30], hurricanes [31,32] and wildfires [33]. It has proved its usefulness as a source of information about where, when and how individuals are being impacted by a specific event [28,30,34]. Therefore its use as a source of social and economic impact information in the wake of a natural disaster, such as widespread flooding, would greatly assist with the collation of data to support improved emergency management, response and resilience.

Here we explore the utility of social sensing to measure impacts of flooding using the 2018 Kerala event as a case study. We focus on social media data from the platforms Telegram and Twitter, with comparison to data from the Kerala Rescue citizen relief coordination web application[3] and the Rebuild Kerala database of government aid for property damage.[4] Despite having the third-highest number of active Twitter accounts (24 million [35]), India's immense population means that the proportional coverage of Twitter is low. This has limited the resolution of previous Twitter social sensing studies in India [36–38]. Conversely, Telegram has 220 million users in India, with the platform having recently experienced significant growth [39]. Telegram is a messaging platform that is widely used in some parts of the world, including Kerala, and permits large groups of up to 200,000 users to be created [40]. Specifically in this project, we study the "KeralaGram" group on Telegram, which had 15,000 users at the time of the 2018 flood and was focused on issues/events/news related to the state of Kerala. While Twitter has been extensively used for social sensing, the use of Telegram is less common. Most relevant Telegram research involves either the production of Telegram bots for distributing hazard information on the platform [41,42], or contains limited investigation and validation of hazard impacts and spatio-temporal trends from the event [43–45]. Additionally, to the best of our knowledge, all previous Telegram social sensing studies were carried out on Telegram Channels, where an admin broadcasts messages to channel followers (one-to-many communication). Channels are appropriate for news distribution, however, are not conducive to promoting organic conversation for social sensing. This work investigates Telegram chats called supergroups. These allow for many-to-many conversations, providing a much richer form of social impact data. The Kerala Rescue application was set up initially by student volunteers as a mechanism to coordinate requests and offers of help during the 2018 flood event, quickly becoming a widely used service supported by Kerala's state government. This citizen science initiative differs from social sensing using social media, in that the information it utilises is structured into a database format designed for the purpose, whereas social media content is unsolicited and unstructured. The processing requirements for these platforms therefore differ. The

[2] http://dm.mcgm.gov.in/livefullpage.
[3] https://keralarescue.in/.
[4] https://rebuild.lsgkerala.gov.in/rebuild2018/.

Rebuild Kerala database was put together by the Government of Kerala after the 2018 event to manage the distribution of financial aid related to property damage [46]. Reports were collected by citizen surveys, validated by local officials, and then either approved/rejected for aid by the state government. More details on these data sources and how the data from each was processed is given below.

The experiment performed in this study is a triangulation between the different data sources. While accurate "ground truth" information may be constructed for the extent of floodwaters [20], it is very difficult to collate accurate ground truth information for the social impacts of flooding. A disastrous flood event such as Kerala 2018 affected many people in many ways, with each individual having their own story. As such, the true social and economic impacts of the flood may never be known completely. Sources of economic information such as the financial losses and compensation, property damage, damage to agricultural land, and disruption to economic activities and livelihoods, are diverse and difficult to unify into a single metric. Measurement of social impacts beyond the tragic loss of life is also hard to perform; the level of distress, loss of health and wellbeing, and impacts on families and communities are almost impossible to quantify. Therefore our study does not seek to use the social sensing approach to accurately recover some known ground truth observation. Instead, it compares and contrasts the view of the flood impacts that are offered by several different sources: two social media platforms, a citizen relief app, and a governmental database. The rigorous analysis of Telegram in this study is novel, providing clear evidence that the platform can and should be used for social sensing. Additionally, social sensing studies are rarely validated to this depth. This systematic comparison of four very different data sources is a unique approach for investigating and validating flood impacts and social sensing itself. Each of the sources might be expected to reveal a part of the overall situation during the flood event. Their agreement in when and where flood impacts are observed, and the nature of the information they contain, will help to evaluate the potential for such data to be used in future flood events to support decision-making and planning.

The paper is structured as follows. Section 2 covers the Data Collection & Methodology by which information was obtained from each source. This is followed by Results in section 3, covering the findings from the study, before the Discussion in section 4 highlights some of the main implications and limitations of the work.

2. Data Collection & Methodology

Here we explore the potential for detecting the impacts of the Kerala floods using publicly accessible data from Telegram and Twitter. Data from both platforms is collected and filtered for relevance, then used to create time series plots and spatial maps of flood-related activity. These are compared to similar figures based on public submissions to the Kerala Rescue app, a citizen-generated web service set up during the flood event to allow affected people to submit requests for help [47]. A further comparison is made to damage data collected after the event by the government-sponsored Rebuild Kerala initiative [46]. While it is known that WhatsApp was widely used during the event to coordinate rescue and relief efforts [48], we are not able to study WhatsApp data due to privacy restrictions on the platform. As well as considering the ability of Telegram and Twitter data to accurately map the social impacts of the flood event, we also examine content from each platform as a source of data about what kinds of impact were experienced.

2.1. Telegram

Telegram is a social media messaging platform which, like many other messaging services, allows for end-to-end encrypted conversations and groups. The feature that differentiates Telegram from its competition is public group chats called *supergroups*. With up to 200,000 members [40], supergroups promote conversation within local communities [49]. For this work, conversations from the supergroup 'KeralaGram' were downloaded using the inbuilt 'Export Chat' function in the Telegram desktop application. Messages were received in JSON format, and shared media was received in its original format (i.e. MP3 for audio and JPEG for images). With around 15,000 members, 'KeralaGram' is the largest public community discussion group in Kerala. The data used within this study is publicly accessible, with all processing and analysis following the Telegram terms of service.

During the flood period, a rise in the amount of posts sharing other media was detected, with a 450% increase in shared images compared to the same date range within the previous month. These images consisted largely of flood photos, alongside screenshots of warnings and messages from other platforms. To utilise this media increase, basic computer vision was applied, automatically adjusting the image color and contrast to optimise the extraction of English and Malayalam text using the OCR library Tesseract [50]. Next, the language detection Python library "langdetect" was applied to the messages [51]. While this package occasionally misclassifies text, its purpose here is to provide a rough overview of the languages present. With only 28% of the dataset being detected as English, machine translation was needed. To translate the remaining text, the XLSX document translation Google Translate[5] feature was used. Bots are often found in social media data, however, due to the presence of admins and strict rules within this chat, the conversation post-translation was highly regulated and did not require bot filtering. Additionally, after location inference (see section 2.5), relevance filtering was not required as the flood caused a large shift in messaging content and structure, with a 98% relevance of located messages (see Fig. 1). A post was only considered relevant if it was directly discussing the 2018 Kerala flood, for instance requesting help or sharing information about the event. If there was any ambiguity, the entry was classified as irrelevant. A sample filtered Telegram message is shown in Table 1. The final dataset contained 13,614 messages and 1188 images containing text between 1st August - 23rd August.

[5] https://translate.google.co.uk/?sl=auto&tl=en&op=docs.

Table 1
Overview of Telegram and Twitter messages after filtering.

Platform	Fictitious posts using similar language to the datasets	Average post length (words)	Percentage of messages in English pre-translation
Telegram	We have 50 students trapped in the Munnar engineering college. While we are safe, we need water and food urgently. Please call us on the number below.	77.52	28%
Twitter	@PMOIndia @CMOKerala, please help us to arrange rescue for over 5000 people stuck in Kuttanad, Alappuzha. We need rescue boats and a helicopter immediately.	24.95	90%

2.2. Twitter

Twitter is a social media microblogging platform, where users produce character restricted messages called tweets. Tweets are stored in JSON format and can be collected for free in real-time or purchased retrospectively from Twitter's Historical PowerTrack service. For this work, the PowerTrack service was used to obtain all tweets identified by Twitter as originating from India (that is, having an 'IN' country code in their location metadata) between 1st August to 23rd August. This collection contained 1,363,659 tweets. All tweets in this collection contain either exact tweet coordinates from GPS enabled devices or a user-defined tweet 'place' attribute, which is used by Twitter to designate their country code. The following analysis and investigation is consistent with the Twitter terms of service.

The first filtering step was extracting English language tweets. While machine translation of tweets in other local languages is possible, it is slow to perform and therefore infeasible for the high volume of tweets in the dataset. Despite this, the language attribute within the tweet JSON objects highlighted that English was the most common language in our dataset, with 90% of tweets in English; therefore, simply extracting the English tagged tweets was deemed sufficient. Location inference (see section 2.5) was used to locate and identify tweets that both concerned and originated from Kerala. This decreased the dataset to 24,414 tweets. Inspection of tweet volumes suggested that bot removal was not required as no account had tweeted excessively (following the >1% rule successfully implemented by Arthur et al. [28]. The tweet collection was geographic, rather than thematic, so to extract relevant tweets a list of keywords was produced, containing both common English-language flood terms and words frequently used in Kerala Rescue requests (Appendix A.1). If the tweet text did not overlap with this list (i.e. contained no words in the keyword list), the tweet was ignored. This resulted in 7097 tweets related to floods.

Next, a manual inspection of the remaining tweets highlighted numerous duplicates, with only minor differences such as the tagged user or a hashtag. The majority of these posts originated from automated and institutional Twitter accounts and did not include useful on-the-ground observations of flood impacts. To remove these, tweet texts were vectorised by word frequency using the scikit-learn package [52] in Python and the cosine similarity between them was calculated. To decrease computation times, stop words and punctuation were removed from tweets before vectorisation. If any two vectors had a 97% or above cosine similarity between them, then they were deemed as duplicates, and only the tweet posted first was kept. This successfully removed obvious duplicates. However, large groups of messages that differed by only a few words remained. To remove these, if at least 10 vectors showed a cosine similarity above 90%, the corresponding messages were classified as duplicates and once again only the original tweet was kept in the dataset. While the 97% and 90% thresholds were heuristically chosen, manual inspection showed they were appropriate for removing unwanted messages. Removal of duplicates left 6936 tweets for further analysis. Manual inspection (using the same criteria as section 2.1) of a randomly selected 20% sample of the remaining tweets showed a 96% relevance indicating that further relevance filtering was not required. A sample filtered tweet is shown in Table 1.

2.3. Kerala Rescue

Kerala Rescue is an emergency request platform, allowing those impacted by the flood to submit a request for services including rescue, food, water, clothing, and medicine. It was created by a college student from Kerala during the early stages of the flood [47]. After 10 days, an additional 200 software developers were actively developing the site, with 50,000 registered Kerala volunteers responding to and validating the 45,000 requests [47]. The data was collected with permission from the developers in JSON format from a public Slack channel active during the floods. Since the data is potentially sensitive, data was anonymised prior to analysis and no identifiable results are displayed.

The Kerala Rescue database contains help requests formatted as records that each contain 27 columns including location fields, rescue details, and requester contact number. The first processing step was the removal of duplicate entries. Due to the ambiguity between duplicate requests and follow up requests, if two entries with the same contact number were posted within an hour of each other, and if either their location or requester attributes were duplicated, then only the original entry was kept. This removed over 5000 entries, which upon manual inspection were correctly classified as duplicates. Location information was provided as both co-ordinate fields and as entries in a text field. To remove inaccurate location coordinates, if the 'coordinate_accuracy' attribute was over 1000 m, the corresponding coordinates were removed and other location data was favoured. The aforementioned language translation technique was applied to the text-form location field, before applying location inference to provide accurate coordinates for mapping (see below).

2.4. Rebuild Kerala

The Rebuild Kerala Initiative (RKI) was set up by the Government of Kerala to assist in the rebuilding and recovery efforts after the flood event [46]. The RKI includes a diverse range of projects, including the development and promotion of the Rebuild Kerala

Development Programme, the strategic roadmap for reconstruction across the state. Relevant to this study, RKI collected data by surveying affected citizens about property damage and made it publicly available as a database, with each report then verified by a local overseer. From this database, we extracted records for verified flood-damaged houses that were accepted for financial aid by the state government.

This data is stored in a comma-separated format with each row corresponding to one of the 1034 Kerala local administrative bodies (Municipalities, Corporations, and Gram Panchayats). Each record is the number of damaged properties approved for financial support within that local body. Manually assigning each location to a corresponding map polygon was the only preprocessing required. A few assumptions were necessary due to alternative spellings and duplicate place names in the data; however, these are unlikely to make an impact due to the volume of the data. In the dataset, most local bodies had corresponding values in the 'Total', 'Verified', 'Approved' and 'Rejected' surveys columns. For this work, the difference between the 'Verified' and 'Rejected' columns was taken as a conservative estimate of the true 'Approved' value, as many entries did not have a value for 'Approved'.

2.5. Location inference

To find on-the-ground observations rather than long-distance "news" posts, the data was geolocated, keeping only the data located in Kerala. Location inference is an automated procedure that identifies place names in the post and then assigns a map polygon (shapefile) for each location. To perform this location inference process, a package developed by Arthur et al. [28] was used, which is based on the work of Schulz et al. [53]. This involved cross-referencing words within a tweet/message/request to various gazetteers (GADM, DBpedia and Geonames [54–56]), before inferring the most likely location. Previous work has found this approach to provide highly accurate locations but, due to the lack of ground truth, this is difficult to quantify. Manual checks of outcomes were performed on an ad hoc basis to ensure validity throughout the process. Due to differences in the platforms, this final filtering step was applied as follows:

- **Telegram:** Location inference was applied to the extracted text from translated messages and images. If this approach was unsuccessful, and if a Kerala landline phone number was present in the message's 'phone' attribute, the location could be inferred using the area code as sourced from the Kerala Government website.[6]
- **Twitter:** Despite the dataset being a geographic collection, manual inspection showed that the exact location discussed in the tweet text frequently did not correspond to the (often large and unspecific) bounding box within the 'place' field in tweet metadata. While the cause of this is unknown, two potential explanations are from the incorrect entry of the 'place' attribute from the user, or from tweets being posted on behalf of someone, with the place tag related to the user location, rather than where they are referring to. Therefore, location inference was applied to the tweet text, with a manual sample showing that this approach had a higher accuracy in determining the location referenced by the post than when the 'place' attribute was used. If the tweet had geotagged coordinates, and if no location was inferred from the text, these coordinates were taken as the inferred location.
- **Kerala Rescue:** A similar approach to that used for Twitter was taken, with the user-entered 'location' field prioritised over the automated 'coordinate' field due to occasional coordinate inaccuracies. These inaccuracies were primarily from requests on behalf of those affected, with the coordinates corresponding to the individual sending out the request, and the 'location' field corresponding to the individual in need of help. If geolocating with this approach was unsuccessful, phone area codes were used as with Telegram posts.
- **Rebuild Kerala:** Location inference was not required as the data was reliably located and verified.

Within the administrative divisions of Kerala, there was ambiguity regarding duplicate place names, resulting in assumptions being made during location inference for the social platforms. For instance, there is a district, taluk (administrative subdivision) and city all called Thiruvananthapuram. As a general rule, unless the administrative level was stated, it was assumed that people were always being specific about their mentioned location (i.e. Thiruvananthapuram city in the previous example). The final filtered and geolocated data counts are shown in Table 2.

2.6. Characterising flood impacts

When a request was made through Kerala Rescue, the requester had the option to select 'True' or 'False' concerning several specific needs they may have had. The options were for rescue, food, water, medicine, clothes, kitchen utensils and toiletries, with each having a follow-up section for more information. For this investigation, the food and kitchen utensils columns were combined due to their similarity, as were the medicine and toiletries columns. To compare the specific needs of those using Kerala Rescue to the other platforms, all of the location inferred Telegram messages and Twitter posts (tweets) were manually categorised. This categorisation removed messages that were not directly requesting help (i.e. messages providing information/offering help), before categorising the remaining data into the five groups: rescue, food, water, medicine, and clothes. As Kerala Rescue users had the option to select multiple requests (i.e. food and medicine), the categorisation of tweets/telegrams was not limited to single classes. The data tended to be relatively easy to categorise, so a single human coder was employed to perform this task. To ensure validity, a sample of 50 labelled messages was independently categorised by 3 human coders, finding an almost perfect agreement (Fleiss's Kappa: 0.897, Krippendorff's Alpha: 0.865). This was seen as good evidence that the manual coding process was robust.

[6] https://kerala.gov.in/std-codes.

Table 2
Data volume before and after filtering from the four separate platforms.

Platform	Type	Initial volume	Filtered and geolocated volume
Telegram	Messages and Images	14,898	962
Twitter	Tweets	1,363,659	6936
Kerala Rescue	Requests	45,322	39,950
Rebuild Kerala	Surveyed Houses	330,578	330,578

3. Results

Fig. 1 shows a time series for the volume of filtered and located messages from Telegram, Twitter, and Kerala Rescue binned into 12-h time intervals from 1st August until 23rd August. Kerala Rescue was not created until 12th August so that platform only has data after that date. The data from Rebuild Kerala was not timestamped so is not plotted. There is a large difference in the volume of records from each platform, with Kerala Rescue having by far the highest number of relevant posts. All time series follow a similar trend, with a sharp increase in messages/requests from 15th August, peaking around 18th August, before decreasing again. This is consistent with the general consensus that the rainfall peaked in severity around 16th August [3]. Note that the 12-h period shows apparent decreases in activity during night time (2400–1200), seen most sharply on 17th August.

Fig. 2 shows filtered and located posts from the four data sources as a choropleth map (heatmap) of Kerala. Data is grouped at both district and taluk level to capture administrative units relevant to the region. To reflect the uncertainty in location inference at the scale of taluks, a moving average filter was used to assign to each taluk the average value across itself and its neighbouring taluks within the same district. This was done for all four data sources. As there was only uncertainty between place names within individual districts and not between multiple districts (as discussed in section 2.5), this filter was applied to each district separately. Data from the three timestamped platforms has been restricted to the main flood period between 12th - 23rd August. Since the counts between platforms vary considerably, values for each map were linearly scaled into the range [0,1] to create the color scale. Qualitatively, there are clear spatial similarities between maps, with two main clusters of activity seen in the taluk-level comparison - the first around Kochi, and the second between Alappuzha and Kollam. There is a visual agreement with these hotspots and the flooded areas identified using satellite imagery by Tiwari et al. [20]. The activity on Telegram and Twitter is more evenly distributed throughout Kerala, whereas on Kerala Rescue and Rebuild Kerala the activity hotspots are more tightly focused.

Fig. 3 quantitatively summarises the relationships between the variables plotted on the district and taluk maps in Fig. 2. Fig. 3 is a logarithmic scatter plot, with each subplot comparing the data from the corresponding column and row. Statistical analysis shows that all pairs of data sources have strong positive correlations at both taluk scale (Pearson's $r > 0.69$, $p < 0.001$) and district scale (Pearson's $r > 0.76$, $p < 0.001$). District-level correlations are stronger than taluk-level correlations. However, this was anticipated as location inference is more reliable at the larger district scale, where small errors are less important and the data less noisy. To check whether these results are explained as an artefactual consequence of population, with highly populated areas producing a stronger 'flood signal'

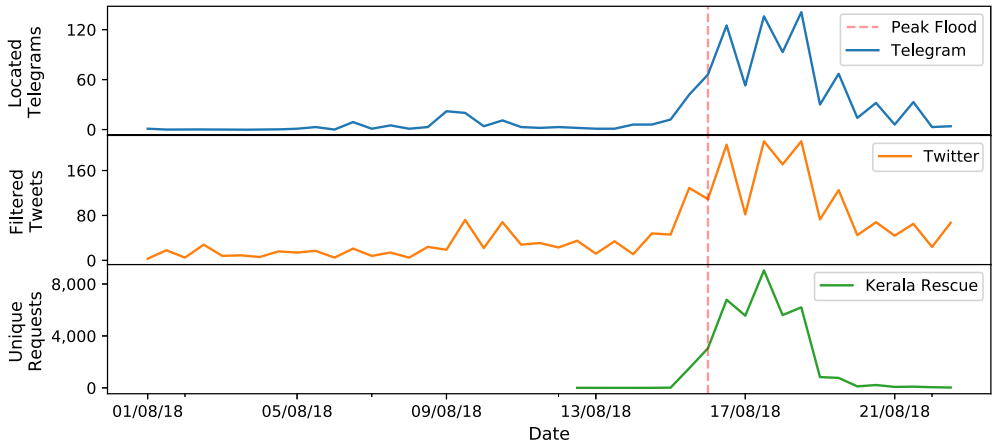


Fig. 1. Time series comparison of filtered social data sources, split into 12-h intervals.

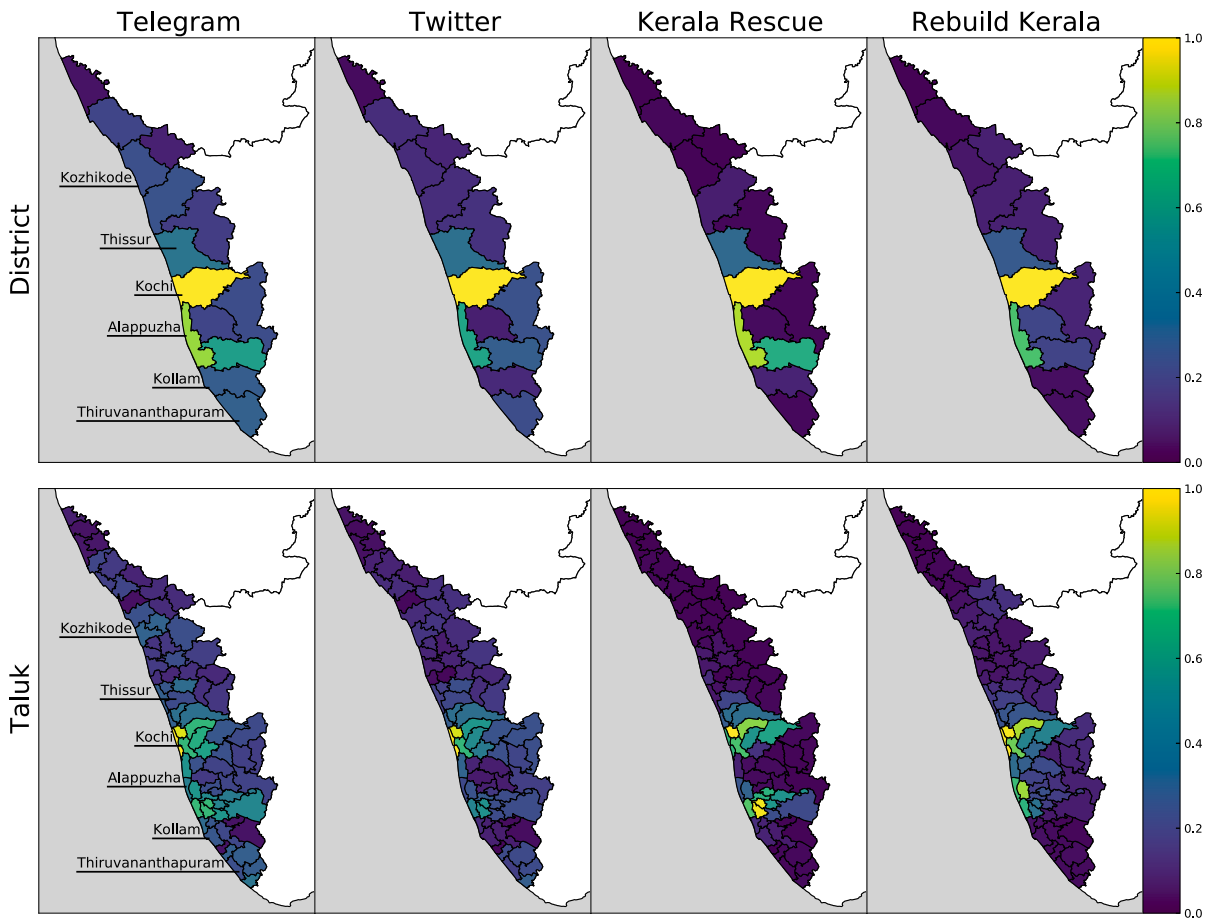


Fig. 2. Map of Kerala comparing the spatial distribution of data from the four platforms. Top: District-scale distribution. Bottom: Smoothed Taluk-scale distribution.

due to larger user populations, we calculated correlations with population size. Appendix A.2 shows that observation counts for both taluks (Fig. 2) and districts (Fig. 3) are uncorrelated with population ($p > 0.24$).

Fig. 4 shows time series for help requests communicated on the timestamped platforms, alongside time-aggregated distributions of request types. Despite the distribution of needs on the platforms being different, the timing of requests across platforms is similar. For instance, requests for rescue peak before requests for provisions on all platforms, with the exception of medicine requests through Twitter, where they peak on the same day. From the bar charts, we can see that excluding rescue requests, the distribution of request types on Twitter and Telegram is similar. On all platforms, clothing is the least requested, and excluding Telegram, rescue is the most requested. It is worth mentioning that both Telegram and Twitter contained two additional categories, namely, requests for volunteers and links for financial aid. As these categories were both lower in volume than the others, and were not present in Kerala Rescue, they have been omitted from this analysis. Rebuild Kerala has also been excluded as it does not provide additional impact information beyond property damage.

Table 3 shows a qualitative assessment of the strengths and weaknesses of each platform for measuring flood impacts. It should be noted that these platforms were not used independently of each other, for instance, messages from the same people were found on Telegram, Twitter and Kerala Rescue. Additionally, links to Kerala Rescue were widely distributed on Telegram and Twitter. Furthermore, these platforms are not the only platforms that were used to report impacts of the Kerala 2018 floods, with others such as WhatsApp being widely used.

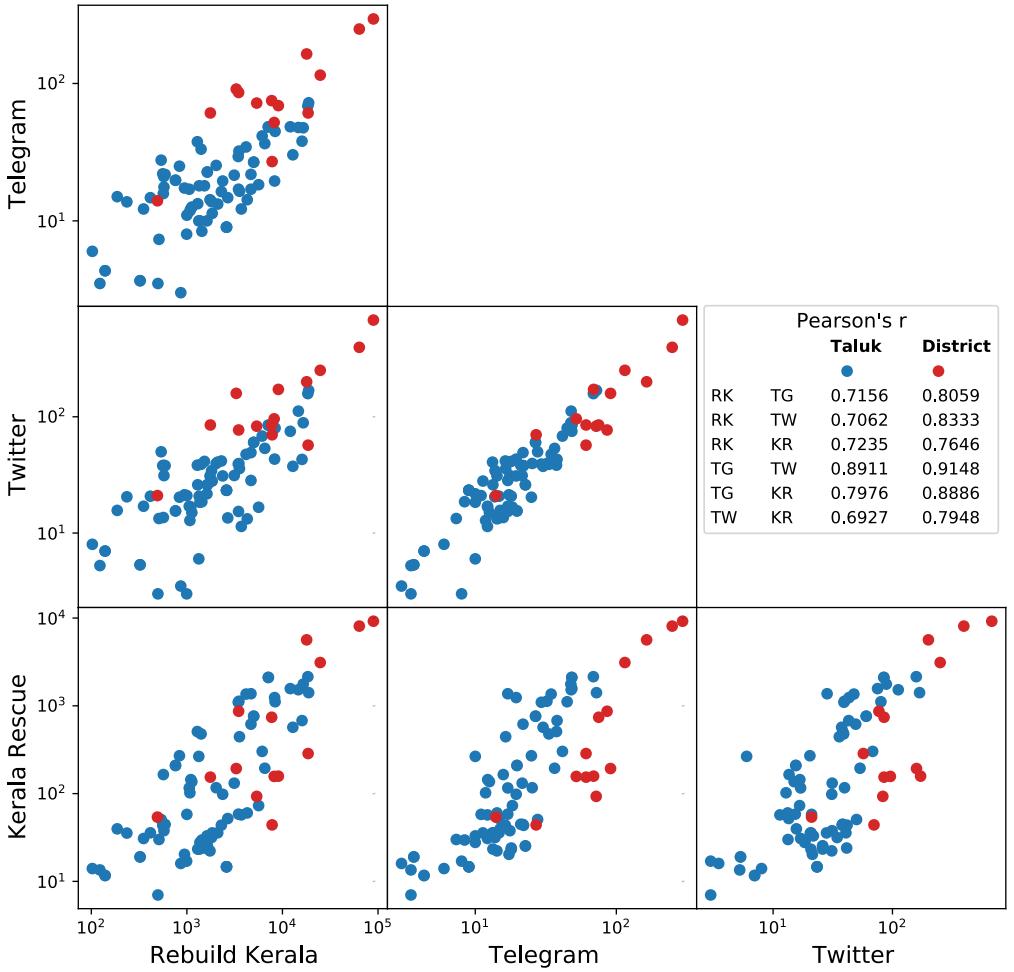


Fig. 3. Log scatter plot, comparing the Taluk (blue) and District (red) counts for flood observations from each data source: Rebuild Kerala (RK), Kerala Rescue (KR), Telegram (TG), Twitter (TW). The legend shows the Pearson's r value between the platforms, with all values having $P < 0.001$. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4. Discussion

In this paper, we have compared social media data from two platforms (Twitter and Telegram), as well as one citizen-produced web application (Kerala Rescue) and a governmental disaster relief survey (Rebuild Kerala). The aim was to map and characterise the social impacts of the major flood event in Kerala in 2018. Results show that there is good agreement between the outputs from social sensing using Telegram/Twitter, the semi-formal relief requests database of Kerala Rescue, and the formal property damage assessment performed by Rebuild Kerala. Data from Twitter and Telegram also showed similarity between the types of impact reported, to the types of help request collated by Kerala Rescue. These findings build confidence that observation using unsolicited social media data can be an effective way to understand the effects of flooding. While Kerala Rescue was a platform rapidly created by volunteers specifically to coordinate requests for help during the flood, and might therefore be expected to perform well in this scenario, it is noteworthy that data derived from the general-use social media platforms Twitter and Telegram also show strong correlations with the structured observations provided by Rebuild Kerala. This study, therefore, adds to the evidence base (e.g. Arthur et al. [28,29] demonstrating the potential for informal social media data to assist with flood response and impact assessment. Here the use of Telegram data offers additional novelty, since (to our knowledge) this platform has not previously been used for this purpose.

The Kerala Rescue dataset contains a large amount of highly relevant data about flood impacts, relative to the smaller (post-

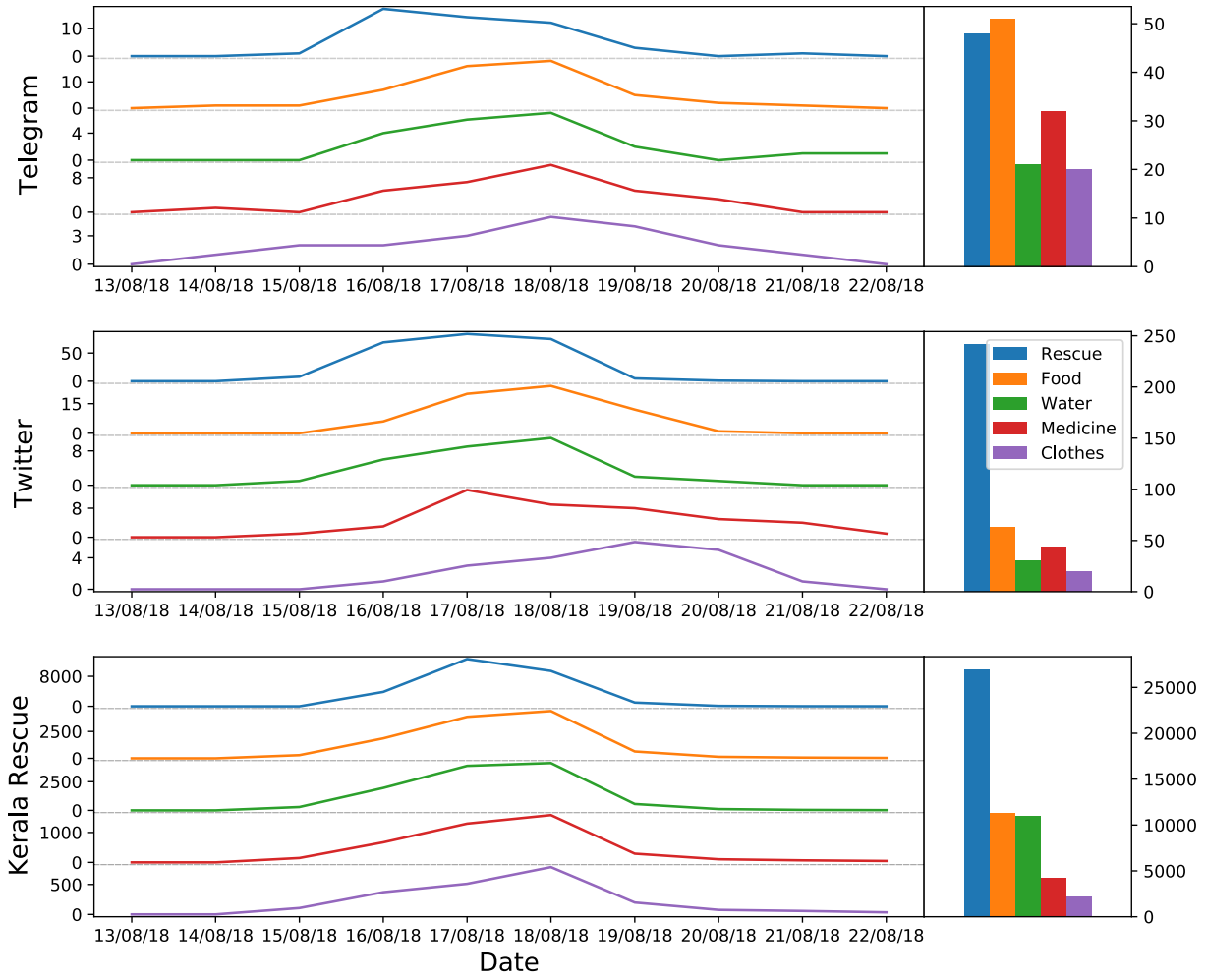


Fig. 4. Temporal comparison of the 5 largest Kerala Rescue request types, compared to the same request types found within Telegram messages and tweets.

filtering) volumes and greater ambiguity of data available from Twitter and Telegram. However, Kerala Rescue was a one-off platform, created in the early days of the flood by volunteer software developers. This ad hoc and spontaneous aspect to Kerala Rescue may add to its authenticity as a source of on-the-ground observations, but also brings uncertainty in that we do not know how widely it was adopted and by whom. By contrast, Twitter and Telegram both have established user populations in Kerala with an observable history of activity on those platforms. This means that while Kerala Rescue is very useful for situational awareness and community organisation during the event, it does not offer an effective "baseline" of activity against which observations can be situated. For example, it is hard to understand why more requests were made on Kerala Rescue in the Kollam region relative to activity on Twitter and Telegram. This could be due to Kerala Rescue genuinely capturing more impacts in that region, or could just reflect greater adoption of the application in that area; without a historic baseline of platform usage in that area, it is hard to determine. For citizen-generated apps such as Kerala Rescue there is no guarantee that the skilled volunteer workforce that created the app will be replicated in future floods or similar extreme weather events. However, Kerala Rescue shows that such applications can be effective in coordinating relief efforts during the event and also as a data source for *post hoc* analysis.

While the social media platforms Twitter and Telegram generate fewer observations of flood impacts in the present case study, they offer a number of contrasting advantages. As referenced above, their established user populations, geographic coverage and generality permit the analysis of the 2018 Kerala flood event in a wider context. This might allow the 2018 Kerala floods to be assessed against other flood or weather events at different times or places, or against other types of social disruption, with a reasonably consistent underlying study population. The use of a temporal baseline allows the severity of an event to be estimated based on the level of social media activity relative to the long-term expected level; cf. identification of storm events using a percentile threshold for social media activity [34]. The establishment of social media as a routine form of communication avoids any special effort to set up a platform when an event occurs since events are observed within the continual flow of communications. Another interesting feature of Twitter and Telegram is that they are relatively unstructured in the ways they support communication between users. They enable conversational interactions between pairs or groups of users, in which a greater depth of understanding can be achieved about an unfolding event.

Table 3
Qualitative comparison of data types.

Platform	Format	Strengths	Weaknesses
Telegram	Social Media: Messaging	- Data is available for both real - time and retrospective collection - Conversations are often regional i.e. from those directly impacted - Low volume of unsolicited content/"fake news" in messages as admins often regulate chats	- Group chats have a trade-off between volume and quality - Conversational nature of messages makes it hard to extract robust measurements - Users often delete their messages, leaving large gaps in conversations/data
Twitter	Social Media Microblogging	- Data can be collected for free in real - time through the Twitter API - Original posts are mostly independent providing separate observations - High volume of raw data	- Relevant data is mixed with larger volumes of irrelevant data, requiring multiple filtering steps - Expensive to obtain tweets retrospectively without a research account - The retweet function and trending topics can distort the signal, preventing balanced impact analysis
Kerala Rescue	Rescue Request Platform	- Data was publicly accessible - Requests were highly structured - High relevance as the platform was created for the 2018 flood - High volume	- Created during the flood so difficult to determine the extent of adoption - One-off platform so cannot be relied on for future events
Rebuild Kerala	Government Housing Damage Data	- Highly structured - Data is reliable as the surveys were verified by government volunteers	- Surveys were manually collected post-event over the following months, expending significant human effort - The platform only measures one impact type (property damage)

This more natural form of discussion may provide insights that more constrained platforms such as Kerala Rescue do not. The converse of this lack of structured communication is that a greater effort is required to derive robust observations. It is not feasible to manually analyse large datasets that may contain hundreds of thousands of messages, while natural language processing can be difficult to apply successfully for short-form messages such as social media posts.

In summary, all of the platforms studied here offer distinct opportunities but also unique drawbacks to resilience planners and flood responders. These are summarised in Table 3 above. Twitter and Telegram allow the potential for routine monitoring and can provide highly relevant impact information. Kerala Rescue was a crucial platform during the flood, generating flows of highly structured and highly relevant data. However, it was created out of necessity and the next flood, or floods in other areas, could easily use a different platform or coordinate through channels like Telegram or WhatsApp. Finally, Rebuild Kerala is probably the most reliable and systematic data collection, but is retrospective, much slower, and focuses on only a single impact type.

One important consideration for any systematic usage of social media or other citizen-generated data sources for flood monitoring and emergency response is that of sample bias. While Internet penetration and smartphone usage are rapidly increasing in Kerala, India and most other parts of the world, it is still more frequent amongst younger, more affluent and urbanised populations. Reliance on such data sources could result in an unintended bias towards their user populations and away from those without access to digital resources. Careful management of data and appropriate consideration of bias might help to alleviate some of the harmful effects of such "digital divides", but should be considered from the outset of any project to operationalise these methodologies.

5. Conclusion

This paper shows that social sensing via popular social media platforms like Twitter and Telegram, or bespoke platforms like Kerala Rescue, can be a useful way to gather validation data and study flood resilience in India. While social media (typically Twitter) has been studied previously to show that it can generate useful insights around natural disasters, here we have focused on the comparison of different platforms, including novel sources such as Telegram and KeralaRescue, and validation against qualified governmental sources. We have shown that social media sources allow the rapid collection of data from people on the ground which is difficult to gather in other ways. A combination of channels could provide excellent coverage across a range of flood impacts - e.g. routine monitoring of Twitter and Telegram for situational awareness and to pick up early impacts. Kerala Rescue or similar could then be used as a central place for the coordination of community action and relief. Post-disaster analysis of all data sources, including formal efforts like Rebuild Kerala, could then provide a detailed record of the range of impacts and enable the creation of better flood resilience planning and infrastructure. As the global penetration of the internet and social media increases, we believe this study's methodology is not geographically bound to Kerala and can be of real value in other flood-prone areas. Additionally, as the platforms are not specifically tailored to flood events, this study's approach could be applied to other extreme weather hazards.

Funding

The authors acknowledge funding from the UK Government's Newton Fund via the Weather and Climate Science for Services Partnership India (WCSSP India). This work was conducted as part of the FRESCO project funded by WCSSP India. H.T.P.W. also acknowledges funding from UK Natural Environment Research Council (NE/P017436/1). J.C.Y. and M.S. are funded by a PhD studentship from the UK Engineering and Physical Sciences Research Council. No funding bodies had any influence over the content of this report.

Data availability

The Twitter data used in this word was purchased using the official Twitter PowerTrack API (https://developer.twitter.com/en/docs/twitter-api/enterprise/powertrack-api/overview (accessed on 15 December 2020)). The Telegram data was collected from the Telegram desktop application (https://telegram.org/blog/export-and-more (accessed on 13 October 2020)). The Kerala Rescue data was initially sourced from the RebuildEarth Slack channel (https://rebuildearth.slack.com/(accessed on 15 October 2020)). The Rebuild Kerala data was collected from the Rebuild Kerala Database site (https://rebuild.lsgkerala.gov.in/rebuild2018/(accessed on 6 November 2020)).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

J.C.Y. conducted the formal analysis, investigation, methodology, visualisation and original draft writing. R.A. and H.T.P.W. supervised the research. R.A., M.S., H.T.P.W. reviewed and edited the final version. All authors have read and agreed to the published version of the manuscript.

A. Appendix

A.1.

Keywords for filtering tweets. Tweets were omitted from analysis if they did not contain any of the words in either list.

Flood Terms	landslide, flood, rain, storm, water, boat, drown, dam, disaster, weather
Kerala Rescue Derived Terms	stuck, please, help, stranded, rescue, trapped, family, emergency, families, terrace, urgent, location, kid, pregnant, homeless, baby, old, elderly, immediate, evacuation, child, father, mother, aunt, uncle, brother, sister, friend, evacuate, pray, death, relief

A.2.

Correlation between district and taluk population and the flood signal from the four platforms (Figs. 2 and 3). The weak correlations are not statistically significant, showing that the levels of observed activity related to floods is not explained by underlying variation in population density across taluks/districts.

	Pearson's R		P Value	
	Taluk Population	District Population	Taluk Population	District Population
Telegram	−0.0145	0.2244	0.9004	0.4406
Twitter	−0.0104	0.2223	0.9285	0.4449
Kerala Rescue	−0.1350	0.0861	0.2417	0.7697
Rebuild Kerala	−0.0389	0.1648	0.7371	0.5734

References

[1] C.B. Field, V. Barros, T.F. Stocker, Q. Dahe (Eds.), Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, 2012.
[2] Z.W. Kundzewicz, S. Kanae, S.I. Seneviratne, J. Handmer, N. Nicholls, P. Peduzzi, R. Mechler, L.M. Bouwer, N. Arnell, K. Mach, R. Muir-Wood, G.R. Brakenridge, W. Kron, G. Benito, Y. Honda, K. Takahashi, B. Sherstyukov, Flood risk and climate change: global and regional perspectives, Hydrol. Sci. J. 59 (1) (2014) 1–28.
[3] V. Mishra, S. Aaadhar, H. Shah, R. Kumar, D.R. Pattanaik, A.D. Tiwari, The Kerala flood of 2018: combined impact of extreme rainfall and reservoir storage, in: Preprint, Hydrometeorology/Modelling approaches, 2018.
[4] N.P. Nagendra, G. Narayanamurthy, R. Moser, Management of humanitarian relief operations using satellite big data analytics: the case of Kerala floods, Ann. Oper. Res. (2020).
[5] Government of Kerala, Post-disaster Needs Assessment - Kerala, Technical report, United Nations Development Programme, 2018.
[6] G. Arunkumar, R. Chandni, D.T. Mourya, S.K. Singh, R. Sadanandan, P. Sudan, B. Bhargava, Outbreak of nipah virus disease in Kerala, India, 2018, SSRN Electron. J. (2018).
[7] S. James, B. Sathian, E.V. Teijlingen, M. Asim, Outbreak of leptospirosis in Kerala, Nepal J. Epidemiol. 8 (4) (2018) 745–747.
[8] L. Hao, A. R, C. van Westen, S.K. S, T.R. Martha, P. Jaiswal, B.G. McAdoo, Constructing a complete landslide inventory dataset for the 2018 monsoon disaster in Kerala, India, for land use change analysis, Earth Syst. Sci. Data 12 (4) (2020) 2899–2918.
[9] H. Ali, P. Modi, V. Mishra, Increased flood risk in Indian sub-continent under the warming climate, Weather Clim. Extr. 25 (2019) 100212.
[10] S. Mukherjee, S. Aadhar, D. Stone, V. Mishra, Increase in extreme precipitation events under anthropogenic warming in India, Weather Clim. Extr. 20 (2018) 45–53.
[11] United Nations, World Urbanization Prospects: the 2018 Revision, 2018. https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html.
[12] K. Gupta, Challenges in developing urban flood resilience in India, Phil. Trans. Math. Phys. Eng. Sci. 378 (2168) (2020) 20190211.
[13] K.M.R. Hunt, A. Menon, The 2018 Kerala floods: a climate change perspective, Clim. Dynam. 54 (3–4) (2020) 2433–2446.
[14] S. Paul, S. Ghosh, K. Rajendran, R. Murtugudde, Moisture supply from the western ghats forests to water deficit east coast of India, Geophys. Res. Lett. 45 (9) (2018) 4337–4344.
[15] T.V. Padma, Mining and dams exacerbated devastating Kerala floods, Nature 561 (7721) (2018) 13–14.
[16] K. Ray, P. Pandey, C. Pandey, A.P. Dimri, K. Kishore, On the recent floods in India, Curr. Sci. 117 (2019).
[17] V. Avashia, A. Garg, Implications of land use transitions and climate change on local flooding in urban areas: an assessment of 42 Indian cities, Land Use Pol. 95 (2020) 104571.
[18] K. Gupta, V. Nikam, Technological and innovative measures to improve flood disaster recovery following Mumbai 2005 mega-flood, in: R. Shaw (Ed.), *Disaster Recovery: Used or Misused Development Opportunity*, Disaster Risk Reduction, Springer Japan, Tokyo, 2014, pp. 287–297.
[19] P. Lal, A. Prakash, A. Kumar, P.K. Srivastava, P. Saikia, A.C. Pandey, P. Srivastava, M.L. Khan, Evaluating the 2018 extreme flood hazard events in Kerala, India, Rem. Sens. Lett. 11 (5) (2020) 436–445.
[20] V. Tiwari, V. Kumar, M.A. Matin, A. Thapa, W.L. Ellenburg, N. Gupta, S. Thapa, Flood inundation mapping- Kerala 2018; Harnessing the power of SAR, automatic threshold detection method and Google Earth Engine, PLoS One 15 (8) (2020), e0237324.
[21] N. Goyal, Disaster governance and community resilience: the law and the role of sdmas, Int. J. Dis. Risk Manag. 1 (2) (2019) 61–75.
[22] N. Bhuvana, I. Arul Aram, Facebook and Whatsapp as disaster management tools during the Chennai (India) floods of 2015, Int. J. Disaster Risk Reduc. 39 (2019) 101135.
[23] M.R. Nair, G.R. Ramya, P.B. Sivakumar, Usage and analysis of Twitter during 2015 Chennai flood towards disaster management, Procedia Comput. Sci. 115 (2017) 350–358.
[24] R. Campbell, D. Beardsley, T. Sezin, Impact-based Forecasting and Warning: Weather Ready Nations, 2018. https://public.wmo.int/en/resources/bulletin/impact-based-forecasting-and-warning-weather-ready-nations.
[25] A.L. Taylor, T. Kox, D. Johnston, Communicating high impact weather: improving warnings and decision making processes, Int. J. Disaster Risk Reduc. 30 (2018) 1–4.
[26] S. Vieweg, C. Castillo, M. Imran, Integrating social media communications into the rapid assessment of sudden onset disasters, in: L.M. Aiello, D. McFarland (Eds.), *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2014, pp. 444–461.
[27] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, *WWW 2010*, Raleigh, North Carolina, USA, *April 26-30, 2010*, 2010, p. 10.
[28] R. Arthur, C.A. Boulton, H. Shotton, H.T.P. Williams, Social sensing of floods in the UK, PLoS One 13 (1) (2018), e0189327.
[29] FloodTags (n.d.), About us – FloodTags. https://www.floodtags.com/about-us/.
[30] J.C. Young, R. Arthur, M. Spruce, H.T.P. Williams, Social sensing of heatwaves, Sensors 21 (11) (2021) 3717.
[31] Y. Kryvasheyeu, H. Chen, E. Moro, P.V. Hentenryck, M. Cebrian, Performance of social network sensors during hurricane sandy, PLoS One 10 (2) (2015), e0117288.
[32] D. Wu, Y. Cui, Disaster early warning and damage assessment analysis using social media data and geo-location information, Decis. Support Syst. 111 (2018) 48–59.
[33] C. Boulton, H. Shotton, H. Williams, Using social media to detect and locate wildfires, in: Tenth International AAAI Conference on Web and Social Media, 2016.
[34] M. Spruce, R. Arthur, H.T.P. Williams, Using social media to measure impacts of named storm events in the United Kingdom and Ireland, Meteorol. Appl. 27 (1) (2020) e1887. Number: 1 _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.1887.
[35] Statista, Twitter: Most Users by Country, 2021. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.
[36] D. Karmegam, B. Mappillairaju, Spatio-temporal distribution of negative emotions on Twitter during floods in Chennai, India, in 2015: a post hoc analysis, Int. J. Health Geogr. 19 (1) (2020) 19.
[37] F. Cecinati, T. Matthews, S. Natarajan, N. McCullen, D. Coley, Mining social media to identify heat waves, Int. J. Environ. Res. Publ. Health 16 (5) (2019) 762.
[38] P. Mishra, R. Rajnish, P. Kumar, Sentiment analysis of Twitter data: case study on digital India, in: 2016 International Conference on Information Technology (InCITe) - the Next Generation IT Summit on the Theme - Internet of Things: Connect Your Worlds, 2016, pp. 148–153.
[39] TechCrunch, Telegram Surpasses 1 Billion Downloads, 2021.
[40] Telegram, The Evolution of Telegram, 2021.
[41] A. Redondo, J.M. Haut, M.E. Paoletti, X. Tao, J. Plaza, A. Plaza, Analysis of remotely sensed images through social media, IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 14 (2021) 3026–3039.
[42] S.E. Ahmady, O. Uchida, Telegram-based chatbot application for foreign people in Japan to share disaster-related information in real-time, in: 2020 5th International Conference on Computer and Communication Systems (ICCCS), 2020, pp. 177–181.
[43] A. Dargahi Nobari, M.H.K.M. Sarraf, M. Neshati, F. Erfanian Daneshvar, Characteristics of viral messages on Telegram; the world's largest hybrid public and private messenger, Expert Syst. Appl. 168 (2021) 114303.
[44] J.M. Cecilia, J.-C. Cano, C.T. Calafate, P. Manzoni, C. Periñán-Pascual, F. Arcas-Túnez, A. Muñoz-Ortega, WATERSensing: a smart warning system for natural disasters in Spain, IEEE Consum. Electr. Mag. 10 (6) (2021) 89–96.
[45] S.Z. Razavi, M. Rahbari, Understanding reactions to natural disasters: a text mining approach to analyze social media content, in: 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), 2020, pp. 1–7.
[46] RebuildKerala (n.d.), Rebuild Kerala initiative- RKI. https://rebuild.kerala.gov.in/en/about.
[47] IEEE (n.d.). The story behind Keralarescue.in – IEEE Kerala Section.

[48] R. Varghese, Y.T. A, Role of Social Media during Kerala Floods 2018, Library Philosophy and Practice (e-journal), 2019.
[49] Telegram, Location-Based Chats, Adding Contacts without Phone Numbers and More, 2019. https://telegram.org/blog/contacts-local-groups.
[50] M. Lee, Pytesseract: Python-Tesseract Is a python Wrapper for Google's Tesseract-OCR, 2020.
[51] M.M. Danilak, Langdetect: Language Detection Library Ported from Google's Language-Detection, 2020.
[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[53] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, M. Mühlhäuser, A multi-indicator approach for Geolocalization of tweets, in: Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013, 2013.
[54] GLOBE, Global Administrative Areas, 2012.
[55] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: a nucleus for a web of open data, in: Proceedings of the 6th International the Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, ISWC'07/ASWC'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 722–735.
[56] GeoNames (n.d.), The GeoNames geographical database covers all countries and contains over eleven million placenames that are available for download free of charge. https://www.geonames.org/.