Environmental Modelling and Software 133 (2020) 104828

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: http://www.elsevier.com/locate/envsoft

Position Paper

A national scale big data analytics pipeline to assess the potential impacts of flooding on critical infrastructures and communities

N. Donratanapat [a], S. Samadi [a,*], J.M. Vidal [b], S. Sadeghi Tabas [c]

[a] Department of Agricultural Sciences, Clemson University, USA
[b] Department of Computer Science and Engineering, University of South Carolina, USA
[c] Department of Civil and Environmental Engineering, Clemson University, USA

ARTICLE INFO

Keywords:
Internet of things
Flood analytics information system
Machine learning
Big Data
Flood situational awareness
The Carolinas.

ABSTRACT

With the rapid development of the Internet of Things (IoT) and Big Data infrastructure, crowdsourcing techniques have emerged to facilitate data processing and problem solving particularly for flood emergences purposes. A Flood Analytics Information System (FAIS) has been developed as a Python Web application to gather Big Data from multiple servers and analyze flooding impacts during historical and real-time events. The application is smartly designed to integrate crowd intelligence, machine learning (ML), and natural language processing of tweets to provide flood warning with the aim to improve situational awareness for flood risk management. FAIS, a national scale prototype, combines flood peak rates and river level information with geotagged tweets to identify a dynamic set of at-risk locations to flooding. The prototype was successfully tested in real-time during Hurricane Dorian flooding as well as for historical event (Hurricanes Florence) across the Carolinas, USA where the storm made extensive disruption to infrastructure and communities.

1. Introduction

The south and southeast United States (US) are subjected to a series of intense storms throughout the year as well as deadly Atlantic hurricane events during hurricane season (June–November). These events can happen in quick succession (~2 weeks apart) and produce catastrophic flooding in wide geographic areas (~1000 km swath) and within short time-spans (less than a 48-h period). As a consequence of these successive events, many lives were lost, and numerous critical infrastructure and communities were vastly disrupted. To reduce the risk of damages, accurate and real time flood assessment is critical for emergency management and to improve two-way communication and understanding of potential impacts.

At present, National Weather Service (NWS) and National Hurricane Center (NHC) provide river and storm path forecasts and issue early warning system for potential areas of flooding; however, it is acknowledged that these forecasts are large scale and have less skills with respect to detecting localized floods and identifying specific areas at risk of flooding (Samaniego et al., 2017; Adams and Dymond. 2019). Flooding in the south and southeast regions are often highly localized and intense which cause inundation in low lying roads and poor drainage areas and the level of individual properties (e.g., Philips et al., 2018). Furthermore, providing geographically targeted early flood warnings in time is hampered by a lack of data and real-time information for stakeholders and residents to take protective actions for themselves, their property, and livestock. Improved data collection and real-time assessment of at-risk locations allow more efficient mutual aid in the operational theater for warnings and evacuations, and more effective search and rescue plans while enabling automatic dispatching of relief resources and evacuation plans.

Further development, validation, and implementation of viable and accurate flood warning systems requires a step change in the methodologies used for data collection and analysis. With the rapid development of earth observation technology and ground-based monitoring systems that produce time-lapse videos and images, high spatial and temporal Big Data have been recently tapped into flood early warning assessment (e.g., Barker and Macleod, 2019). Although, these images require innovative enabling technologies to improve the integration, retrieval, analysis and presentation of large amounts of information (Grolinger et al., 2013; Nativi et al., 2013). Smart technologies such as Internet of Things (IoTs), image processing, and machine learning (ML) can provide the "intelligence" to analyze real-time data and alleviate information overload for flood early warning system (e.g., Rao et al., 2017).

* Corresponding author.
E-mail address: samadi@clemson.edu (S. Samadi).

https://doi.org/10.1016/j.envsoft.2020.104828
Accepted 12 August 2020
Available online 26 August 2020

IoT is one of the fastest developing fields in the history of computing, with an estimated 50 billion devices by the end of 2020 (Al-Garadi et al., 2018) that can integrate billions of smart devices to communicate with one another with minimal human intervention. The connectivity in IoTs and high-speed data transfer capabilities can be used to implement real time image processing and ML analytics system for flood risk studies. Both image processing and ML algorithms are powerful methods of data exploration for real time monitoring and learning about 'normal' and 'abnormal' condition of a watershed system. Recently, image processing and ML algorithms have been demonstrated to label time laps camera imagery, crowdsourcing, tabular data, and user generated texts and photos to extract road flooding inundation extend and depth (de Albuquerque et al., 2015; Starkey et al., 2017; Feng and Sester, 2018; Barker and Macleod, 2019; Feng and Sester 2018, 2018; Erfani and Samadi, 2019). In addition, analysis of social geodata during floods can provide actionable intelligence to assist first responders to identify at-risk communities.

Stakeholders need place-based, geotagged, and crowd-sourced information about flooding being analyzed rapidly in real-time. Accessibility to voluntarily generated and often publicly published content on social networking and social media provides a strong draw for disaster related research. Crowdsourced social media data, particularly Twitter, is increasingly used to improve situational awareness and two-way communication during hurricane and flood events (e.g., Kryvasheyeu et al., 2016; Barker and Macleod, 2019). Social crowdsourcing data can help with the identification of flood extend especially during pluvial and fluvial flooding in urban settings (e.g., Smith et al., 2015; Eilander et al., 2016; Arthur et al., 2018). Developing a pipeline to gather the data and identify tweets relevant to flooding proved to be useful to assess real-time flooding impacts and damage in Sao Paulo- Brazil (De Assis et al., 2016), Jakarta-Indonesia (Eilander et al., 2016), the River Elbe-Germany (Herfort et al., 2014), and across Great Britain (Barker and Macleod, 2019).

As needs for real time flood impact assessment increase, stakeholders are facing fragmented data environments and warehouses with multiple technologies—often on multiple web services. There is a need to automate Big Data and crowd sourced information collection in real-time and create a map-based dashboard to better determine at-risk locations and flood situations. Indeed, with the new advancement in technologies, there is an opportunity to gather and combine social media data with ground-based observations and imagery and translate this information into a web-based application to monitor and assess flooding hazards and to communicate this information with citizens in real time.

This paper introduces Flood Analytics Information System (FAIS) as a data engineering and analytics pipeline, based on real-time flood warnings and river level information, natural language processing of tweets, and river and traffic web cameras imagery. FAIS allows the user to directly download flood related data from USGS and visualize the data in real time. The outcome of the river measurement, imagery, and tabular data is displayed in a web based remote dashboard and the information can be plotted in real-time. A Twitter Application Programming Interface (API) and a bot software were developed and incorporated into the prototype as part of the real time crowd intelligence for Twitter data gathering. The developed Twitter bot allows user to monitor every tweet being tweeted and can automate all or part of Twitter activities. Indeed, our developed pipeline allows the user to query tweets from Twitter by a specific user and/or keyword using both Search and Streaming APIs. A Search API gathers historical geotag data while a Stream API monitors real-time geotagged tweets with shortlisting at-risk areas based on provided keywords. FAIS system can be used equally efficiently by stakeholders as a pervasive early warning system to take smart action such as warning and evacuation, deployment of emergency assets, search and rescue, and planning. The prototype was tested for Hurricanes Florence and Dorian driven flood situational assessment across the Carolinas.

This paper is organized as follows. In Section 2, the research questions and motivation of this research work are explained. The procedures, algorithms, and the functionality of FAIS application are introduced and discussed in Section 3. Section 4 discusses the implementation and case studies. Conclusions and future works and limitation of the prototype are provided, respectively in Sections 5 and 6.

2. Research Questions and Motivation

Developing FAIS required addressing three research questions which are:

1. How to programmatically and automatically identify areas at-risk of flooding based on crowdsourced data, real-time flood peak rates, and river level information? The first research question was whether we could compute programmatically the areas of at risk to flooding using various data sources. Vieweg et al. (2014) and Barker and Macleod (2019) indicated that such an automatic task is difficult to implement in real time due to the volume of data to identify relevant information for decision-making process. However, Streaming APIs proved to be useful for prioritizing a list of at-risk locations (Barker and Macleod, 2019).
2. How to spatially display the retrieved data and implement this information for alert and warning system? The second research question was to investigate the viability of displaying the retrieved data in a timely and continuous way. Previous studies showed how cross-referencing tweets can be used for prioritizing at-risk locations to flooding (Middleton et al., 2014; Barker and Macleod, 2019) as well as arranging location-based queries during floods, using georeferencing/geotag tweets (Laylavi et al., 2016).
3. How to seamlessly retrieve data from various sources and how to use this information for making actionable decision? The third research question investigated the viability of automated retrieval of data and images from ground-based monitoring gauges as well as live traffic and river webcams data. Previous studies highlighted that APIs are particularly helpful in gathering various Big datasets (text, tabular, and images) and could filter social media messages during flooding events (Spielhofer et al., 2016; Barker and Macleod, 2019). Internationally, there are an increasing number of data sources with a data service APIs that can be integrated with any software application.

Our aim was to develop and test a pipeline integrated with historical and real-time information based on these three research questions and visualize at risk locations during a series of flooding events across the Carolinas. The authors also discussed the design of the prototype with federal and state stakeholders to more proficiently develop and implement the workflow. During several visits to SC Emergency Management Division (SCEMD) as well as virtual discussions with federal agencies such as USGS and Federal Emergency Management Agency (FEMA), we demonstrated the need for a national scale pipeline that: (i) combines historical and real-time river level information with crowdsourcing data, (ii) automates Big Data gathering and information collection in real-time, and (iii) creates a map-based dashboard to better determine at-risk locations and flood situations across the United States. These needs and discussion along with deficiencies in existing Big Data pipelines provided comprehensive roadmap tasks for performing this research.

3. Prototype Design and Development

FAIS is initially designed as a Python package targeting two sources of data i.e., USGS and Twitter. The package was then transferred to a web Python platform to collect the data during historical and real-time events and visualize impacted areas. The pipeline uses IoTs-APIs and machine learning for transmitting, processing, and loading Big Data through which the application gathers information from various data servers and replicates it to a data warehouses for use with crowd intelligence approaches. FAIS filters flood-relevant tweets using location-filtering and word embedding of tweets. User can stream or search for tweets using proper keywords for any region in US. FAIS provides both custom data and analytics-as-a-service offerings to help users gain insights about flood situation, data environment and start driving informed decisions. The prototype also performs flood frequency analysis (FFA) to assist engineers in designing safe structures. Below, we systematically describe a series of major design components and algorithms designed within the FAIS application.

3.1. Machine Learning and Image Processing Approaches

This study used Google Vision API to detect objects in time lapse images. Google Vision API uses image processing and machine learning approaches to detect and extract information about objects and entities in an image, across a broad group of categories. This tool encapsulates machine learning models in an API approach that allows developers to use computer vision technology for classifying images into thousands of categories and assign them sensible labels and scores. Vision API detects objects in the images using (i) multiple objects including the location of each object within the image; and (ii) fast, high-accuracy models to classify images or detect objects at the edge, and trigger real-time actions based on local data.

FAIS allows the user to use the Vision API directly or use AutoML Vision to train machine learning model for image annotation and label images. The application detects the objects in the image using google-cloud-vision as a Python package to deal with the API. We included Google cloud sdk along with gsutil tools in the FAIS algorithm to easily upload large dataset of images to a google bucket. The tool then creates a bucket in Google cloud storage and user can upload image folder from the local desktop to Google bucket. The API then utilizes machine learning tools to perform label detection on a request image and sends the result back to the FAIS application. The tool can detect individual objects and pieces of text and information within an image directly from the application, analyze images, and build custom models using the API to accommodate more flexibility for particular use case.

FAIS uses Label Detection to annotates an image with a label (or "tag") based on the image content and then name them. For example, a picture of a flooded road may produce a label of "flood", "road", or some other similar annotation. Label Detection determines broader category contexts in different ways—for example, an image can be labeled as "flood", "water", "river", "floodplain", etc. that cover broader categories of water resources objects. To create high-quality training datasets of annotated images, 100–200 or more flood occurrences (across all images) is required to train the Vision model and label the objects. The more occurrences of an object such as "flood" in time lapse images, the better the model trains and performs. After the user creates the labels, FAIS API calls to create an object detection dataset and populates the images and labels them in the JSON format. The labels constantly store on the MangoDB database and display on the presented image.

3.2. Flood Frequency Analysis

FFA is a technique used by hydrologists and engineers to predict flow values corresponding to specific return periods or probabilities along a river. The tool uses "dataRetrieval" (DeCicco et al., 2018) and "xts" (Jeffrey et al., 2020) libraries to retrieve annual peak flow rates for provided years and calculates statistical information such as mean, standard deviation and skewness which are further used to create frequency distribution graphs. The tool currently fits Gumbel distribution to the annual maximum flood data and plots frequency curves. These graphs are then used to estimate the design flow values corresponding to specific return periods which can be used for designing structures such as dams, bridges, culverts, levees, highways, sewage disposal plants, waterworks, and industrial buildings.

Gumbel is a proper distribution if: (i) the river system is less regulated with less significant reservoir operations, diversions or urbanization effects, (ii) flow data are homogeneous and independent (lack of fluctuations and long-term trends; Philips et al., 2018), (iii) peak flow data cover relatively long records (>10 years), and (iv) no major tributary exists whose inflow may affect the flood peak rates (e.g., Raynal and Salas, 1986). Gumbel distribution and the procedure with a return period T is given as,

$$X_T = \overline{X} + K.\sigma_x \tag{1}$$

where σ_x represents standard deviation of the sample time series. K denotes frequency factor which is formulated as $K = \frac{Y_T - \overline{\overline{Yn}}}{S_n}$, in which Y_T is reduced Variate, $Y_T = -\left[Ln.Ln.\left(\frac{T}{T-1}\right)\right]$. The values of $\overline{\overline{Yn}}$ and S_n are selected from Gumbel's Extreme Value Distribution that depends on the sample size (e.g., Raynal and Salas, 1986). It should be noted that the theoretical definition of return period is the inverse of the probability that an event will be exceeded in a given year. For example, a 10-year return period corresponds to a flood that an exceedance probability of 0.10 or a 10% chance that the flow will exceed in one year.

3.3. Development of Twitter APIs

During recent hurricane events, many citizens in the Carolinas used Twitter to share flood information such as local damage, road closure, and shelter information. The government agencies such as NWS, NHC, SC Department of Transportation (SCDOT), and USGS also used Twitter to provide updates about the storm path, environmental condition, damaged infrastructure, emergency situations, evacuation route, and resources in a continuous and timely manner during and after the event. A tweet can provide a variety of information, such as text, images, videos, audio, and additional links. In addition, there is also a significant amount of metadata that is attached to each tweet. This metadata includes information regarding geolocation (either a place name or coordinates), the author name, a defined location, a timestamp of the moment the tweet was sent or retweeted, the number of retweets, the number of favorites, a list of hashtags, a list of links, etc. This information is valuable and has the potential to provide intelligence when attempting to extract information for use in crisis response. Twitter APIs also offer a varying number of filters and filtering capabilities including additional filter operators and tweet enhancements (e.g., profile location and un-shortened URLs).

The Twitter platform provides various APIs for searching Tweets including (i) the standard Twitter APIs consisting of REST APIs and Streaming APIs, and (ii) the enterprise APIs including filtered firehose, historical search and engagement APIs for deeper data analytics. FAIS uses the standard Twitter APIs because it is free and less challenging to gather Twitter flow of information and Hashtag (#) driven topics. Standard API provides an endpoint to return time-series counts of Tweets matching user query. The interested geotagged data that can be gathered using the standard API are images, videos, text, and numeric (e.g., flood depth) from citizen inputs. For privacy issues and other security concerns regarding personal information, our developed Twitter APIs do not employ any user-related features (such as number of followers on Twitter), rather focused on the message-related attributes. The APIs also calculates sentiment of the tweet and the identified categories. Sentiment is a text/tweet analysis that is used to categorize and classify the opinions and sentiments expressed in text. Three classes of sentiment were implemented using Twitter APIs including (i) Positive—a positive sentiment has been expressed, (ii) Negative—a negative sentiment has been expressed, and (iii) Neutral—a neutral or no reaction sentiment has been expressed.

3.3.1. Twitter Streaming API

We used Tweepy package and integrated it with FAIS as an easy-to-use Python library for accessing the Twitter data. Twitter developer account was used to access Token, Token Secret, Consumer Key, and Consumer Secret to manipulate Twitter functionalities. To protect the credential, the authors decided to develop a Twitter Streaming bot (functions on both iOS and Mac) and deployed it at Heroku cloud platform outside of the application access which can be controlled by the Heroku User Interface. Heroku is a cloud platform as a service (PaaS) that enables system-level supervision and coordination of Twitter APIs, crowd sourced data, and tweets. Our developed Streaming Twitter bot automated all Twitter data gathering and continuously watched all Twitter activities during real time implementation. To be able to watch Twitter activity in real time, the bot gets notified when new content, such as tweets, that matches certain criteria (such as "Dorian Floods") is created. This is particularly important when dealing a vast amount of real time tweets. We created a reusable Python module (a module config) containing the logic common to the bot functionalities. This module reads the authentication credentials from environment variables and creates the Tweepy API object. By reading the credentials from environment variables, user avoids hard coding them into the source code, making it much more secure. The bot reads the credentials from four environment variables, including CONSUMER_KEY, CONSUMER_SECRET, ACCESS_TOKEN, and ACCESS_TOKEN_SECRET. After reading the environment variables, the bot creates the Tweetpy authentication object that eventually uses to create API object. The bot uses the logging Python module to inform errors and information messages that help user debug them if any issue arises. The tweets save constantly in the MangoDB database when the bot is in operational use. The administrator can choose to activate or deactivate the bot and change the keywords for Streaming services. Our developed Twitter bot contains three components (Fig. 1) notably:

1. Twitter Client: This component talks to the Twitter API and authenticates the connection to use its functionality. This also hosts a function called tweets_listener, which will continuously stream tweets and listen for the matched keywords. Once it finds the match, it will then talk to the other two components.
2. Tweet Analyzer: It analyzes the tweets and gives it a score after a match is found.
3. Twitter Streamer: This module streams tweets from pre-specified keywords, analyzes the data, and organizes them into a data frame. The collected tweets will then store in a MongoDB database waiting to be extracted. A conceptual process about how to gather real time tweets using the developed Streaming Twitter bot is shown in Fig. 1.

Due to the size of queried data, the Twitter bot filters the data and only keeps text, location, author, and date of tweets which are eliminated for over 95% of uninterested data. FAIS application has access to MongoDB cloud database without having access to Twitter bot. This allows the user to see the result of our services while protecting FAIS Twitter account privacy and information allowing the users to have access to database resources whenever needed.

Using Twitter Streaming bot, FAIS was able to identify at-risk locations to flooding. The application first cycles through a set of USGS web addresses for river gauge height readings, parsing these flat files using Python web scraping technique and obtain all the latest river levels. Each river level reading is compared with its respective long term cached average level, to identify the highest relative river levels in real-time. The highest river level then intersects with watershed polygons as well as geotagged tweets to identify at-risk locations to flooding. Geotagged tweets coordinates considered as a center-point for approximately 16 km wide square boxes. This size is arbitrarily chosen to cover the areas nearby to each flood gauge. The retrieved tweets are constantly stored in a MangoDB database during operational use which provide an ideal open-source database to store JSON format files.

3.3.2. Twitter Search API

Twitter Search API is a functionality to search past tweets that match the search criteria. Twitter contains the search functionality but has a limited amount of search result, search frequency, and time constrained. To upgrade the functionality, the user requires to pay for the upgraded Twitter account. With the goals of keeping FAIS application free and open source, an alternative search method was implemented which was using the REST service freely from Twitter search function. Twitter search operation allows the users to look for tweets based on interested user account, keywords, language, and time period but this is not an ideal procedure for the user to go through the website and gather all the interested tweets that they need. Therefore, this project used "urllib" python library for Universal Resource Locator (URL) handling modules,

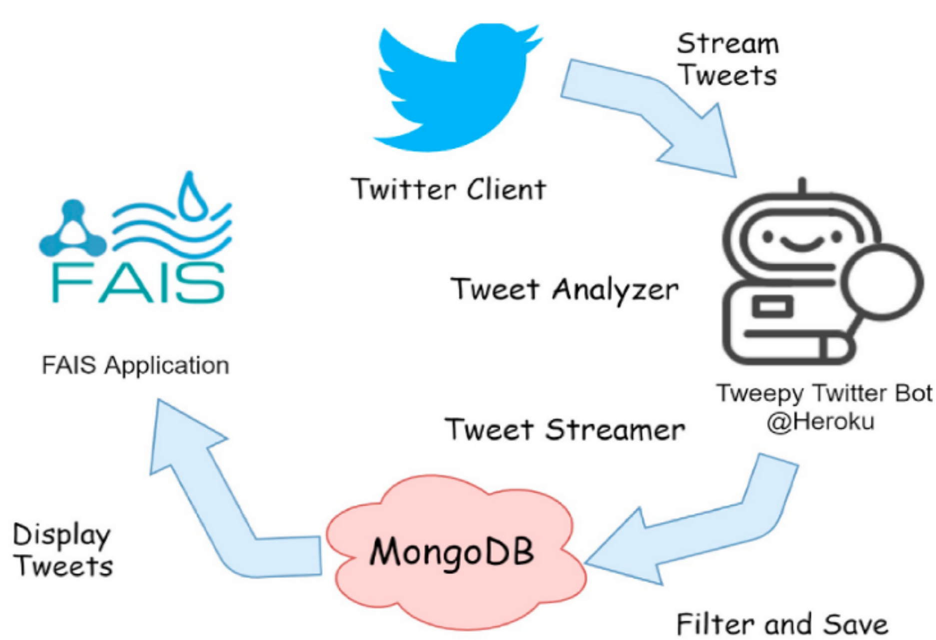


Fig. 1. The workflow and integration of developed Twitter bot components.

to send specific request to Twitter search web service and to gather the tweets. The URL can be then modified based on the user search criteria.

The Search API continuously sends request URL to Twitter and waits for the JSON response from Twitter. After receiving the JSON file, the Search API extracts interested tweets including username, text, retweets, likes, date, id, Permalink, user_id, media, URL, and the sentiment. The response JSON file contains a 'min_position' data file which can be used to move iteration forward. The iteration will keep running until the response exceeds maximum number of tweets limit set by the user. Note that the date limit is set to the request URL sent to Twitter. Twitter outcomes such as the sentiment analysis can be performed for a tweet. The sentiment shows whether or not a tweet has positive (+1), neutral (0), or negative tones (−1). This information is important because it can be used by social scientists to study the impacts of flooding on the citizens and how the residents responded to the flooding and damage.

3.4. Development of USGS and 511 Traffic Data APIs

USGS collects and stores multiple river system data across the United States. These are river flow data (flood, streamflow, gauge heights), water quality, ground water levels, and precipitation at defined gauging stations which are strategically placed at the outlets of rivers and lakes. These placements allow USGS to correctly monitor and collect the data and compute several statistical indices related to the river flow. USGS server provides two different types of flow data including real-time and historical records based on datetime. FAIS gathers USGS discharge data (daily streamflow and flood data) as well as gauge heights, and river web camera images. The stream data is value of flow rate in cubic feet per second (cfs) and water level (gauge height) in feet. FAIS first gathers both historical and real time USGS data for any state in US. It then analyzes the information and plots the historical and real time data. These data are recorded based on data recording time step such as every 15-min to 1-h intervals. FAIS made USGS data collection seamless and straightforward by providing access to the station's information including station name, ID name, latitude, longitude, and URL to USGS server.

In addition, FAIS gathers USGS real-time cameras and 511 traffic images. DOT in each state provides traffic data in real-time to the users through 511 web application. The cameras are strategically placed on the bridges and roads and along the interstate, allowing operators to continuously monitor road conditions. They also monitor rising and falling water stage over critical infrastructures such as roads and bridges. In addition, USGS established real-time web cameras for critical rivers where there is high chance for flooding and inundation issues. FAIS currently collects six USGS Sevillian cameras images across South Carolina (SC). The application uses a dynamic mapping interface to allow the user to select specific traffic camera and access the data in real-time. Indeed, there are embedded URLs at both USGS and 511 Traffic websites that were used to gather both USGS and traffic images in real time. Images constantly store at MangoDB database when the application is in operational use.

3.5. Web Development Platform

With the use of Python in the FAIS API tools, it was straightforward to use the same language and develop a web application. The seamless interaction between the API and Python motivated this study to use Django as a web framework. Django is a rapid development platform with clean, and pragmatic design that provides a widest range of libraries including Twitter APIs and machine learning tools. The key point of making it an ideal development framework is due to its focus on automation as much as possible and adhering to the "don't repeat yourself (DRY)" principle aiming at reducing repetition of software patterns and avoid redundancy. Django allows the user to develop a web application using Python element along with a classic web development language like HTML, JavaScript, jQuery, etc. This means that developers are not stuck with limited approaches of solving the problem. This interaction can also be a downside of this development framework as it requires the knowledge of different languages to debug like Ruby on Rails. FAIS is version -controlled using GitHub since Git is a widely used version control system and an open source repository hosting service.

Fig. 2 shows the overall workflow of the FAIS pipeline development. The workflow includes collecting the data from various resources, analyzing the potential at risk areas and providing actionable results to users/stakeholders. As illustrated, FAIS combines multiple APIs along with machine learning and image processing algorithms (Google Vision API) and FFA script (written in R) to provide both historical and real-time information about flood-risk incidents. Our developed prototype offers an end-to-end, open source, web-based, pipeline architecture to address the crucial issue of how first responders and decision makers can be smartly informed and acted in emergency situations. To achieve our aim and answer three research questions stated above, we tested FAIS operationally during Hurricane Dorian flooding event (September 04–06, 2019) as well as during historical event, Hurricane Florence flooding (September 13–16, 2018), across the Carolinas.

4. Results and Discussion

4.1. Retrieval of Social Geotagged Tweets Using Twitter APIs

This section addresses the first and second research questions: How to programmatically and automatically identify areas at-risk of flooding based on crowdsourced data, real-time flood peak rates, and river level information? And how to spatially display the retrieved data and implement this information for alert and warning system? To address these questions, two options, i.e., the Search and the Streaming APIs, were included to the FAIS application for data gathering. We used Tweepy library as a Search API and developed a Streaming API using Twitter bot. Search API was more suited to singular and specific queries for tweets, whereas the Streaming API provided a real-time stream of tweets.

There were widespread flooding events across the Carolinas when we designed and Beta tested FAIS application. These events caused localized and major flooding as well as above average river level for most basins in the Carolinas. FAIS was tested for historical hurricane driven flood events such as Hurricane Florence (September 13–16, 2018) across the Carolinas. In addition, the prototype was operationally tested during Hurricane Dorian event (September 04–06, 2019) and georeferenced tweets were gathered in real-time to identify at-risk locations. An example from each hurricane driven floods is presented with intersecting at-risk locations and geotagged tweets. A web-based console and a visualization tool- GeoJSONLint13 were used to view results and inspect the polygons.

4.1.1. Hurricane Dorian Case Study

We monitored georeferenced tweets, filtered by keywords and queries across the shortlisted areas in the Carolinas during Hurricane Dorian event (September 04–06, 2019). To identify at-risk locations in real-time, a shell script in Python ran on a local computer server, the script was reset every 3 h in order to update areas at-risk of flooding from the latest national and environmental data sources as well as Twitter feeds. A period of 15 min was initially chosen as intended trade-off between tracking the latest at-risk area forecasts. API updates varied between 15 min to several hours but based on real-time testing the period extend to 3 h to allow some reaction time from impacted citizens on Twitter. However, the choice of time period depends solely on the project requirements as well as the severity and impacts of flooding events. A period of 6–12 h was chosen during Hurricane Dorian flooding event as an intended trade-off between tracking the latest at-risk areas to flooding (API updates varied between 15 min and 6 h depending on flood data time step), and that allowed capturing reaction from those at-risk areas in real time.

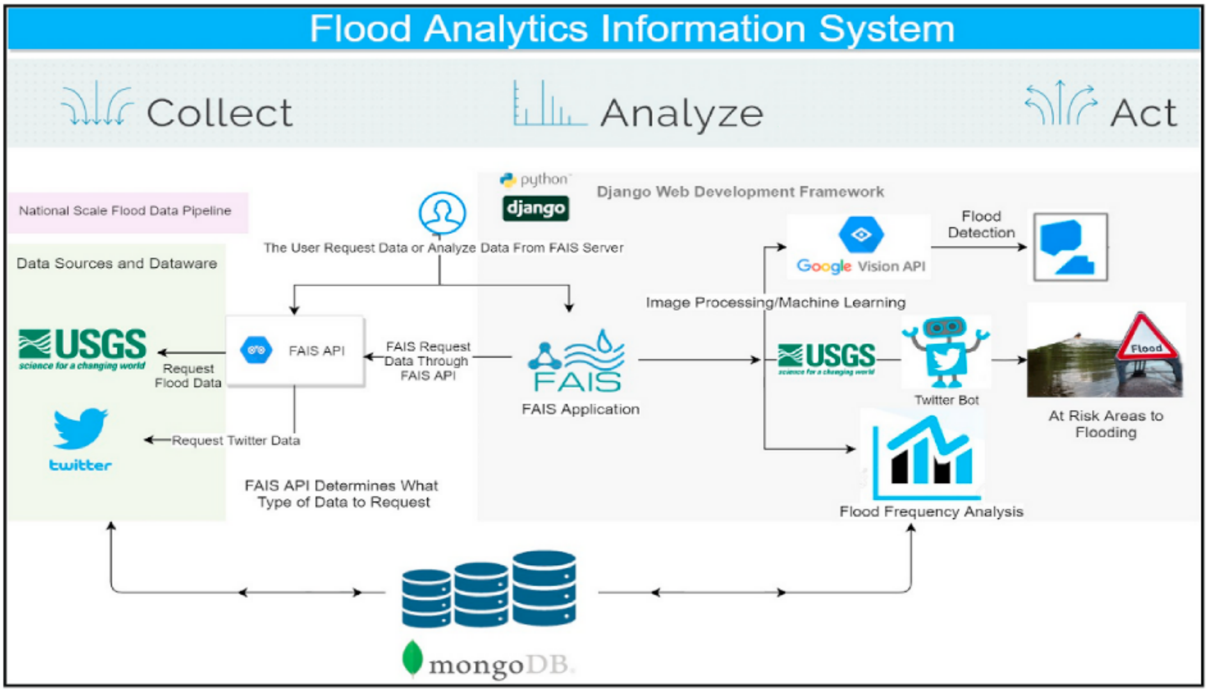


Fig. 2. FAIS workflow and structure.

The collected tweets provided a real-time dataset which explored further and used to prioritize at-risk locations for Dorian flood simulation (results not shown here). This provided a general indication of the proportion of potentially relevant tweets that could be used to identify flooded areas and improve flood situational awareness for first responders. FAIS filtered georeferenced tweets returned from the Streaming API with keywords and a blend of geographic data sources (geolocations) to show areas affected by Dorian floods, at a regional scale.

During Hurricane Dorian, we constantly ran the Twitter bot to collect tweets and to determine at-risk locations over time. The stored tweets provided a real-time dataset which then explored and used to prioritize at-risk locations to Dorian flooding. This provided a general indication of the proportion of potentially relevant tweets that could be used to identify flooded areas and improve flood situational awareness. Data retrieval of Dorian flooding was explored via basic text query searches such as "Hurricane Dorian Floods" and centroids were added to tweet's geotag for creation of map-based visualizations. Currently, the size of bounding boxes of tweets are 16 km from the centroid that determines the number of tweets intersect with the bounding boxes to yield and visualize at-risk areas. The size for 16 km boxes is arbitrarily chosen to cover the areas nearby to each USGS gauge in the coastal Carolinas and may differ location by location based on the proximity of the tweets to the nearest flood gauge. Previous studies showed that messages within 10 km of severely flooded areas had a much higher likelihood of being related to such events (see de Albuquerque et al., 2015). Fig. 3 illustrated collected geotagged tweets intersected with watershed boundaries for South Carolina during Hurricane Dorian flooding.

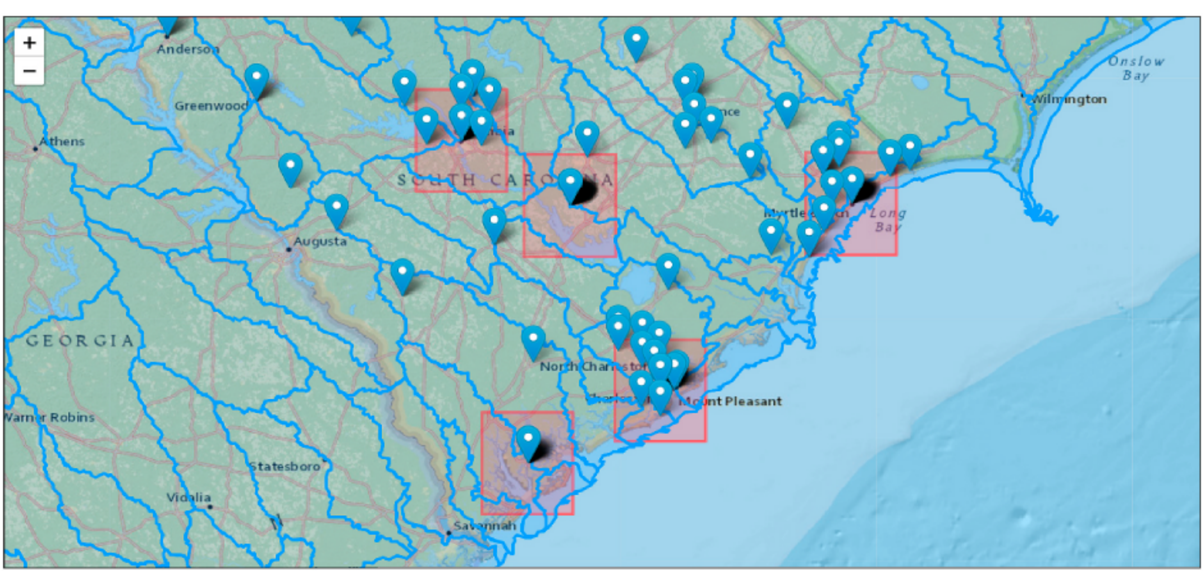


Fig. 3. Geotagged tweets retrieved via Twitter Search API in real-time during September 04, 2019 for Hurricane Dorian event across the coastal SC using specific location searches combined with keywords.

Delineating at-risk areas could be further improved by populating critical infrastructure (bridges, roads) and flood defense structures (levees, dams, reservoir) which are out of the scope of this research. Table 1 presents several real time tweets, along with their geolocation information, Twitter account, etc. which were retrieved during Hurricane Dorian flooding in South Carolina (September 04–06, 2019). The classifier was developed by an annotator (the first author) manually labeled a subset of >800 tweets from September 04–06, 2020 as either relevant or irrelevant tweets.

4.1.2. Hurricane Florence Case Study

We implemented FAIS to extract geospatial footprint of Hurricane Florence flooding event using georeferenced tweets. The user has an option to select a proper Twitter account for the historical tweets gathering. We chose NWS Twitter account (@NWS) to gather Hurricane Florence historical tweets and map the geospatial footprint (map view wasnot shown here). FAIS was able to successfully identify a dynamic set of at-risk areas using Twitter Search API for Hurricane Florence event. At-risk areas to flooding were identified by intersecting the geotagged tweets with watershed polygons and river gage heights. Search API that is designed within the FAIS application was able to collect the tweets and visualize them in on a Leaflet map via Folium Python API. Fig. 4 showed the FAIS design and outcomes for Twitter Search API implementation of Hurricane Florence flooding event. The Search API provided the source of the tweet, image, the sentiment, and map view of the tweets. If the user changes the keywords, the Twitter Search API will then filter the tweets and save the data in a MangoDB database. We used National Geographic base map tiles, and geolocated each tweet based on its geographic coordinates. Each tweet is represented by a clickable marker, which provided a pop-up box of the tweet information, locations, coordinates, time, etc. Alternative interactive base map can be included to the prototype such as OpenStreetMap, Imagery, Topographic, etc. FAIS can be applied nationally across the US, running for both real time and post flood event tweet gathering and assessment.

Table 2 displays several tweets gathered using NWS account for Hurricane Florence flooding in SC. The tweets were retrieved via location filter that varied in terms of spatial precision (less or no exact “Place” metadata or coordinates). Location filtering reduced mismatch of off-topic tweets (irrelevant), which is an issue for keyword-based retrieval (de Albuquerque et al., 2015). Although, the word “flood” and its translations were also frequently used figuratively or in a transferred sense (e.g., “I am flooded with many tasks”). We performed the filtering mostly by location filter capabilities designed within Twitter APIs that seemed to be an effective filtering approach to improve collecting the relevant tweets and reducing manual labor. Overall, we collected over 800 “location-filtered” tweets between September 13–16, 2018 across the Carolinas.

4.1.3. USGS Data Collection

This section addresses the third research question: How to seamlessly retrieve data from various sources and how to use this information for making actionable decision? FAIS collects and displays USGS data that include the date, discharge (cfs), and gage height (ft) along with their associated plots. Gathering USGS historical data involves selecting the target state, the interested station, and the date. After query criteria is entered, FAIS creates a request URL and sends it to the USGS server for collecting the data. The prototype displays the data as “Table View” (Fig. 5(a)) as well as “Map View” (Fig. 5(b)) and plots the results (Fig. 5(c)). The user can also upload a csv format file of all the collected data that contains station name and ID, latitude, longitude, discharge, gauge height, and USGS original URL (Fig. 5(d)).

In addition, FAIS collects six USGS Sevillian cameras images across South Carolina. These cameras are located at Rocky Creek near Wade Hampton, Rocky Branch at Whaley St., Columbia, PeeDee River near Florence, Lake Monltrie Trailrace Canal at Moncks Corner, SC, Tearcoat Brach at I-95 near Manning, SC, and Pocotaligo River at I-95, above Manning, SC. The image contains the meta data such as when and where the image was captured. FAIS also collects North Carolina (NC)’ 511 images in real-time. Several cameras that are located in the coastal region of NC were selected to record road flooding conditions in real-time. These images are crucial for monitoring of water level and providing early warning to the local community in case the water level increases above a predefined threshold value. Both USGS and 511 traffic images cane be stored at MangoDB database when the application is in operational use. Readers are referred to FAIS web server (https://floodanalytics.clemson.edu/) for more detailed information and to stream and store the camera images.

4.2. Image Processing and Label Detection

We integrated Google Vision API with FAIS application for object detection of flooded and non-flooded images. The tool first trains the automated ML Vision model and then labeled the datasets. This provides custom label detection data with scores. We used flooded and non-

Table 1
Examples of tweets as gathered during Hurricane Dorian flooding in South Carolina. Note: These geotagged tweets are manually labeled as “relevant” by an annotator (the first author).

Tweet/Text	Twitter Account	Geolocation	Date/Time	Latitude	Longitude
Hurricane Dorian flattened and flooded Great Abaco Island …	@AnitaNelam	South Carolina, USA	Thursday, September 05, 2019 02:58:52 p.m.	33.68	−80.43
Low country SC hit by hurricane Dorian, which peaked at Cat. 5.	@ b7ameister	South Carolina, USA	Thursday, September 05, 2019 20:04:52 p.m.	33.68	−80.43
2019 is the fifth consecutive year (2015–2019) in which 10 or more billion-dollar weather and climate disaster events have impacted the United States. Hurricane Dorian was one of them - NOAA	@ bopete	Charleston, Charleston County, South Carolina, USA	Friday, September 06, 2019 17:52:35 p.m.	32.78	−79.94
I'm about to! The people who run our HOA have no clue what their (they are!) doing. I had to dispute a $350 charge because I didn't want to power wash my house a week before hurricane Dorian and now they're trying to charge it to me again.	@BrubakerSteven	Pawleys Island, Georgetown County, South Carolina, USA	Friday, September 06, 2019 17:22:59 p.m.	33.42	−79.12
This is awesome! I followed all the buoys and Data Collection Research during Hurricane Dorian with wave heights and timings.	@LisaGethard	Charleston, Charleston County, South Carolina, USA	Friday, September 06, 2019 17:45:43 p.m.	32.78	−79.94
… remembering those in the Bahamas still feeling impact Hurricane Dorian, especially the hundreds of kids orphaned by the storm.	@F3Stumble	Columbia, Richland County, South Carolina, USA	Friday, September 06, 2019 11:24:28 a.m.	34.01	−81.03
Hurricane #Dorian skirting the Carolina Coast from a satellite perspective.	@EdPiotrowski	Myrtle Beach, Horry County, South Carolina, USA	Wednesday, September 04, 2019 15:24:14 p.m.	33.69	−78.89

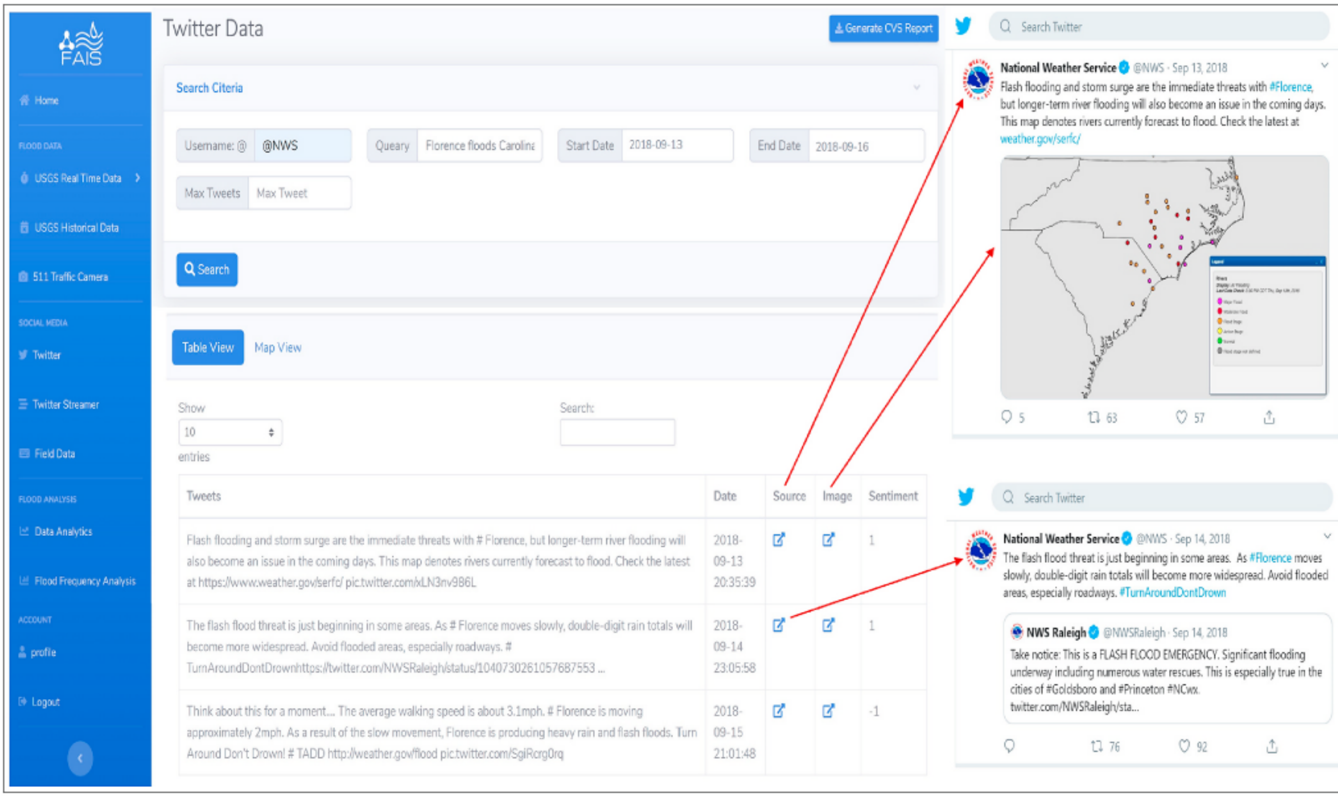


Fig. 4. FAIS illustration for Florence driven flood data gathering, geotagged source, image, and the sentiment.

flooded images of Hurricane Florence to train the Google Vision API and detect the objects in the images. We extracted high quality still images from Hurricane Florence videos. Overall, we analyzed a range of 100–240 time-lapse images to detect the labels and sore them. The higher quality images user delivers and the better the design of the model user uses; the smarter outcome will be produced.

Fig. 6 shows the detected objects for the Person Street bridge, Cape Fear River, NC. Flood (91%), land lot (91%), and Asphalt (89%) are detected as major labels in flooded images while the major labels and scores are given to sport venue (89%) and residential area (85%) in non-flooded images. Fig. 7 showed the score and the number of images that we used to detect the label for this location. As illustrated, the algorithm was able to detect water/flood (0.97) during and after hurricane event. The algorithm also detected bridge infrastructure (0.92) for pre-event while distinguished it as a reservoir (0.9) during post flooding event due to overtopping issue (see Fig. 8). A number associated with every returned label annotation, representing the Vision API's assessment of the label's accuracy. Scores range from 0 (no confidence) to 1 (very high confidence) (Fig. 9).

Overall, Google Vision API detected 11 labels for the flooded and non-flooded images at the New Bern, NC while captured 17 labels for the Person Street bridge flooding event at the Cape Fear River, NC. Despite the lower number of labels, 92.6% of Vision's labels turned out to be relevant (8 errors). It would be worth mentioning that Google Vision API could further improved by cluster equivalent labels together ("Flood", "Water", and "Water Resources/River") whenever a water being is detected in the image. By collapsing such labels into one, the number of detected labels will definitely decrease, and it may also have some impacts on relevance scores. Our analysis suggests that Google Vision API has detection problems whenever flooded areas are too small (below 50px) partially out of the image, or occluded by other obstacles such as vehicles and trees. This might improve over time with a more specialized pattern recognition layer and approaches such as including an image segmentation to the tool based on watershed algorithm that builds barriers in the locations where water merges. These barriers can provide segmentation results that could be used to estimate inundation extend and flood level using a reference such as a bridge pier or a car. It should be noted that we didn't focus on other accuracy parameters such as location, direction, and special traits (Vision doesn't provide such data). Further work and a considerable dataset expansion may provide useful insight into flood location and direction accuracy, although the difference of a few pixels is usually negligible for most applications. The best scoring fragments for a given label matched well with visual appearance of the object in the image. Indeed, the pairs of flooded and non-flooded images should have the same view angle and geometry.

Although image processing and machine learning algorithms designed within Google Vision API yielded successful results as indicated by the scoring metrics, there are several challenges in detecting the flooded or inundated areas. For example, this study found discrepancies in labeling dry/flooded image pairs specifically for differentiating "flood", "floodplain", "waterway", and "road". The algorithm detected some labels that weren't existed in the image such as roof and floodplain. To prevent this issue, the model can be validated on a simple dataset where a single object (e.g., flood) occupied most of the image. A wide variety of images with annotated objects that co-occur in the same images can be used to test the accuracy of the tool. Further, the algorithm can be enhanced by including in painting procedure to efficiently remove unwanted objects such as the sidewalk and streetlight to detect the inundation border as studied elsewhere (Witherow et al., 2019). Although, differences in image resolution and lighting, and environmental conditions have significant impacts on annotating an image with a reliable label and score.

Flood detection provides important information to stakeholders because the results can be proactively used for flood emergency management. Indeed, the capability to detect temporal changes in image sequences is crucial and this information can be combined with other datasets such as USGS flood peak rates and rainfall (radar) data to develop an automated image-based flood alarm system as a disaster-

Table 2
Examples of tweets, and date and time as gathered for Hurricane Florence flooding in South Carolina using NWS Twitter account.

Tweet/Text	Twitter Account	Date/Time
Here is the latest (5PM EDT) forecast track and key messages with Hurricane Florence from @NHC_Atlantic. NOAA Hurricane Hunters discovered Florence strengthening and could become a major Hurricane on Monday. #HurricaneFlorence #HurricanePreparedness	@NWS	5:26 p.m. · Sep 9, 2018
Hurricane #Florence continues to strengthen; max winds are now 140 mph. Further strengthening is forecast, and it is expected to be a large and powerful hurricane as it nears the East Coast. See below for a 1-stop page for the latest #Florence info: http://weather.gov/wrn/florence	@NWS	5:04 p.m. · Sep 10, 2018
Flash flooding and storm surge are the immediate threats with #Florence, but longer-term river flooding will also become an issue in the coming days. This map denotes rivers currently forecast to flood. Check the latest at https://weather.gov/serfc/	@NWS	4:35 p.m. · Sep 13, 2018
Here is the latest forecast track and key messages for Hurricane #Florence from the @NHC_Atlantic. Damaging, hurricane force winds likely along portions of the coasts of North Carolina and South Carolina beginning this evening. #HurricaneFlorence	@NWS	5:35 p.m. · Sep 13, 2018
Hurricane #Florence is resulting in communication difficulties in some areas. Having multiple ways to receive warnings though this event could save your life.	@NWS	10:52 a.m. · Sep 14, 2018
WIDESPREAD, LIFE-THREATENING flash flooding and storm surge ongoing or possible today. Have multiple ways to receive warnings. It could save your life! #Florence #HurricaneFlorence	@NWS	10:52 a.m. Sep. 14, 2018
The flash flood threat is just beginning in some areas. As #Florence moves slowly, double-digit rain totals will become more widespread. Avoid flooded areas, especially roadways. #Turn Around Don't Drown	@NWS	7:05 p.m. · Sep 14, 2018
Think about this for a moment …. The average walking speed is about 3.1 mph. #Florence is moving approximately 2 mph. As a result of the slow movement, Florence is producing heavy rain and flash floods. Turn Around Don't Drown! #TADD http://weather.gov/flood	@NWS	5:01 p.m. · Sep 15, 2018

monitoring application. Image-based flood warning information can facilitate proactive monitoring and damage assessment, and early warning to rising water levels and associated inundation areas in real time. However, current deficiencies in Google Vision API and overall image processing algorithms explained above may limit Vision application for image-based early warning system.

4.3. Flood Frequency Analysis of Two USGS Gauges

FAIS provides flood frequency analysis to estimate flood quantiles that combines elements of observational analysis, stochastic probability distribution and design return periods. FAIS currently uses Gumbel distribution to compute FFA for any given flood gauging station in US. Fig. 10 shows FFA for the USGS 02147500 Rocky Creek at Great Falls, SC, SC as well as the USGS 02196000 Stevens Creek Near Modoc, SC. As illustrated annual flood peak of ~15000 cfs for the Rocky Creek represents a design return period of 25-year while the same flood peak shows a design return period of <5 years for the Stevens Creek gauge. The difference is related to the size of the drainage system. On small watersheds, a 25-year rainfall event may produce a 25-year flood event. On large watersheds, however, the 25-year flow event may be produced by a series of smaller rainfall events. This distinction should particularly be kept in mind by the practitioner while dealing with design projects in large watersheds. The likelihood of a 100-year flood (50% annual chance) occurring at both gauging stations ranging between 32000 to 35000 cfs. Both gauging stations appeared to experince large flooding events particularly USGS 02196000. This gauge is part of the Savannah Basin where frequent flooding is a major threat for the residents. FFA for this location proved that high peak values made critical contributions to the upper tail of the Gumbel probability distribution function. This analysis is useful in providing a measurement parameter to assess the damage corresponding to specific flow during flooding event. Along with civil infrastructure design, flood frequency analysis can be used in flood insurance and flood zoning activities. Accurate estimation of flood frequency not only helps engineers in designing safe infrastructure but also in protection against economic losses due to maintenance of structures. However, the accuracy of FFA estimates may vary using different probability distributions such as Pearson type III, Gamma, and Normal distributions. We recommend using Gumbel function for a river system with less regulation and significant reservoir operations, diversions or urbanization effects. Efforts are underway to include other distribution functions to the prototype. Therefore, care should be given when the current function uses for any structure design purposes.

5. Conclusions

In this paper, we developed a prototype for flood data analytics and assessment using IoT APIs, crowd intelligence, and Big Data gathering approaches. Our study presented the first step towards identifying at-risk areas to flooding in real time and defining the geospatial footprint of a flood event using georeferenced tweets. We aimed to use IoT-APIs and collect environmental data (river levels and discharge) at the national scale in combination with Twitter APIs to identify the areas affected by a flood. The application also uses image processing and machine learning to detect label and scores objects in time-lapse images. FFA algorithm was also developed in R and embedded with the FAIS to perform flood design metrics and peak rates that could be combined with Twitter and image processing results for studying the design metrics of overtopped structures as well as economic assessment.

Overall, our proposed pipeline proved to be robust and user-friendly tool for both real-time and post-event analysis of flooding at regional scale that could help stakeholders for rapid assessment of the situation and damages. Gathering the data and analysis of flood situation take on average few minutes to select the data periods and set the designed algorithm to run; thereby it reveals to be promising for meeting first responders' needs during emergencies. During real-time event, the time between a tweet appears online and visually plots for a stakeholder as being potentially relevant (in terms of location and content) would be in the order of few seconds to minutes, thereby this rapid analysis can provide an early information channel for asset allocation and rescue. The application proved to be proficient for real time flood assessment as it was tested during a 2-day Hurricane Dorian event across the Carolinas with over 15,000 geotagged tweets collected from 38 dynamic, potentially at-risk areas.

Using Streaming API during Hurricane Dorian, tweets were collected and labeled as "relevant" because they were related to ongoing hurricane and flood event. Streaming API assigned a tweet as "relevant" class when a user tweeted about ongoing Dorian floods at the time of posting. This includes, for example, tweets referring to response, rescue, road closure, and failure of critical infrastructure/water supply. All other tweets were assigned being "irrelevant". Examples include historic flood event commemorations, the use of the word "flood" in figurative or transferred sense, etc. It is interesting to note that the Streaming API showed a great performance to maximizing the retrieval of geotagged tweets via location filtering as well as using more detailed key words, and this finding is in agreement with other studies (see Barker and Macleod, 2019; Ekta et al., 2017; Tsou et al., 2017; Morstatter et al., 2013). The alternative Search API preferentially relied on a tweet's geotag presence and therefore reverts to the more general and less

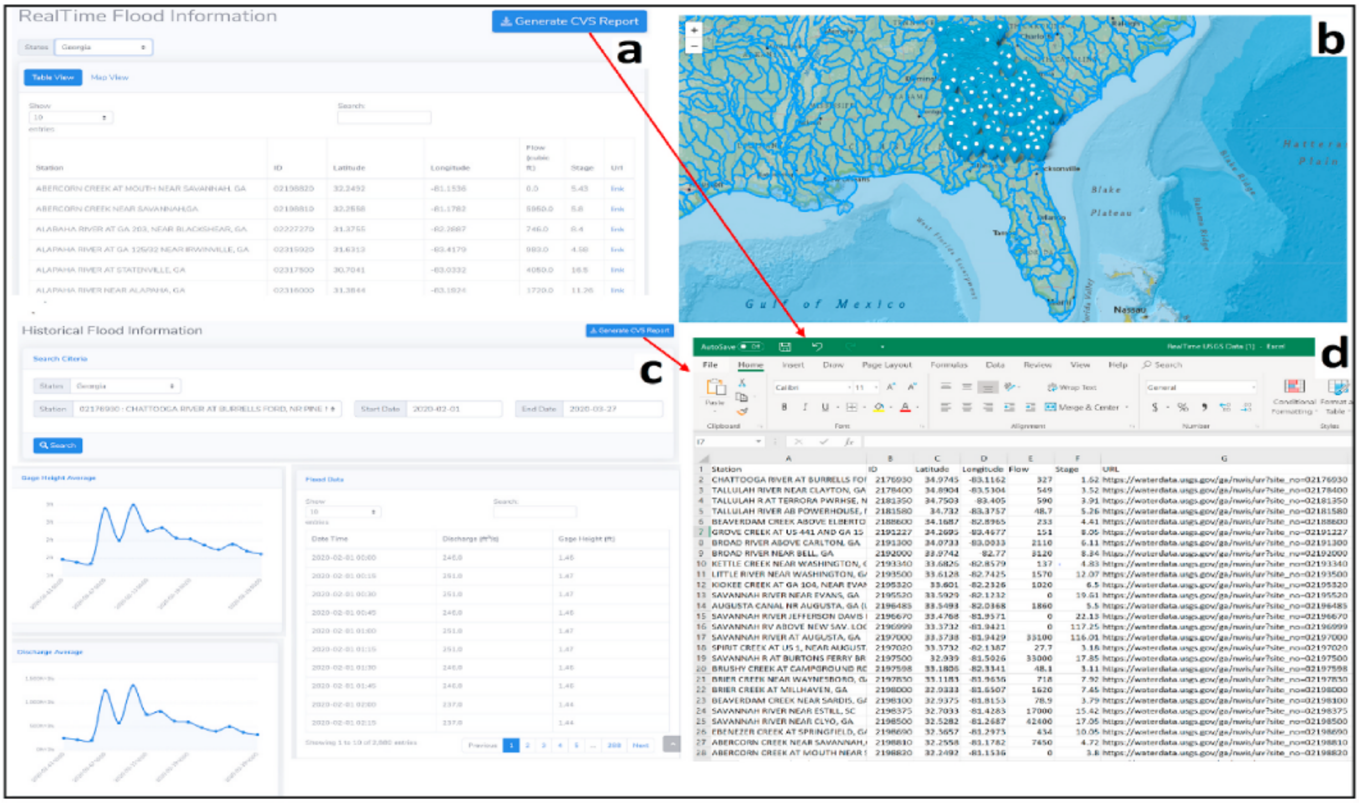


Fig. 5. Real time flood data collection (a), map view (b), historical flood data and the plots (c), and a csv format data (d) for both real time and historical data for Georgia, USA.

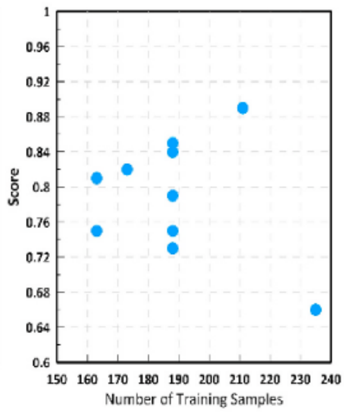
Fig. 6. Flooded (right) and non-flooded (left) images with detected labels (roads, residential area, water, etc.) and scores during Hurricane Florence event at the Person Street bridge, Cape Fear River, NC.

spatially accurate user profile location. The user profile location is manually set by the user, and maybe different to the specific tweet's actual location. Furthermore, tweets are often contained noisy and redundant data and that cleaning and geoparsing might be difficult and present a time-consuming task. Most importantly, the Search API permits just a single location per query and is not suitable for spatiotemporal assessment of flooding event in real-time. Our developed Streaming API and a Twitter bot allowed simultaneous monitoring of multiple locations, necessary for real-time, national-scale flood assessment.

In addition, we integrated Google Vision API to parcel images into labels and scores them. As discussed, image processing and machine learning is not perfect: it often splits objects into separate labels, labels objects that are not presented in the image, or includes multiple objects in the same label. However, over a large dataset of time lapse images, we found that image processing approaches perform better for many object categories. We showed improved object discovery for a more complicated set of flooded and non-flooded images for recent hurricane driven floods in the Carolinas. Our research revealed that images and image sequences (videos) make up about 80 percent of all corporate and public unstructured Big Data for flood related studies, providing an excellent data sources for flood analytics research. Although, the result described herein is encouraging for time-lapse images/videos analysis, the following open issues still exist in image recognition: (i) how do we effectively improve recognition and label detection approaches to cope with numerous object categories that exist and are recognizable by humans? and (ii) How can we apply image processing and machine learning approaches to motion and actions in video and label them? We hope that the techniques developed in this study will offer a good starting point in addressing these issues and developing more intelligent

Before Flood

Object Class	Score	Training Samples
Sport Venue	0.89	211
Residential Area	0.85	188
Mode of Transport	0.84	188
Asphalt	0.82	173
Waterway	0.81	163
Roof	0.79	188
Tree	0.75	188
Walkway	0.75	163
Land Lot	0.73	188
Suburb	0.66	235



After Flood

Object Class	Score	Training Samples
Water	0.97	163
Atmospheric Phenomenon	0.94	233
Flood	0.91	193
Land Lot	0.91	188
Asphalt	0.89	173
Water Resources	0.86	163
Storm	0.84	163
Geological Phenomenon	0.83	233
Floodplain	0.82	163
Wind	0.82	233

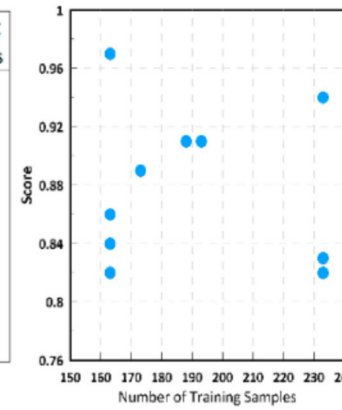


Fig. 7. The scores and the number of images that were used to detect the label for the Person Street bridge, Cape Fear River, NC.

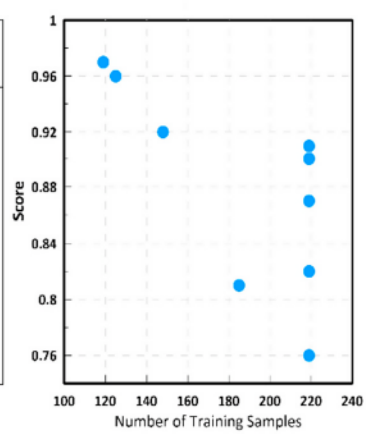
Label and Score
Water : 0.97
Water resources : 0.96
Bridge : 0.92
River : 0.91
Bank : 0.9
Floodplain : 0.87
Watercourse : 0.82
Flood : 0.81
Waterway : 0.76
Event : 0.72

Label and Score
Body of water : 0.99
River : 0.98
Water resources : 0.98
Water : 0.97
Nature : 0.96
Bank : 0.95
Waterway : 0.9
Reservoir : 0.9
Floodplain : 0.9
Lake : 0.88

Fig. 8. Flooded (right) and non-flooded (left) images with detected labels and scores during Hurricane Florence event at New Bern, NC.

Before Flood

Object Class	Score	Training Samples
Bridge	0.92	148
Water	0.97	119
Water Resources	0.96	125
River	0.91	219
Bank	0.90	219
Floodplain	0.87	219
Watercourse	0.82	219
Flood	0.81	185
Waterway	0.76	219



After Flood

Object Class	Score	Training Samples
Reservoir	0.90	148
Water	0.97	119
Water Resources	0.98	125
River	0.98	219
Bank	0.95	219
Floodplain	0.90	219
Watercourse	0.82	219
Flood	0.81	185
Waterway	0.90	219
Nature	0.96	211
Lake	0.88	198

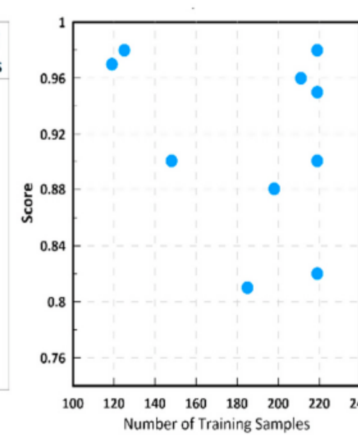


Fig. 9. The scores and the number of images that were used to detect the labels for flooded and non-flooded images during Hurricane Florence in New Bern, NC.

and proficient image segmentation and processing methods for flood studies.

The authors believe that data analytics applications in the field of flood risk management should adopt a modular view, moving from a component based to a national scale. At present, data analytics research remains in its developing phase into existing workflows and practices. There appears a major gap particularly in seamless integration of different data sources as number of datasets keep increasing over time. Although, the main issue in integrating these datasets is to ensure data consistency, accuracy and completeness for informed decision-making processes. As data collection through various heterogeneous sources in real-time is highly susceptible to noise and uncertainty. To this end, security as well as privacy issues in data transmission and analytics, and storage also need to be under constant control to ensure the authenticity of data and citizen-based crowd sourced information while keeping the confidentiality of people's sensitive information. In this study we set out the major design choices and decisions made by acknowledging the inputs from many decision makers and first responders. We appreciate the discussions with SCEMD and believe FAIS can be a more transparent and efficient tool than many other similar pipelines. We acknowledge there is a need to evaluate the pipeline in other case studies and real-time events. Also integrating other approaches such as deep learning

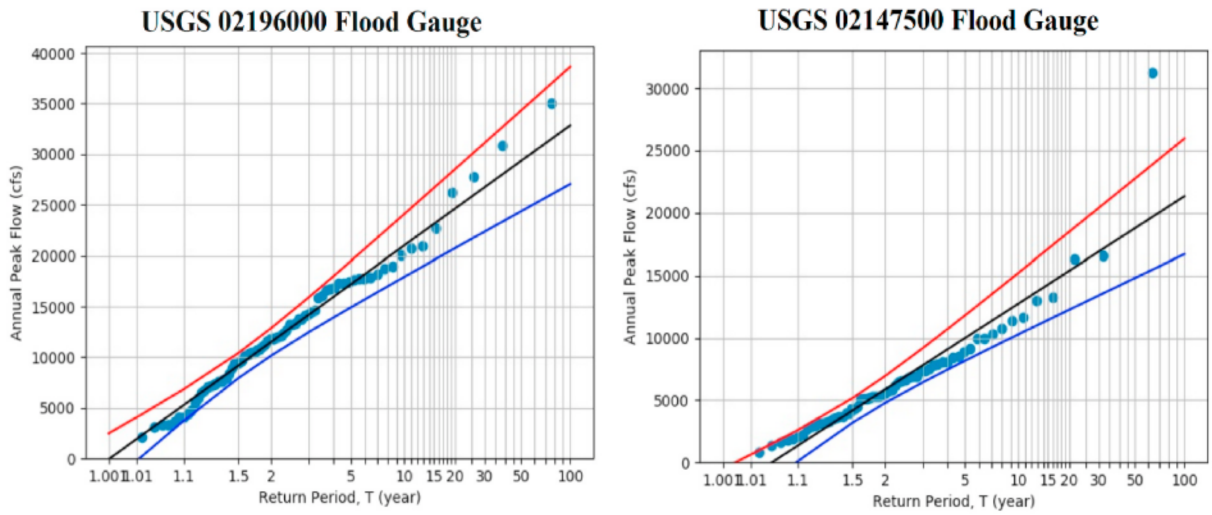


Fig. 10. Flood frequency analysis for the USGS 02147500 as well as USGS 02196000 using Gumbel distribution. Red and blue lines represent 95% confidence interval while black line represents fitted Gumbel distribution. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

techniques for word-embeddings natural language processing in deriving vector representations for non-English tweets. This would be crucial since we gathered tweets in Spanish and other languages when we tested the prototype operationally during Dorian flooding.

6. Limitation of FAIS Prototype and Future Work

This national-scale Big Data analytics pipeline could be further improved in several ways. In the FFA section, we recommend computing nonstationary flood frequency models by incorporating external covariates (López and Francés, 2013; Philips et al., 2018; Su and Chen, 2019). Meteorological modes such as El Niño-Southern Oscillation (ENSO) and/or physical properties (reservoir index) are key covariates that could be taken into account for non-stationary FFA assessment. In addition, compound flood calculation can be included to compute FFA based on multivariate extreme variables (see Renard et al., 2013) for the coastal region. In addition, FAIS workflow and design for identifying at-risk areas to flooding can be further improved. The size of bounding boxes of tweets (currently 16 km from the centroid) requires further attention and evaluation, given the boundary effect on the number of tweets intersect bounding boxes and yielding a match as well as considering the density of USGS flood gauges including ungauged areas. In an urban setting, increasing the size of bounding boxes would likely include more significant tweets thereby more accuracy on the chosen proximity. The selected tweets can be also overlaid with other data layers such as census data, population density, rainfall radar data, and Federal Emergency Management Agency (FEMA) flood hazard map.

In addition, FAIS image processing and machine learning approaches need further attension. Google Cloud Vision is a mature detection tool and comes with more flexible API conventions, multiple image formats, and a native batch support. Its Object Detection functionality generates much more relevant labels, and its label detection currently seems mature enough, as well. Although it's not quite perfect, yet. The biggest issue of this tool seems to be rotational invariance, although it might be transparently added to the deep learning model in the future. Vision has a wide margin of improvement regarding batch/video support and more advanced features such as image search, object localization, and object tracking in video. Being able to fetch external images (e.g. by URL) might help speed up API adoption, while improving the quality of flood detection features will inspire greater trust from users. For further FAIS development, a linking approach could be proposed to increase system autonomy during real time flood risk assessment and data retrieval tasks. This automation would help satisfy the need for timeliness and reliability during emergency responses and management. Efforts are currently underway to combine IoT sensor rainfall and water level information to the prototype.

7. Software and Data Availability

Flood Analytics Information System (FAIS): Various Scripts as they apply to pipeline development.

Description: Data Gathering application designed for flood and Twitter data analyses.

Developer: Nattapon Donratanapat.

Contact: pleuk5667@gmail.com.

Software Access: https://pypi.org/manage/project/fais/releases/

Year First Available: 2019.

Hardware required: Windows, Linux, MacOS.

Hardware required: Intel i3 or mid-performance PC with multicore processor and SSD main drive, 4 Gb memory recommended.

Cost: Free. Software and source code are released under the Massachusetts Institute of Technology License.

Software availability: All source code can be found at GitHub public repository as well as at the Python Package Index (PyPI).

https://github.com/VidyaSamadi/Flood-Analytics-Information-System-FAIS

https://pypi.org/project/fais/

https://floodanalytics.clemson.edu/

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.

7. Acknowledgements

This work was supported by the U.S. National Science Foundation (NSF) Directorate for Engineering under grant CBET 1901646. Any opinions, findings, and discussions expressed in this study are those of the authors and do not necessarily reflect the views of the NSF. The authors acknowledge the early and continuous discussion with South Carolina stakeholders, first responders and decision makers. FAIS prototype has been developed using cluster computing at the University of South Carolina as well as Clemson University; therefore, their support and assistant are gratefully acknowledged. FAIS Python package is

publicly available at the HHR (HydoSystem and Hydroinformatics Research) GitHub account (https://github.com/HHRClemson). The prototype is currently running on the IBM cloud computing service (https://floodanalytics.clemson.edu/).

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2020.104828.

References

Adams, T.E., Dymond, R., 2019. The effect of QPF on real-time deterministic hydrologic forecast uncertainty. J. Hydrometeorol. 20, 1687–1705. https://doi.org/10.1175/JHM-D-18-0202.1.

Al-Garadi, M.A., Mohamed, A., Al-Ali, A., Du, X., Guizani, M., 2018. A survey of machine and deep learning methods for Internet of Things (IoT) security [Online]. Available: https://arxiv.org/abs/1807.11023.

Barker, J.L.P., Macleod, C.J.A., 2019. Development of a national-scale real-time Twitter data mining pipeline for social geodata on the potential impacts of flooding on communities. Environ. Model. Software 115, 213–227.

de Albuquerque, J.P., Herfort, B., Brenning, A., Zipf, A., 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. Int. J. Geogr. Inf. Sci. 29 (4), 667–689.

De Assis, L.F.F.G., De Albuquerque, J.P., Herfort, B., Steiger, E., Horita, F.E.A., 2016. Geographical prioritization of social network messages in near real-time using sensor data streams: an application to floods Brazilian. Journal of Cartography 68, 1231–1240.

DeCicco, L.A., Hirsch, R.M., Lorenz, D., Watkins, W.D., 2018. DataRetrieval: R packages for discovering and retrieving water data available from Federal hydrologic web services. https://doi.org/10.5066/P9X4L3GE.

Eilander, D., Trambauer, P., Wagemaker, J., van Loenen, A., 2016. Harvesting social media for generation of near real-time flood maps. Procedia Eng 154, 176–183.

Ekta, P.K., Bundela, P., Dewan, R., 2017. Tweet analysis for real-time event detection and earthquake reporting system development. Int. Res. J. Eng. Technol. 3113–3119, 04 (05).

Erfani, M., Samadi, S., 2019. \. The American Geophysical Union, San Francisco, California, 9-13 December 2019.

Feng, Y., Sester, M., 2018. Extraction of pluvial flood relevant volunteered geographic information (VGI) by deep learning from user generated texts and photos. ISPRS Int. J. Geo-Inf. 7 (2), 39.

Grolinger, Katarina M., Capretz, E., Mezghani, E., Exposito, E., 2013. Knowledge as a service framework for disaster data management. In: WETICE '13 Proceedings of the 2013 Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 313–318.

Herfort, B., Schelhorn, S.J., Albuquerque, J.P.D., Zipf May, A., 2014. Does the Spatiotemporal Distribution of Tweets Match the Spatiotemporal Distribution of Flood Phenomena? A Study about the River Elbe Flood in June 2013 Proceedings of the ISCRAM 2014 Conference–Pennsylvania. USA.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., Cebrian, M., 2016. Rapid assessment of disaster damage using social media activity. Sci. Adv. 2 (3), e1500779.

Laylavi, F., Rajabifard, A., Kalantari, M., 2016. A multi-element approach to location inference of twitter: a case for emergency response. ISPRS Int. J. Geo-Inf. 5 (5), 56.

López, J., Francés, F., 2013. Non-stationary flood frequency analysis in continental Spanish rivers, using climate and reservoir indices as external covariates. Hydrol. Earth Syst. Sci. 17, 3189–3203. https://doi.org/10.5194/hess-17-3189-2013.

Middleton, S.E., Middleton, L., Modafferi, S., 2014. Real-time crisis mapping of natural disasters using social media. IEEE Intell. Syst. 29 (2), 9–17.

Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M., 2013. June. Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. In: ICWSM. June 2013.

Nativi, S., Mazzetti, P., Geller, G.N., 2013. Environmental model access and interoperability: the GEO model web initiative. Environ. Model. Software 39, 214–228.

Philips, R.C., Samadi, S., Meadows, M.E., 2018. How extreme was the October 2015 flood in the Carolinas? An assessment of flood frequency analysis and distribution tails. J. Hydrol. 562, 648–663. https://doi.org/10.1016/j.jhydrol.2018.05.035.

Rao, R., Plotnick, L., Hiltz, R., 2017. Supporting the Use of Social Media by Emergency Managers: Software Tools to Overcome Information Overload Proceedings of the 50th Hawaii International Conference on System Sciences. January 2017.

Raynal, J.A., Salas, J.D., 1986. Estimation procedures for the type-1 extreme value distribution. J. Hydrol. 87, 315–336.

Renard, B., Sun, X., Lang, M., 2013. Bayesian methods for non-stationary extreme value analysis. In: Extremes in a Changing Climate: Detection, Analysis and Uncertainty. Springer Netherlands, Dordrecht,, pp. 39–95.

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., Eisner, S., Müller Schmied, H., Sutanudjaja, E.H., Warrach-Sagi, K., Attinger, S., 2017. Toward seamless hydrologic predictions across spatial scales. Hydrol. Earth Syst. Sci. 21, 4323–4346. https://doi.org/10.5194/hess-21-4323-2017.

Smith, Luke, Liang, Q., James, P., Lin, W., 2015. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. J. Flood Risk Manage. 10, 370–380, 2015.

Spielhofer, T., Greenlaw, R., Markham, D., Hahne, A., 2016. December. Data mining Twitter during the UK floods: investigating the potential use of social media in emergency management. In: Information and Communication Technologies for Disaster Management (ICT-DM), 2016 3rd International Conference on. IEEE, pp. 1–6.

Starkey, E., Parkin, G., Birkinshaw, S., Large, A., Quinn, P., Gibson, C., 2017. Demonstrating the value of community-based ('citizen science') observations for catchment modelling and characterisation. J. Hydrol. 548, 801–817.

Su, C.J., Chen, X.H., 2019. Assessing the effects of reservoirs on extreme flows using nonstationary flood frequency models with the modified reservoir index as a covariate. Adv. Water Resour. 124, 29–40. https://doi.org/10.1016/j.advwatres.2018.12.004.

Tsou, M.H., Jung, C.T., Allen, C., Yang, J.A., Han, S.Y., Spitzberg, B.H., Dozier, J., 2017, July. Building a real-time geo-targeted event observation (geo) viewer for disaster management and situation awareness. In: International Cartographic Conference. Springer, Cham, pp. 85–98.

Vieweg, S., Castillo, C., Imran, M., 2014, November. Integrating social media communications into the rapid assessment of sudden onset disasters. In: International Conference on Social Informatics. Springer International Publishing, pp. 444–461.

Witherow, M.A., Cem, Sazara, Winter Irina, M., Arboleda, Mohamed I., Cetin Elbakary, Mecit, Iftekharuddin, Khan M., 2019. Floodwater detection on roadways from crowdsourced images. Comput. Methods Biomech. Biomed. Eng.: Imaging & Visualization 7 (5–6), 529–540. https://doi.org/10.1080/21681163.2018.1488223.