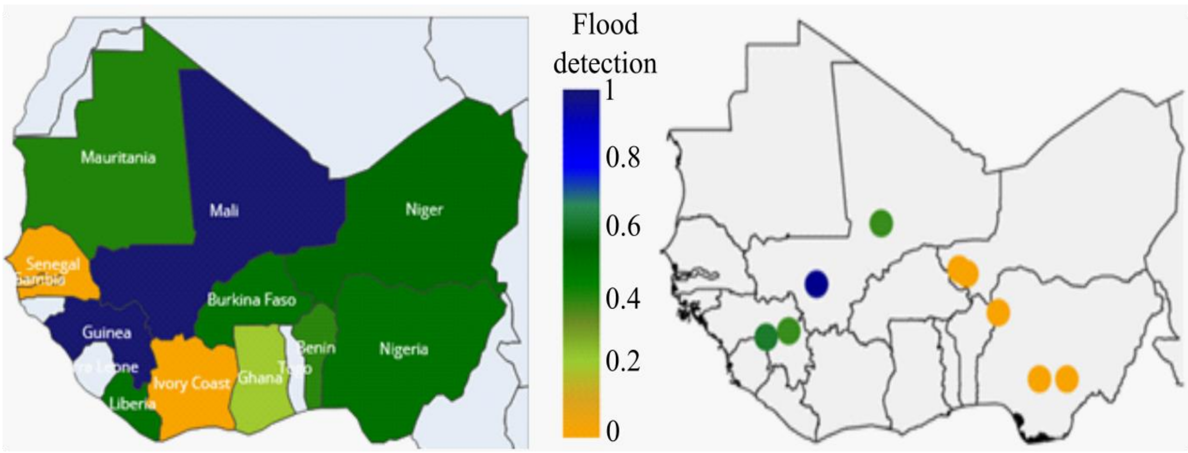
Review of Early Warning Dissemination in Media and Assessment of Flood Early Warning Systems with Media



A case study in West Africa

Master Thesis

Federal Institute of Technology Zurich (ETHZ)

Institute of Environmental Engineering (IfU)

&

Swiss Federal Institute of Aquatic Science and Technology (eawag)

Decision Analysis (DA)

6th of January 2021

Supervisors:

Prof. Dr. Peter Molnar, ETHZ, IfU

Dr. Martijn Kuller, eawag, DA

Dr. Darcy Molnar, ETHZ, IfU

Dr. Judit Lienert, eawag, DA

Author:

Thierry Hohmann

Abstract

Climate change is projected to exacerbate flood hazards in the future. Around 20% of all floods worldwide occur in Africa. The impact of floods can be mitigated with Flood Early Warning Systems (FEWS). Effective FEWSs need early warning dissemination and sound forecasting modeling. Researchers found that media, and recently also social media, are key stakeholders in warning dissemination. Furthermore, researchers investigated the potential of social media in assessing the accuracy or improving early warning dissemination in FEWS. However, little research on these two aspects has been conducted in West Africa. In this study, we aim to address the current warning dissemination in media and social media in West Africa. Furthermore, we investigate if social media can be used to assess the performance of FANFAR, a FEWS in West Africa predicting streamflow. We queried archives of newspapers, radio, and TV and tweets based on the flood event disaster database EMDAT and identified if early flood warnings are issued and whether they differ between events of different magnitude and between countries. Furthermore, we compared the daily number of tweets to the daily forecasted flood risk by FANFAR on a country scale. We employed a flood event detection algorithm developed by de Bruijn et al., 2019, which uses machine learning to identify flood-related tweets. Results show that radio and TV archives that can be systematically queried are, to the best of our knowledge, not available. However, the assessed newspapers often disseminate the warning that has been issued by a governmental meteorological agency. We found, that in Nigeria, more early warnings are issued in newspapers and tweets with an increase in people affected during a flood. We did not observe this trend in Ghana and could not find any early warnings issued in the Ivory Coast or Burkina Faso. Nevertheless, we found that tweets contain links with potential information about flood warning dissemination. Results also show that tweets indicate flooding reliably. However, we found that the assessment of FANFAR based on only flood-related tweets is not adequate. Additional factors, such as rainfall, or systematic lags between tweets and forecasts, could play a role. These factors should be determined on a regional scale. Mostly the influence of rainfall on tweets limits their usage in FEWSs that predict streamflow, such as FANFAR. Nonetheless, in combination with an authoritative dataset, such as streamflow, tweets could provide additional information. Our findings imply that media and social media can be used as a tool for warning dissemination, a source for warning dissemination research, and as additional data to support the assessment of FEWSs. Especially the projected increase of social media in the future will facilitate the improvement of early warning research and FEWSs.

Acknowledgments

Throughout the writing of my thesis, I have been fortunate to receive great support and assistance along the way.

I would first like to thank Peter and Darcy Molnar from ETHZ, who enabled me to travel to Ghana in course of the Network of Water and Life program. Besides the friends I made in this program, I learned valuable lessons about research for impact and research in low-income countries. This experience assisted me greatly in completing my thesis. Also, you have been more than helpful with your constructive criticism and feedback concerning my thesis.

I would further like to thank Judit Lienert and Martijn Kuller from eawag. You supported me with day-to-day supervision and valuable inputs regarding report writing, conceptual thinking, and presentation of results. Your shared knowledge will also help me in future projects. Furthermore, I would like to thank the Decision Analysis cluster and the whole Environmental Social Sciences department at eawag for questions and inputs during presentations I held.

Also, I would like to acknowledge the support I received from Jafet Andersson, Léonard Santos, and Lorna Little from the Swedish Meteorological and Hydrological Institute (SMHI). You provided me with the necessary data along the way and pitched ideas and inputs for my thesis.

Furthermore, I would like to thank Jens de Bruijn from the Vrije Universiteit Amsterdam. You provided neatly structured data indispensable for my thesis and were always available when I had questions.

I also would like to thank my parents Beatrice and Charles Hohmann for their emotional support during the troublesome time of COVID-19. Finally, I would like to acknowledge the support of my friends from different branches of studies for invigorating and fruitful discussion related to my thesis. Namely, Vlad Bunkin, who helped me a great deal with programming-related issues, and Maurice Grunder, who assisted me in proof-reading, played a vital part in the development of this thesis.

This thesis utilized data from the FANFAR project, which has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 780118.

Table of Contents

List of Acronyms

List of Tables

List of Figures

1. Chapter: Introduction

Water is indispensable for life. However, every year water-related disasters, such as floods, cause death, and billion-dollar losses all over the world (Perera et al., 2019). 23% of all floods hit Africa, which makes it the second most affected continent after Asia (Perera et al., 2019). In 2010, more than 1 million people were affected by floods in 20 African countries (Dessalegn and Akalu, 2015) and in 2015-2016 the continent faced the most severe El Niño in decades. Besides physical damage, floods can lead to an outbreak of infectious diseases, such as Malaria, and expose people to toxic substances (Dessalegn and Akalu, 2015). Furthermore, the situation in Sub-Saharan African (SSA) countries will likely be exacerbated in the future (van Niekerk and Nemakonde, 2017). The majority of SSA's disasters are of hydrometeorological origin and are expected to increase in frequency and magnitude due to climate change (van Niekerk and Nemakonde, 2017) and lead to unprecedented floods (Besada and Sewankambo, 2009). This high exposure to flood disasters in tandem with a low adaptive capacity makes Africa one of the most vulnerable continents (Dessalegn and Akalu, 2015), not least due to the rapid human settlement growth (Ahadzie and Proverbs, 2011).

In general, two types of measures are available to protect the population from flooding: Structural and non-structural measures. On the one hand, structural measures, such as dams or weirs, store runoff (Madamombe, 2004) or pose a physical barrier to the flood. Non-structural measures, on the other hand, include flood early warning systems (FEWS) and settlement policies (Madamombe, 2004). In general, an early warning system (EWS) should disseminate warning information to individuals so that they can respond to an imminent threat (UNISDR, 2009) and harm and economic losses can be mitigated (UNEP, 2012) . An increase in preparedness for disasters of the population decreases deaths (Gwimbi, 2007). FEWS are already operational in various countries, e.g. the European Flood Awareness System (EFAS) in continental Europe or the Flood Forecasting & Warning Service in Australia (Emerton et al., 2016). The UN Office for Disaster Risk Reduction advocates for the strengthening of FEWS (Perera et al., 2019), also in light of the sustainable development goals (SDGs). Yet there is a lack of adequate FEWS in low-income countries (UNEP, 2012).

Challenges in the implementation of FEWS include communication with stakeholders and the forecast model accuracy. Firstly, the communication from the FEWS derived information to the stakeholders is vital for the success of an EWS (UNEP, 2012), as the communication to those most at risk influences the effectiveness of natural hazard management (van Niekerk and Nemakonde, 2017). The information must be effectively passed down to the end-user via various media, like newspaper, television, and radio (UNEP, 2012). Therefore, media, are key stakeholders in warning dissemination (Ganiyu et al., 2017a). Nevertheless, the involvement of e.g. West African media in flooding has been criticized to be focused on the provision of relief measures rather than warning the population (Ahadzie and Proverbs, 2010; Ahadzie and Proverbs, 2011; Ganiyu et al., 2017a; Ganiyu et al., 2017b). Secondly, there is a trade-off between the warning lead-time and the prediction accuracy of a FEWS (UNEP, 2012). On the one hand, false alarms cause people to not take flood warnings seriously (Gwimbi, 2007; Madamombe, 2004). On the other hand, a flood without a preceding warning leaves the population unprepared. Therefore, the predictive capabilities of FEWSs need to be improved (UNEP, 2012; Perera et al., 2019).

Communication to reach potential flood victims needs improvement (Madamombe, 2004). Also, people must be educated at a local level on how to react to flood hazards (Madamombe, 2004). For example, newspapers in Nigeria feature disaster risk communication. However, around 80 % of the flood-related texts are not prominently placed and around 20 % of texts are published before an actual event (Ganiyu et al., 2016; Ganiyu et al., 2017a). In Ghana, around 3 % of news contains flood warnings and 8 % of news raises awareness on how to behave during a hazard (Ahadzie and Proverbs, 2011). Furthermore, social media is increasingly used as an information source for disasters (de Bruijn et al., 2017). Tweets related to flooding correlate with flood events in space and time (Fuchs et al., 2013) and can therefore be used to detect and locate floods (Rossi et al., 2018; de Bruijn et al., 2019). Tweets also proved helpful to predict areas at risk of flooding (Smith et al., 2015), augment flood risk information in the form of maps (Lorini et al., 2019), or serve as a proxy variable for rainfall (Restrepo-Estrada et al., 2018). Flickr also facilitates event detection with tags appended to images (Naoko Nitta and Babaguchi, 2014; Ling Chen, 2009). The use of social media in Africa is, except for Facebook, increasing (Statcounter, 2020).

We found that literature concerning flood warning communication is confined to a specific region and selected newspapers and that literature concerning FEWSs did not address the incorporation of social media in West Africa. We aim to close these research gaps by investigating the current warning dissemination in West African media and by analyzing how forecasts of a FEWS can be assessed with the help of social media. We broadened the spatial extent of news coverage and systematically searched online newspaper databases, radio and TV archives, and social media during flood events in the whole of West Africa. We then evaluated whether flood warning dissemination is included in the media to establish a baseline of the current practice of flood warning dissemination in West Africa. Furthermore, we compared the spatial, temporal, and quantitative occurrence of tweets and forecasted flood severity. We found that Twitter has the most instruments already available for our purpose (flood text classification, flood detection algorithm) compared to other social media. We investigated if tweets can be used to assess the performance of a FEWS in tandem with authoritative data series (rainfall and streamflow). We based our research on the FEWS *Reinforced cooperation to provide operational flood forecasting and alerts in West Africa* (FANFAR), which is currently being developed with local stakeholders and aims to provide flood risk forecasts 10 days ahead (SMHI, 2020b). Based on the research gap we identified, we aim to address the following research questions (RQs):

RQ1: How is flood risk communication currently disseminated in media (newspaper, radio, TV, and social media) in West Africa?

- *RQ1.1:* How does flood warning dissemination in media differ regarding content and channel between flood events of different magnitudes?
- *RQ1.2:* How does flood warning dissemination in media differ regarding content and channel between countries in West Africa?

RQ2: Can social media be used to assess the performance of FANFAR in West Africa?

- *RQ2.*1: How much does increased social media activity in a specific region in West Africa correspond to registered flood events from a flood event database?
- *RQ2.2*: How much does enhanced tweet activity correspond to an increased flood risk by FANFAR in a specific region in West Africa*?*
- *RQ2.3*: Can other media support the assessment of the performance of FANFAR?

2. Chapter: Literature Review

We reviewed both academic and grey literature. For academic literature, we used the online databases of Google Scholar and Web of Science and for grey literature Google. Firstly, we defined a set of keywords with combinations of the words displayed in Table 2-1, to find an initial body of literature. Secondly, we applied the snowballing procedure, including forward and backward snowballing, as described by Wohlin, 2014. Using forward snowballing, we identified papers based on citations in the paper currently investigated. Backward snowballing, in contrast, is conducted by investigating the list of references attached to the document under consideration and identifying potentially suitable literature. Finally, we directly looked up the authors.

*Table 2-1: Research questions (RQ) and the corresponding search terms to find an initial body of literature. The research questions are specified in subsequent sections. A * indicates a wildcard*

RQ	**Keywords**
RQ1	Africa, emergency risk communication, flood*, flood early warning system, flood warning dissemination Africa, flood warning Africa, implementation, evaluat* media reduction flood risk, media analys*, methodology media analy*, news fram*, newspaper analys*, print media flood risk communication, warn*
RQ2	classification, flood*, tweet, Twitter, Facebook, Flikr, social media, online scrap* algorithm, event detect*, detect* algorithm

2.1 FEWS Warning Dissemination in West African Media

2.1.1 Early Warning Dissemination in FEWS

Definition of an EWS and FEWS

An EWS can be understood as "an integrated system of hazard monitoring, forecasting and prediction, disaster risk assessment, *communication* and preparedness activities systems and processes that enable individuals, communities, governments, businesses and others to take timely actions to reduce disaster risks in advance of hazardous events" (UNISDR, 2017, p. 17), or as a "[...] set of capacities needed to generate and *disseminate* timely and meaningful warning information to enable individuals, communities and organizations threatened by a hazard to prepare and to act appropriately and in sufficient time to reduce the possibility of harm or loss." (UNISDR, 2009, p. 12). A FEWS is a type of EWS and consists of "*Provision* of specific forecasts relating to rainfall for both quantity and timing, for which numerical weather-prediction models are necessary" (WMO, 2011, p. 1-6), an "Establishment of a network of manual or automatic hydrometric stations, linked to a central control by some form of telemetry" (WMO, 2011, p. 1 - 6) and "model software, linked to the observing network and operating in real time" (WMO, 2011, p. 1 - 6). However, the terminology around EWSs is not consistent in the literature (UNDRR, 2006). Examples of FEWSs include the European Flood Awareness System (EFAS) in continental Europe, the Flood Forecasting & Warning Service in Australia (Emerton et al., 2016), and FANFAR in West Africa. Note, that in the three EWS/FEWS definitions, the *communication*, *dissemination,* and *provision* of risk are highlighted.

Risk communication in FEWSs

Risk communication is a vital part of any EWS, as a scientifically sound forecast will contribute little to mitigate the effects of a hazard if the communication is not understood (Baudoin et al., 2014, UNEP, 2012). Risk communication is concerned with informing the decision-makers, experts, and the affected public about a risk situation (Höppner et al., 2010; Rollason et al., 2018). Early warning is a subset of risk communication (Figure 2-1). It is the communication step between the forecast itself and the subsequent actions taken by e.g. authorities or local communities (Cools et al., 2016). Warning dissemination typically starts with scientific sources that channel warnings through government decision-makers and the media (UNDRR, 2006). The information should reach multiple actors, including emergency services, operators of utilities, economic service providers, and others, who themselves may act as onward disseminators (UNDRR, 2006). Warnings can be disseminated to a targeted or broad audience on different levels (e.g. local or regional) and via various channels, like newspaper, radio, and TV (UNEP, 2012) and formats, like visual or printed (Macnamara, 2005). The warning message content needs to be tailored to the recipients, such as non-governmental organizations (NGO)s or villagers (Ping et al., 2016; Seeger et al., 2018) and usually takes the form of "top-down" communication (Perera et al., 2019). Also, the timeliness of warning dissemination influences the effectiveness of early warning messages (Seeger et al., 2018). In a nutshell, warning messages are characterized through content, channel, format, and timing and media are key stakeholders in warning dissemination (Ganiyu et al., 2017a).

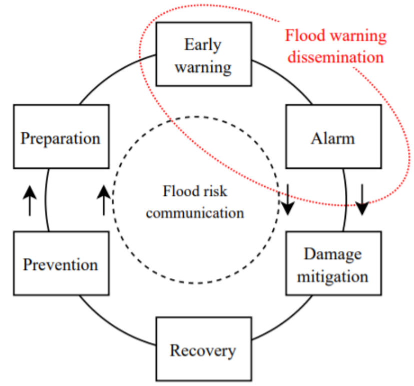


Figure 2-1: Disaster management cycle (black, solid), risk communication cycle (black, dashed), and its subset flood warning dissemination (red, dotted). Adapted from Ping et al., 2016.

The role of media in early warning dissemination

Traditional media, such as television, radio, and newspaper, are commonly used to disseminate emergency information (Dessalegn and Akalu, 2015; Perera et al., 2019; Ruggiero and Vos, 2014; Seeger et al., 2018), and play a crucial role in communicating natural hazards (Höppner et al., 2010; UNDRR, 2006). Also, social media gained importance in disseminating warning messages (Castillo, 2016; Imran et al., 2015; Palen and Anderson, 2016; Smith et al., 2015), and creating situational awareness (Vieweg, 2012). Media can be investigated using content analysis, which entails analyzing text, ranging from interview transcripts to newspapers, television, or advertisement (Macnamara, 2005).

2.1.2 Content Analysis of Flood Related Issues in West African Media

Content analysis of West African newspapers showed that in Nigeria, out of 500 stories about flooding during 2012, predominantly in the form of news, around 20 % were published before an event (Ganiyu et al., 2016; Ganiyu et al., 2017a). Similarly, out of 167 issues from 20 different newspapers in Ghana between 2000 and 2010, around 3 % featured warning messages, and 7 % were concerned with educating the population about floods (Ahadzie and Proverbs, 2011). The other articles mostly reported on causes of floods or relief items provided by governmental and non-governmental organizations, likely due to political incentives (Ahadzie and Proverbs, 2011), i.e. the government wants to show that it cares about the population. The content does not only vary between newspapers, but also between the narrators crafting the news content (government, NGOs, local communities). In Nigeria for example, governmental agencies and local communities tend to blame the cause of flooding on humans for not obeying the law, while NGOs see the cause of flooding in climate change and human factors (Adekola and Lamond, 2017).

We found that the importance of media in warning dissemination is acknowledged by the scientific community. However, we discovered that research mostly focuses on overarching topics such as governance of natural hazards or flood risk practices. Little is known about flood warning dissemination through media, especially in West Africa. Newspaper content analysis is limited on the most prominently circulated issues in West Africa in a few countries and to the best of our knowledge, no research has been conducted on flood warning dissemination in newspapers, radio, TV, and social media. Furthermore, the relationship between flood event magnitude and warning dissemination has not been addressed. We aim to close this research gap by analyzing flood early warning dissemination based on flood events in West Africa.

2.2 Incorporating Social Media in FEWSs

2.2.1 Twitter in Flood Forecasting

Social media and hazard detection

The World Wide Web allows millions of people to access a plethora of online information worldwide (UNEP, 2012). In recent years, "social media, and in particular Twitter, has gained traction as a novel source of information on disaster events" (de Bruijn et al., 2017, p. 1), and in event detection (de Bruijn et al., 2017; Musaev et al., 2015; Rossi et al., 2018). Information disseminated via social media can improve circumstantial awareness in a crisis (Vieweg, 2012). Therefore, services employing Twitter have been developed to create awareness during a disaster. For example, the Emergency Situation Awareness (ESA) increases situational awareness by displaying hazard-relevant tweets on a map, especially for earthquakes (Yin et al., 2012). Another example is the Artificial Intelligence for Disaster Response (Lucas et al., 2014) of harvesting hazard-related tweets to increase awareness of danger. Usually, these services consist of an overview of crisis-related social media and its temporal and spatial characteristics (Imran et al., 2015). Although other social media (e.g. Flickr) were used for general event detection, such as "housewarming" (Ling Chen, 2009) or "fireworks displays" (Naoko Nitta and Babaguchi, 2014), we found that mostly Twitter was used for hazard detection management.

Twitter and flood hazards

Twitter is a microblogging platform where users can post messages containing up to 240 characters, images, and videos and it is the most studied social media platform in hazard management (Simon et al., 2015). Its popularity roots in the user-friendly Application Programming Interface (API), which offers a free extraction of data (Imran et al., 2015; Rossi et al., 2018). The Twitter API allows to programmatically obtain and analyze data and chat on Twitter (Twitter, n.d.). Twitter has also been studied in flood risk research, since recent advances in e.g., smartphones can enhance flood forecasting and monitoring with "big data" (Horita et al., 2017) and increase the spatial extent where data is available (Restrepo-Estrada et al., 2018). For example, from 5 million tweets collected over eight months in Germany, a correlation was found between flooded locations and the location where flood-related tweets were posted (de Albuquerque et al., 2015; Fuchs et al., 2013). On a global scale, tweets reporting flood hazards coincided with the timing and location of the hazard (de Bruijn et al., 2019). It is therefore suggested that tweets (and potentially other social media such as Flickr, YouTube, Facebook) can serve as a complementary source of information (Fuchs et al., 2013; Rossi et al., 2018) for flood monitoring and forecasting (de Bruijn et al., 2019; Lorini et al., 2019; Rossi et al., 2018; Smith et al., 2015).

Twitter and its application to FEWSs

We identified two main uses of social media in the context of FEWSs: (1) Social media can be used reactively to infer where a hazard is taking place and how severe it is (de Bruijn et al., 2019; Lorini et al., 2019; Rossi et al., 2018) and (2) social media can be used to calibrate, validate, and improve hydrodynamic models (Restrepo-Estrada et al., 2018; Smith et al., 2015).

(1) Inference on hazards: Twitter can be incorporated into FEWSs as an additional service. Tweets are collected either when a region is deemed prone to flooding (Lorini et al., 2019, Rossi et al., 2018) or when a threshold of tweet occurrence is reached (Smith et al., 2015). The collected tweets are then used in FEWSs for flood risk communication using maps (Lorini et al., 2019), in warning messages (Rossi et al., 2018) and to identify flood risk areas (Smith et al., 2015). Tweets have proven to be well in line with the reports of emergency personnel (Rossi et al., 2018) or increased news coverage (Fuchs et al., 2013).

(2) Model validation: Social media help in forecast validation. Again, flood-related social media posts are stored and subsequently compared to model results or registered events. For example, Twitter was employed to select a suitable hydrodynamic model in the United Kingdom, based on the criteria depth and velocity that could be extracted from collected tweets (Smith et al., 2015). Also, floods can be located (i.e. when and where a flood is taking place) with Twitter. 88 million tweets in 12 languages were collected globally and only the flood-related tweets were kept (de Bruijn et al., 2019). Results show that floods were detected that were not registered by flood databases (de Bruijn et al., 2019). Furthermore, tweets related to rainfall were used as a proxy variable for rainfall to generate a non-authoritative (i.e. not measured remotely or on the ground) data series in São Paulo (Restrepo-Estrada et al., 2018). The resulting data series were then used jointly with authoritative data series to calibrate a hydrological model (Restrepo-Estrada et al., 2018). This lead to improvements in estimating streamflow compared to only employing authoritative data (Restrepo-Estrada et al., 2018). Similarly, flood forecasts improved with the inclusion of crowd-sourced water levels together with static sensors (Mazzoleni et al., 2017). Tweets can also help to close measurement gaps due to malfunctioning measurement devices or to reduce inconsistencies in data series (Restrepo-Estrada et al., 2018). Regardless of the application of Twitter in FEWSs, flood events need to be detected based on tweets.

2.2.2 Flood Event Detection with Twitter

Three steps are required to detect events based on Twitter: (1) Tweets need to be geo-located to conclude their spatial occurrence, (2) tweets need to be filtered, based on the application (in our case tweets need to be flood-related), and (3) a criterion for the identification of events (in our case flooding) must be defined.

(1) Geolocation

A Twitter user can decide if they allow attaching exact geographic information to their tweet (Twitter, 2009) in the form of GPS coordinates (Twitter, 2020b) which is the case for around 2 % of all tweets (Leetaru et al., 2013). Besides, the user can refine the precise geographic information by tagging a location like a city or neighborhood (Twitter, 2020a). Tweets without GPS coordinates or a shared location need additional steps for a location assignment. Assigning a location to a tweet is referred to as *geotagging* (de Bruijn et al., 2017; Imran et al., 2015) and can be achieved by either using the specified GPS coordinates or by applying geolocation algorithms. Most geolocation algorithms use *toponym recognition*, which extracts a location from the text of a tweet (e.g. “Accra” from “Heavy floods hit Accra”), and *toponym resolution*, which rules out ambiguities (e.g. multiple places in the world are called London) (de Bruijn et al., 2017, Imran et al., 2015). Toponym resolution and recognition require a pre-defined set of locations, called a *gazetteer* (de Bruijn et al., 2017), to which the extracted location is compared. However, gazetteers are usually limited to unambiguous places with a large population (Amitay et al., 2004; Schulz et al., 2013), i.e. only Paris, the capital of France, would be listed. Named Entity Recognition (NER) is a more sophisticated approach, which considers grammar and context (Al-Rfou et al., 2015) to extract locations. However, the common and error-prone language used on Twitter poses a problem for NER (Li et al., 2012), since misspelled words might not be recognized. To mitigate the negative effect of a limited gazetteer on the geolocation process, further information provided in a tweet can be extracted, such as the location field, referenced websites, and the time zone (Schulz et al., 2013). de Bruijn et al., 2017 developed the Toponym-Based Algorithm for Grouped Geoparsing of Social Media (TAGGS) based on Schulz et al., 2013, which employs this additional information.

(2) Classification

It is vital to identify only the hazard-related tweets that contain useful information for further application, e.g. to identify flood events (Imran et al., 2015). The approaches for classifying tweets range from simple keyword filtering to advanced machine learning methods (MLMs). MLMs are mathematical models that can learn to recognize patterns (e.g. in the text) on their own (Keijsers, 2010). Simple keyword filtering, which allows selecting tweets that must contain a substring (e.g. “flood” or “flooding”), was applied by multiple authors to find flood-relevant tweets (Fuchs et al., 2013, Restrepo-Estrada et al., 2018). However, simple filtering methods, such as using keywords, will classify the sentence “my mail is flooded with spam” as flood-related (Lorini et al., 2019). Such false positives can be mitigated with classifying algorithms that were trained on an example set of flood-related tweets. Common practices to collect tweets are described in Imran et al., 2015. A newer approach incorporating MLMs was developed by de Bruijn et al., 2019. The authors also included hydrologic data and managed to improve the precision and recall of their classification algorithm (de Bruijn et al., 2020). Besides the classification of a tweet in a single language, experiments to assess the transferability of an algorithm to new languages were conducted. For example, Lorini et al., 2019 provided a method to apply a classifying algorithm to a new language, and de Bruijn et al., 2019 collected and classified tweets in 12 languages. After tweet classification, we can assess whether they are related to an event.

(3) Event detection

We found that most event detection approaches are based on keyword bursts. However, other methods and domain-specific approaches have also been developed (Imran et al., 2015). Events can be detected under the assumption that an increase in usage of a word in tweets indicates a new event (Imran et al., 2015). For example, authors compared the actual and expected or historical probability or frequency of a word in a tweet occurring during a fixed time window. If the actual probability of frequency surpasses the expected or historic one, a word is deemed to be “bursty” (Power et al., 2014; Yin et al., 2015), and an event is assumed to take place. Also, tweet bursts can be detected by searching tweets sequentially and not in a fixed time window, where a suspected burst is rejected or confirmed based on subsequently posted tweets (SEQAVG, see e.g. Riley, 2008). Further, the criterion when a word is “bursty” can be adapted to the time of day (people likely tweet less at night than during the day) and the general tweet activity in a country (de Bruijn et al., 2019). The use of Twitter in event detection has, however, its drawbacks.

Drawbacks of Twitter in Hazard Detection

Drawbacks to using Twitter as a data source include that reporting a hazard can occur away from the hazard location (Fuchs et al., 2013) or that information useful for inclusion into modeling is usually disseminated after the event (Smith et al., 2015), making harvesting of tweets during an event impossible. Furthermore, only using one form of social media is limiting and does not provide a complete view of a crisis (Imran et al., 2015). Besides, the large number of tweets that is not related to a hazard needs to be filtered before they can be applied to hazard management (Fuchs et al., 2013; Li et al., 2012), and literals and jargon make it difficult for some algorithms to classify tweets correctly (de Bruijn et al., 2017; Fuchs et al., 2013; Li et al., 2012; Yin et al., 2015). Finally, only 2% of tweets are geotagged (Leetaru et al., 2013), calling for algorithms necessary to identify the correct location of the tweets.

2.2.3 Assessment of FEWSs in West Africa

We discovered that social media and especially Twitter have been used in EWSs to increase situational awareness during a crisis, to detect hazardous events, and to improve hydrologic models. However, we found that research in that area was predominantly conducted in Europe. Little research has been conducted in low-income countries. Furthermore, to the best of our knowledge, social media has not yet been used to assess the performance of a FEWS. We aim to close this research gap by investigating if social media can be used to assess the FEWS FANFAR.

3. Chapter: Methods

3.1 Case Study

3.1.1 FANFAR

FANFAR is a FEWS and an EU funded Horizon 2020 project that aims to provide short- and medium-term hydrological forecasts of up to 10 days to key stakeholders (SMHI, 2020a). The forecast system is designed to be reliable and robust and to disseminate relevant information at the right time (SMHI, 2020a). The FANFAR project consists of the Information and Communications Technology (ICT), the decision analysis dimension, and the system sustainability dimension (Figure A-1 in the Appendix). The ICT dimension is concerned with adapting current and developing new ICTs in collaboration with stakeholders (SMHI, 2020b). The behavior and decision dimension investigates the stakeholders' preferences and behavior to enhance the decision-making and technology adoption processes (SMHI, 2020b). Finally, the sustainability dimension scrutinizes the inclusion of West African institutions, the long-term challenges and constraints, and opportunities (SMHI, 2020b). Inside the ICT dimension lays the forecasting and alert system (Figure 3-1). River flow and earth observations are fed into the hydrological model. With the help of meteorological analysis, the past and the initial state of the modeling domain are determined. The initial state is then used to generate future conditions. Based on past and future conditions, a risk level in the form of a return period (2 years, 5 years, or 30 years) is determined and communicated to the stakeholders.

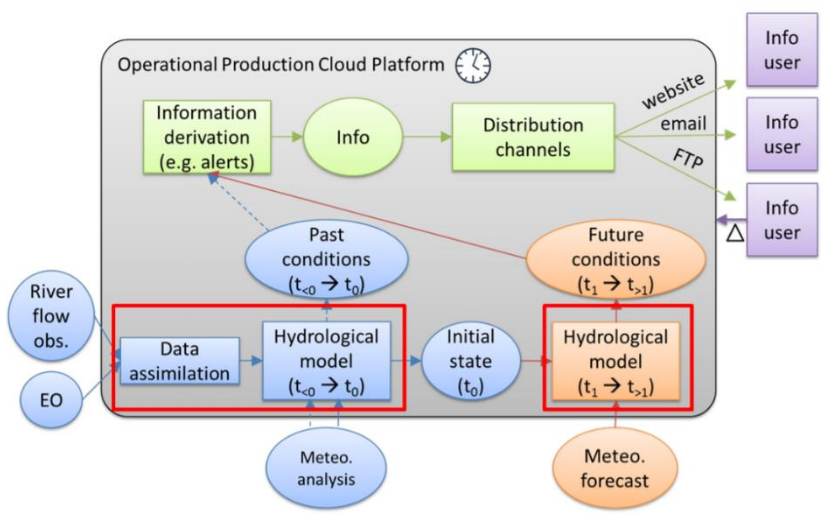


Figure 3-1: Components and data flow of the forecasting system within the Information and Communications Technology (ICT) dimension of FANFAR, taken from Andersson, 2020a. The red squares contain the hydrological modeling components. The blue shapes are data leading up to the initial state to model the future conditions (orange shapes). The green shapes show the dissemination of the result to the stakeholders (purple shapes). EO: Earth observation.

3.1.2 Forecast Model

It is crucial to understand the forecast model behind any FEWS to understand its strengths and weaknesses and include them in the FEWS evaluation. The subsequent information is, if not specified otherwise, taken from Andersson, 2020a. In the context of FANFAR, the hydrological model predicts the effects of meteorology on the *river flow*, *water level*, and *soil moisture content*. A suitable model for FANFAR should fulfill three points: Firstly, the model should already be operational somewhere else. Secondly, the input files should be updated daily, and calibration and enhancing catchment delineation should be conducted if seen appropriate. Thirdly, the model should have adequate performance on a daily scale. Therefore, Andersson, 2020a, deemed the model Hydrological Predictions for the Environment (HYPE) developed by the Swedish Meteorological and Hydrological Institute (SMHI) appropriate to be used in the FANFAR context.

3.1.3 The HYPE Model

General

HYPE is a semi-distributed hydrological model mainly used to simulate water flows and quality (SMHI, 2018). It is driven by temperature and rainfall data and its domains are organized in the form of sub-catchments (SCs) or sub-basins that can be attributed to hydrological response units (HRUs), e.g. land use, soil type, and elevation (Lindström, 2017, Figure 3-2). The HRUs can be structured with a maximum of three soil layers with a specified thickness (see Lindström, 2017 for details). Multiple SCs then build a river catchment (Figure 3-2).

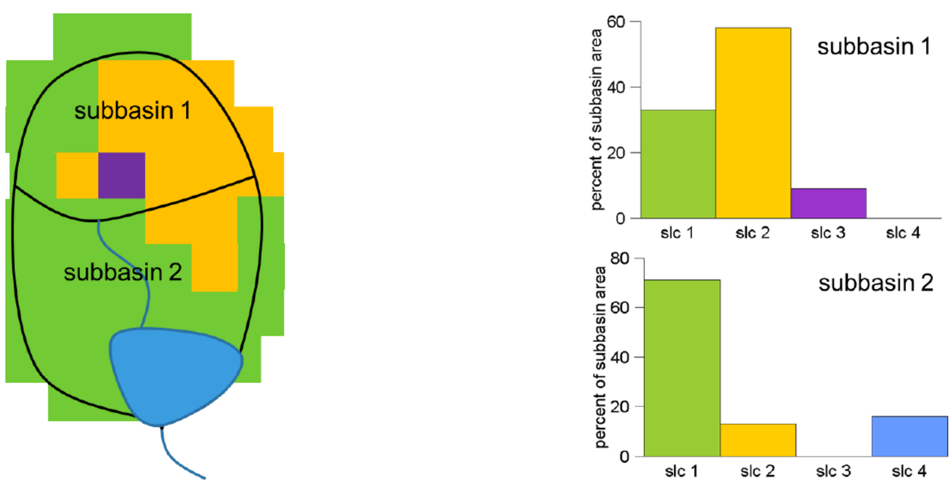


Figure 3-2: Example of a catchment divided into sub-catchments (SCs) or sub-basins (left) and the distribution of the corresponding HRUs (here named slc) for every SC (right). Taken from SMHI, 2018.

Elements of a sub-catchment

Horizontal components: The water moves, after falling on land, through the different storages (Figure 3-3). Firstly, a fraction *f* of the runoff Q_{land} from the land storage is routed to the wetland, whilst the rest is routed directly to a local stream. Again, the flow is diverted, and one part *g* of the flow Q_{Stream} is routed via an internal lake, whilst the other part $(1-g)*Q_{Stream}$ of the local stream runoff flows directly into the main stream. The main stream exchanges water bi-directionally via a floodplain. Furthermore, the inflow from the upstream SC is added. Finally, the water flows into a lake or wetland and is routed to the next downstream SC.

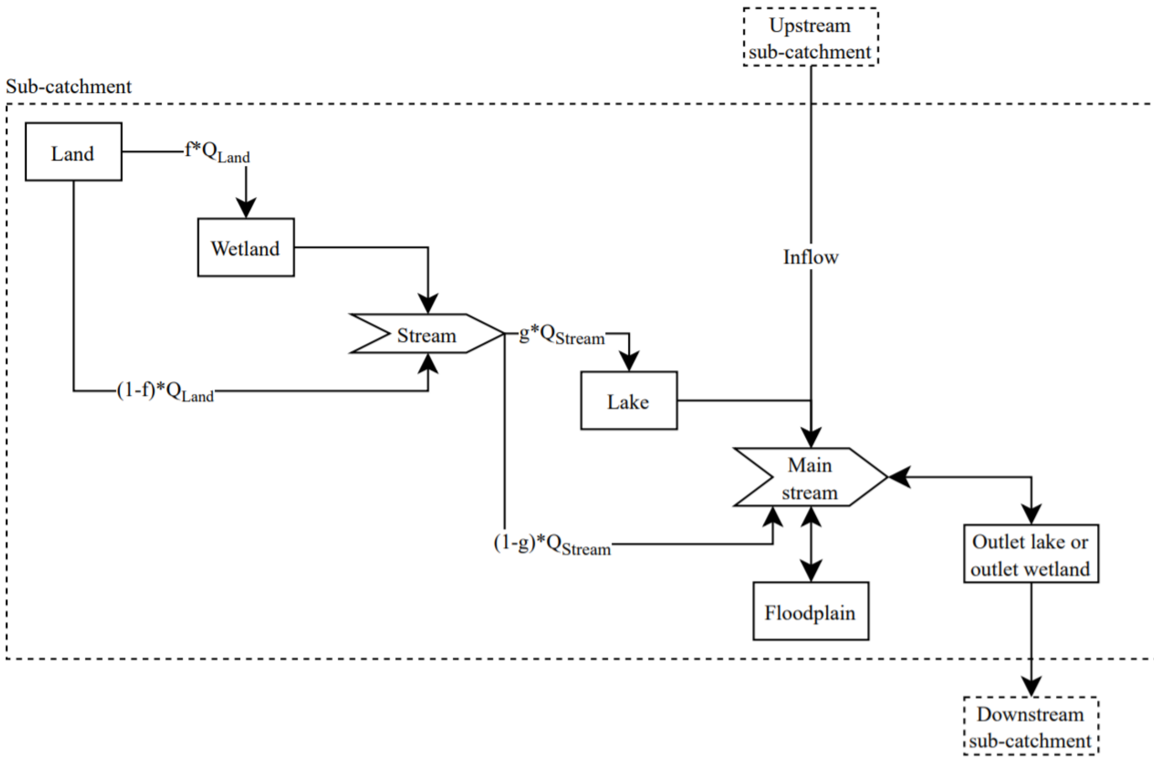


Figure 3-3: Horizontal elements within a sub-catchment (SC) modeling flow routing. Storages are displayed as rectangles and arrows. Q_{Land} - water from overland flow, f - fraction of overland flow reaching wetland, Q_{Stream} - water from local flow, g - fraction of local flow reaching lake. Adapted from SMHI, 2020d.

Vertical components: Every storage (Figure A-2 in the Appendix) is explained in detail in SMHI, 2020d. In general, only the land-storage can contain multiple components, such as snow, glacier, and up to three soil layer components. Processes such as precipitation, evapotranspiration, surface runoff, etc. influence the water balance. All other storages are single component storages and their water balance is dependent on inflows, outflows, or water abstractions, amongst others. The detailed formulas for evapotranspiration, temperature, and other adjustments are provided in SMHI, 2020c.

The adaptation of HYPE to West Africa

Andersson, 2020a adapted the HYPE model for the usage in West Africa, as described in the following. Two versions of HYPE are currently used in the context of FANFAR: The World Wide HYPE model (WWH) and the Niger-HYPE model. We focus on WWH since we used it in our analysis. However, the Niger-HYPE model is included in the subsequent explanation for the sake of completeness. WWH covers a larger domain than Niger-HYPE (Figure 3-4) and contains more SCs at a finer spatial resolution (Table 3-1).

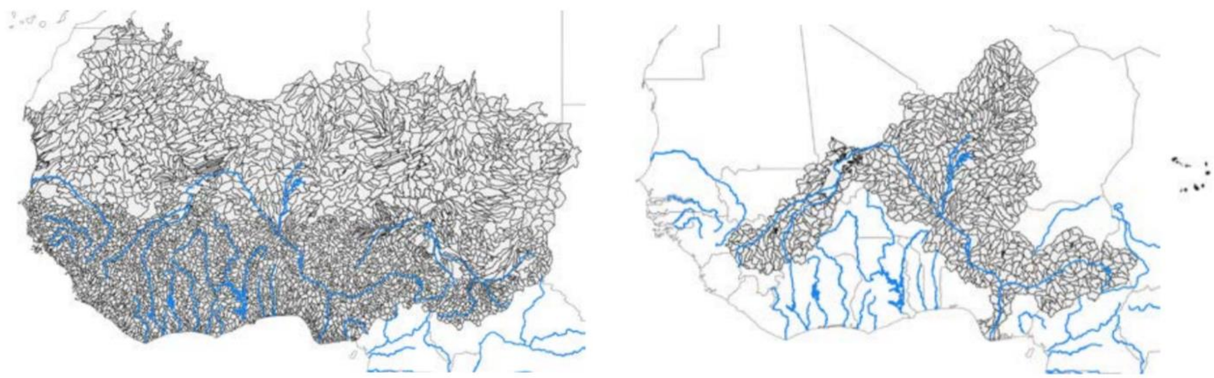


Figure 3-4:Catchment delineation of the World Wide Hype (WWH, left) and the Niger-HYPE (right) before the adaptation to West Africa. Taken from Andersson, 2020a.

Table 3-1: Characteristics of the WWH and the Niger-HYPE sub-catchments (SCs). Adapted from Andersson, 2020a.

Model	**WWH**	**Niger-HYPE**
Domain	West African hydrological basins (8.6 million km^2)	Niger River basin (2.1 million km^2)
Number of SCs [-]	4581	803
Average SC size [km^2]	1870	2619

FANFAR requires a daily forecast production to warn the stakeholders concerning future threats. WWH was evaluated against 106 streamflow gauges in West Africa and yielded a Kling-Gupta Efficiency (KGE) of -0.1 on a daily scale (Equation 1). KGE is, like the Nash-Sutcliffe efficiency, a metric in hydrology to summarize model performance, which is increasingly used in model performance assessment (Knoben et al., 2019).

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2}$$

Equation 1: KGE - Kling-Gupta Efficiency, r - linear correlation between observation and simulation, $\sigma_{sim/obs}$ – standard deviation of measured/observed streamflow, $\mu_{sim/obs}$ – mean of simulated/observed streamflow. Taken from Knoben et al., 2019.

A KGE of -0.1 was regarded as unsatisfactory (1 is the best value KGE can achieve, meaning that the simulations exactly match the observations). Therefore, Andersson, 2020a, undertook five major changes. (1) The SC delineation had to be enhanced. Besides, four manipulations on the hydrological model were undertaken to improve performance: (2) The global meteorological forcing data, (3) the West African weather data, (4) the hydrological model processes, and (5) the assimilation of observations.

(1) Basin delineation: In a first calibration step, the basin delineation of WWH was adjusted to match 97 measurement stations located in West Africa. The number of sub-basins increased to 4609 and the average SC size decreased to 1858 km^2.

(2) Global meteorological forcing data: A first improvement to the WWH model in the West African context was achieved by applying Hydrological Global Forcing Data (HydroGFD) 3.0 instead of HydroGFD 2.0 data, which HYPE currently uses. HydroGFD is a merged data set consisting of historical precipitation and temperature based on meteorological reanalysis and global observations (SMHI, 2019). It aims to globally reproduce the observed climate as accurately as possible. On a daily time scale, the results depend on the quality of the reanalysis, and possibly not all events can be captured (SMHI, 2019). The main difference between HydroGFD 2.0 and HydroGFD 3.0 (now used in WWH in the West African context), is the increase in spatial resolution from 0.5 to 0.25 degrees. The change in global forcing data improved the KGE from -0.1 to approximately 0.25.

(3) West African weather data: HydroGFD data can be improved with inputs from additional meteorological data. In the case of FANFAR, data from the AGRHYMET Regional Centre is used. AGRHYMET is an institution of nine member states and its main task is to improve natural resources management and food security in the Sahelian region, see (UN-SPIDER, 2020) for details. AGRHYMET combines satellite and ground data from West African stations. The potential of the merged AGRHYMET and HydroGFD data has yet to be established, as the calibration of the model on the merged data is still in progress.

(4) Hydrological model processes: The hydrological processes of WWH were altered to fit better into the West African domain. Firstly, the evapotranspiration and recession coefficients of the subsurface flow were adjusted. Secondly, the flood plains and reservoir behavior were adjusted. These two adjustments decreased the overreaction of WWH and matched the simulated to the observed peak flows better. The two calibration steps lead to an improvement in KGE at almost all 151 streamflow gauges considered, with more than 2/3 of the stations displaying a KGE between 0 and 1.

(5) Assimilation of observations: Hydrological systems have a memory; past conditions influence future ones. The assimilation of observations aims to combine numerical models with observations and has the potential to improve hydrological forecasts. In FANFAR, two kinds of observations are used: (1) locally measured streamflow and water level data, and (2) remotely measured water levels are used to update the past conditions of HYPE. (1) Wherever available, simulated discharges are replaced with observed ones. If no observation is available, a calibrated auto-regressive component is applied and the "observed" streamflow is calculated from the last known model value. This potentially improves forecasts at and downstream of gauged locations. (2) Satellite-based water levels can be provided even in remote locations. In FANFAR, the remotely measured water levels will be used to firstly generate a rating curve between simulated streamflow and observed altimetry. The rating curves can then be used to deduce discharges that can be fed into the auto-regressive component as described above. This approach of applying a rating curve to the simulated discharge at a certain location provided adequate results for water levels, especially above the warning threshold.

The Accuracy of WWH

It is vital to assess the accuracy of a forecasting model to gain realistic expectations for further applications, e.g. comparing it to the occurrence of social data. The following information is taken, if not specified otherwise, from Andersson, 2020a. Since FANFAR aims to produce flood forecasts for up to 10 days to warn the population about threats by flooding in the form of return periods of streamflow, the measured and simulated severity must be compared rather than the actual discharges to assess the accuracy of WWH. The accuracy of WWH was assessed by the probability of flood event detection (POD) score and the false alarm ratio (FAR) score (Equation 2 and Equation 3 respectively). As mentioned before, the model was assessed considering return periods. These observed and simulated return periods of the discharge RQ_{obs} and RQ_{sim} were deduced from long-term model simulations and measured data respectively. An event is observed if RQ_{obs} is larger than a threshold $RQ_{obs,\Theta}$. Likewise, an event is considered forecasted if RQ_{sim} is larger than a threshold $RQ_{sim,\Theta}$. Note that simulations can only be compared with return periods deduced from simulations. The same is valid for observations. Therefore, the forecasting system is accurate if the forecasted streamflow is above its simulation-based return-period threshold, in tandem with the observed streamflow being above its observation-based return-period threshold. This analysis was conducted for a simulation period of 2017-2019 with WWH and HydroGFD 3.0. The simulation results were compared to nine gauged locations (Figure 3-5). In the Upper Niger, events are detected well (POD close to 1) but also many false alarms are issued (FAR close to 0). The opposite was found for the Middle and Lower Niger region.

$$POD = \frac{Hits}{Hits + Misses}$$

Equation 2: POD - Probability of detection, Hits - true positives, Misses - false negatives

$$FAR = \frac{False\ alarms}{False\ alarms + Hits}$$

Equation 3: FAR - False alarm ration, False alarms - false positives, Hits - true positives

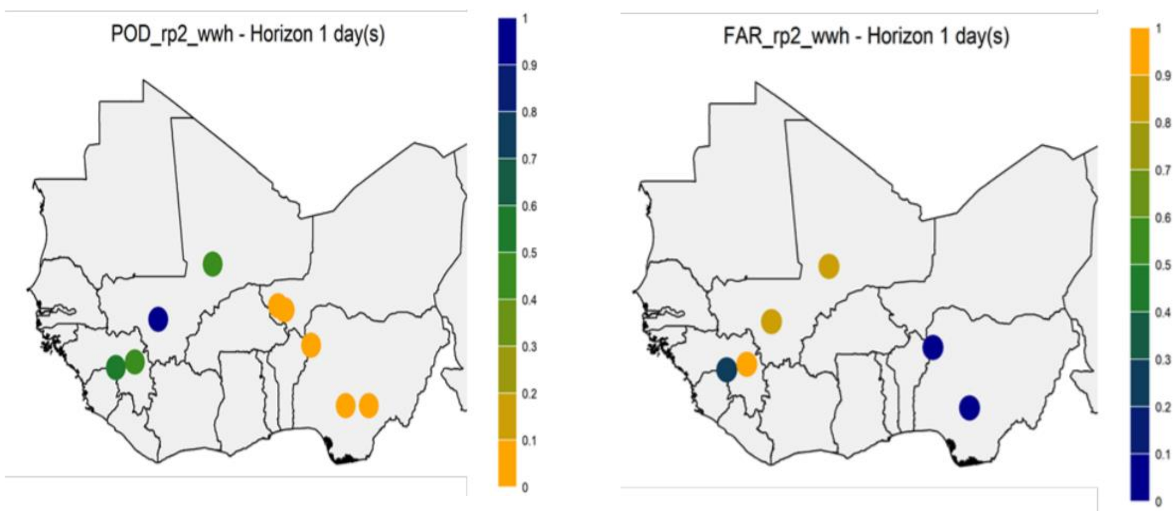


Figure 3-5: Forecast skills of the World Wide HYPE (WWH) regarding probability of detection (POD) on the left and false alarm rate (FAR) on the right , assessed at 9 gauged locations (dots). Taken from Andersson, 2020a.

Comparison of WWH and Niger-HYPE

As noted earlier, two models for FANFAR in West Africa exist; The WWH and the Niger-HYPE. This thesis is using WWH because it covers a larger domain (Figure 3-4 on p. 12) and has been subject to a rigorous calibration. Nevertheless, the difference between the two models is briefly addressed based on the subsequent case study. During a flood in September 2019 in Niamey, Niger, both Niger-HYPE and WWH did not predict a flood that was observed. In the case of Niger-HYPE, the re-forecast would predict a flood if HydroGFD 2.0 instead of HydroGFD 1.0 data were applied. Similarly, WWH would make a correct re-forecast with the second calibration step (see above) implemented.

3.2 Flood Early Warning Dissemination Investigation in West African Media

3.2.1 Data Collection

Data collection process

To investigate the current state of early warning dissemination in media in West Africa, we extracted content from newspapers and tweets based on flood events (Figure 3-6). We selected different events from a flood database between 2017-2019 (inclusive). We then used the location and time of the flood events to search media archives and tweets. We included a buffer period of two weeks before the event was registered by the flood event database, to account for possible delays in the event registration (e.g. because the flood was not strong enough yet to be registered) and to capture the potentially disseminated early warnings. We ended the search on the end date of the flood event from the database. We then counted the number of early warnings based on the media content. We conducted this analysis for the countries Ghana, Nigeria, Ivory Coast, and Burkina Faso, to account for both anglophone and francophone countries. Also, preliminary results showed that these countries have the most data available.

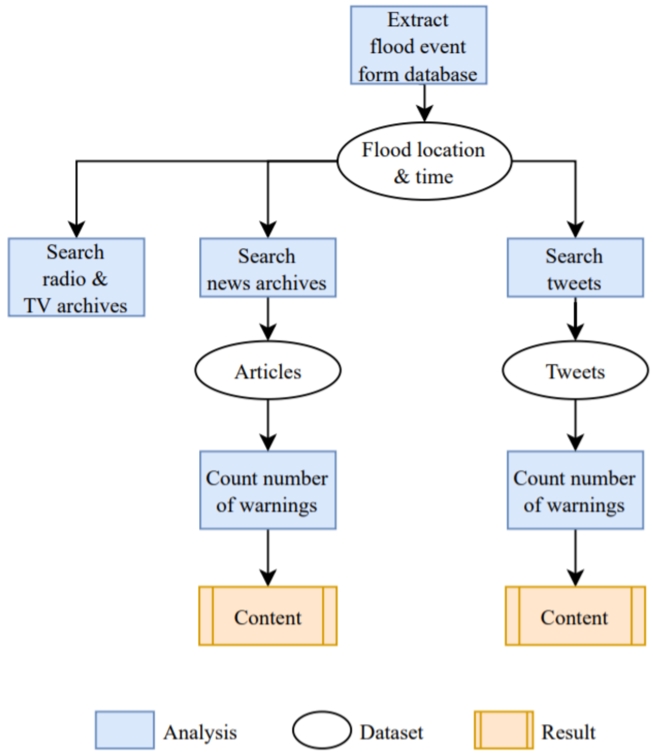


Figure 3-6: Flowchart of the data collection process. Blue boxes show a data analysis step (e.g. the extraction of a flood event or the counting the number of warnings). White ovals show datasets, like articles or tweets, resulting from a data analysis step. This data is then put into the next analysis step. Yellow boxes indicate a result (i.e. early warning related content).

Data sources

Flood event database: To address our RQs, we searched for a flood event database that includes the location, the start and end time, and the magnitude of a flood event. Furthermore, the database should cover all countries in West Africa. Preliminary results showed that the International Disaster Database (EMDAT) fulfills those requirements and is publicly accessible. EMDAT compiles disaster information from sources such as UN agencies, NGOs, and insurance companies (CRED, 2020). The disasters are registered at a country-wide scale and updated daily (CRED, 2020). For an event to be registered, 10 or more people have to be dead or 100 or more have to be affected or a declaration of emergency or a call for international assistance has to be stated (CRED, 2020). We retrieved the flood disasters from EM-DAT Public, 2020, with the filters "Disaster Classification → Flood and Location → Western Africa" and the timespan form 2017-2019 (inclusive).

Online newspaper, radio, and TV archives in West Africa: We searched for archives of newspapers, radio, and TV that contained media from 2017-2019 (inclusive) in West Africa. We included search terms such as *news archive API, Africa, archive journal Afrique, radio, TV, television, Ghana, Burkina Faso* in combination with each other on Google. We found online sources providing an overview of newspaper, TV, and radio stations in West Africa (Table 3-2). However, it proved difficult to find comprehensive news databases that could be queried systematically for the topic (i.e. specifically looking for flood early warning issues), location, and time. Besides, we could not find a database that contained a variety of publishers and that archived newspaper, radio, and TV shows (Table 3-3). We could not find any radio and TV archives and most of the newspaper databases are concerned with journal findings (e.g. Nordiska Afrikainstitutet), historical newspapers (e.g. Centre of Research Libraries), or present-day news (e.g. Abidjan News). Furthermore, in news APIs, oftentimes West Africa can not be specified as a location (e.g. Bing news) or archives only date back one year (e.g. News API). Although Archive.org offers archived radio programs, they are predominantly related to cultural topics.

Table 3-2: Meta-sources ordered alphabetically, that show lists of newspaper, TV, or radio sources. The links to the sources are displayed in Table B-1 in the Appendix.

Name of source	Description
Geopoll	TV and radio stations and overview of African newspapers
ilissafrica	Guide on how to search West African newspapers
Onlinenewspapers	Overview of African newspapers
Stanford Library	African newspaper databases

Table 3-3: News sources ordered alphabetically and their description. The weblinks are displayed in Table B-2 in the Appendix.

Name of source	Description
Abidjan news	Present-day news from the Ivory Coast in French
Afrol News	Present-day news in Africa, option to filter for region or country
All Africa	Over 130 news sources all over Africa
Archive.org	Non-profit library of free books, music, websites, and more
BBC Africa	Present-day news in Africa
Bing news search API	Get relevant news articles based on a query
Center of Research Libraries	Newspaper in Africa, among other continents, in the 19^{th} and 20^{th} century
News API	Retrieve live articles from all over the web
Newsbank	Diverse global coverage of news in over 139 countries, access needed
NexisUni	Newspaper database
Nordiska Afrikainstitutet	Policy-related research on present-day Africa
Panapress	Archive of African news over the last 15 years, subscription needed
Radio Nigeria	Out of order
ReliefWeb	Humanitarian information service, info from more than 4000 sources
Reuters	News, videos, and pictures from all over the world
Stanford Library xSearch	Research on social sciences and humanities relating to Africa

We opted for NexisUni, as it contains a large range of news articles that can be queried. However, no TV and radio news are included. NexisUni gives access to printed and online magazines, news, and blogs, from local to international issues (ETH Library, 2020). The available languages include English and French, among others. We searched with the terms "(flood* AND (warn* OR alert*)) OR (inond* AND (aler* OR avert*))" and set the filters "Location by Publication: International → Africa → *Country*" and "Timeline: *Start date of EMDAT event minus two weeks – End date of EMDAT event*". If no news were found within that query, we expanded the timeline filter to span 2017-2019 (inclusive). *Country* was set to Burkina Faso, Ivory Coast, Ghana, and Nigeria, to account for both francophone and anglophone countries.

Tweets: The collection of tweets is explained in Section 3.3.1. We filtered the collected tweets based on time (the start date of EMDAT minus 2 weeks up to the end date indicated by EMDAT) and only selected the flood-related ones (see Section 3.3.1).

3.2.2 Early Warning Dissemination Analysis

Quantitative content analysis

Definition: Content analysis can be defined as "[…] a research method that uses a set of procedures to make valid inferences from text" (Weber, 1990, p. 9), "a research technique that is based on measuring the amount of something (violence, negative portrayals of women, or whatever) in a representative sampling of some mass-mediated popular form of art" (Berger, 1998, p. 25,) or " a technique for examining the content or information and symbols contained in written documents or other communication media (e.g., photographs, movies, song lyrics, advertisements)" (Neuman, 2013, p. 49). Content can be analyzed to different extents. Descriptive, the most basic analysis, gives "insight into messages and images in discourse and popular culture represented in mass media" (Macnamara, 2005, p. 4), while predictive and inferential analysis "go further and explore what media content says about society and the potential effects mass media representations may have on audiences." (Macnamara, 2005, p. 4). Furthermore, content analysis can be divided into a quantitative or qualitative approach. Quantitative content analysis (QNCA) includes collecting data about issues, the number of mentions, the appearance of keywords, and audience reach (Macnamara, 2005). In QNCA, the form of the medium (visual media, printed text) should be considered (Macnamara, 2005), as "Form characteristics are often extremely important mediators of the content elements." (Neuendorf, 2002, p. 24). Qualitative content analysis (QLCA) considers that the meaning is polysemic, i.e. has a different meaning depending on the audience (Macnamara, 2005). While QNCA can be a scientific method, QLCA can hardly be designed to produce reliable findings (Macnamara, 2005). However, QLCA is paramount to understand the deeper meaning of content, thus a combined approach of QNCA and QLCA is likely ideal (Macnamara, 2005). Since we investigated early warning dissemination in West African media but did not analyze the role society plays in warning dissemination, we opted for descriptive QNCA.

Objectivity, validity, and generalizability of QNCA: The objectivity, validity, and generalizability of a QNCA largely depend on the selected media content sample (Macnamara, 2005). A media sample can be collected based on media forms (newspaper, magazines, radio), a specified date range, and the corresponding content (Newbold et al., 2002). Furthermore, an apriori design of the study is important to keep objectivity (Macnamara, 2005). For example, adding new issues or categories to classify content as one goes along during the study, decreases the objectivity of the study (Macnamara, 2005). To come up with an initial set of categories, the researcher can immerse her- or himself in a subsample of the selected media content sample (Neuendorf, 2002). Finally, the interpretation of the media sample by the researcher (or coder) is subjective. We describe below, how we accounted for objectivity, validity, and generalizability in our investigation, for the channels newspapers and tweets respectively.

Early warning dissemination content analysis in newspapers and tweets

Newspapers: We defined categories based on an initial body of newspapers, that we focused on in the subsequent analysis. Preliminary results using the query for Nigeria and the dates 1/8/2017-19/9/2017 (based on EMDAT, approx. 100 articles) showed that six categories appear (Table 3-4). We then analyzed the output from the search query with descriptive QNCA, i.e. we did not analyze the greater context or deeper meaning of the news articles, and counted how many times per flood event the category *Warning* appeared (Figure 3-6, on p. 15). If an article fit into more than one category, we labeled it with the in our eyes most prevalent category. We then saved the links of the articles related to a warning in Microsoft Office Excel.

Tweets: To not analyze every single tweet, we analyzed an initial set of tweets to define keywords that indicate if a tweet contains an early warning. Based on tweets from 24/6/2017-9/7/2017 in Nigeria (from EMDAT, approx. 200 tweets), we defined *warn*, *predict*, *imminent* and *alert*. If a tweet contained one or more of those keywords, it was considered to disseminate early warning. We then checked the rest of the tweets in Ghana and Nigeria. We also investigated tweets manually from Burkina Faso and Ivory Coast to account for francophone countries. We conducted this analysis in Microsoft Office Excel.

Table 3-4: Categories identified from an initial set of newspapers. The category Warning in bold describes when the content of a news article is considered to be an early warning.

Category	Description
Prevention	• Prevention of any disaster, such as fires or earthquakes, and floods • Analysis of current susceptible areas (e.g. susceptible to floods, fire, cholera, etc.) • Articles about enhancing resilience towards different kinds of disasters. • Warning e.g. about not putting illegal structures somewhere, or how to handle agricultural fields
Relief	• Relief measures for flood victims • Visits of important people (e.g. politicians) to sites struck by flooding
Report	• Report on flood disasters, e.g. how many people were victimized • Recap of an event that happened in the past and now e.g. damage reports are published
Warning	• **Warning population itself** • **Warning forwarded from e.g. local meteorological agency, related to an imminent flood threat or risk, if not projected to longer than one month**
Other	• Other disasters • Social unrest that included flooding and warning in the text, such as illegal mining problems, dumping of illegal waste, or underfunded organizations

3.3 Assessing FANFAR with Twitter

3.3.1 Data Collection and Pre-Processing

We used three datasets to investigate if Twitter can be used to assess the performance of FANFAR:

(1) The FANFAR forecasts and re-analyzed rainfall from HYPE
(2) Posts extracted from tweets, containing the word flooding, and flood events that are registered based on the extracted tweets by a Twitter flood detection algorithm
(3) Events registered by the natural disaster database EMDAT

Datasets (1) and (2) were pre-processed for the subsequent data analysis and are described in detail below.

(1) FANFAR forecasts

The FANFAR forecasts were reproduced for the years 2017-2019 (inclusive) with the help of the re-forecasting function of HYPE. The date of the desired forecast can be set from today back to 01-01-2017 (Andersson, 2020b). A hindcast, up to the date of the forecast is performed by HYPE to get the initial conditions (Andersson, 2020b). Subsequently, the re-forecast was produced. Contrary to the hindcast, where historical meteorological *forcing* data is used, the re-forecasting employs meteorological *forecasts*. The lead time ranges from 1-10 days. (i.e. the re-forecasts range from 1-10 days ahead of the day they are issued for). The results of the re-forecast are predicted streamflow severity levels in the form of return periods (R) for every SC in West Africa. Four severity levels are used: No danger, moderate (R > 2 years), high (R > 5 years), and very high (R > 30 years). We used FANFAR with the WWH and HydroGFD 1.0 that was operational when the re-forecasts were produced. Besides, we limited our analysis to the 1-day re-forecast, as we assume that they are the most accurate compared to e.g. the 10-day re-forecast.

We also used rainfall from HydroGFD 3.0, as described by Berg et al., 2020. HydroGFD 3.0 contains daily precipitation and temperatures (mean, minimum, and maximum) that are bias-adjusted, and it covers the whole world (with exception of the Antarctic continent) at an approximately 25 km resolution. The main method to produce the HydroGFD 3.0 data set includes adding observational monthly anomalies to climatology and adjust the re-analyzed data to the monthly mean. We refer to Berg et al., 2020 for details on the further processing steps, e.g. how consistency between different data sets is assured.

(2) Tweets and flood events from the detection algorithm

Retrieval, classification, and geolocation of tweets: We apply the approach of de Bruijn et al., 2019, extracting the tweets from 01/01/2017-31/12/2019 in West Africa. This approach best suits assessing the performance of FANFAR, as it works on a global scale and covers languages common in West Africa. Furthermore, its code and data are publicly available, and the approach has been validated. The approach can be summarised as follows: Firstly, tweets are retrieved from the Twitter real-time streaming API and filtered for the keyword flood* (English) and another 11 languages (e.g. inond* for French or inund* for Spanish). * indicates a wild card. Secondly, since only 2 % of tweets are precisely located (Leetaru et al., 2013), they need to be geotagged (i.e. assigned a location). The TAGGS (de Bruijn et al., 2017) evaluates the most probable location based on the tweet's time zone, mentioned city names, and other criteria (Figure A-3 in the Appendix). Thirdly, the tweets are then classified with the Bidirectional Encoder Representations from Transformer (BERT) as *flood-related* (i.e. related to an ongoing event) or *flood-unrelated* (related to the commemoration of a flood or a figurative use). BERT is a natural language processing tool that learns relations between words to predict their meaning (de Bruijn et al., 2019). In summary, we now know where the tweets are coming from and if they are related to flooding. In the next step, we register events based on these tweets.

Flood event detection: We apply the flood event detection algorithm (FEDA) developed by de Bruijn et al., 2019 since it considers regional fluctuations in the number of tweets (Figure 3-7). FEDA contains six steps. (1) Δt_c is the for day/night variations corrected time between tweets. (2) The thresholds Θ_{Start} and Θ_{End} for the critical time between two tweets are computed based on 5 % and 30 % of expected tweets during a one-year spin-up period for the start and end of an event respectively. $\Theta_{Start} > \Theta_{End}$ since people usually tweet more at the beginning of a flood event than towards the end (de Bruijn et al., 2019). (3) If $\Delta t_c < \Theta_{Start}$, then an average corrected time $\overline{\Delta t_c}$ over an array a_n of size $n = f(e^{-\Theta\text{Start}}, c \in \{1,2,3\})$ is filled with tweets. c is a constant and indicates, how sensitive the FEDA is. A small c implies that we compute $\overline{\Delta t_c}$ over a smaller number of tweets, leading to an earlier and more frequent event detection (de Bruijn et al., 2019). The available settings of the algorithm are sensitive (c=1), balanced (c=2), and strict (c=3). (4) If $\overline{\Delta t_c} < \Theta_{Start}$, the priorly suspected event is confirmed and the region is set to flood state. (5) During the flood state, an incoming tweet is added to an array a_n and the oldest tweet in a_n is discarded. If $\overline{\Delta t_c}$ surpasses Θ_{End}, or no tweet was received in the last 24 hours, the flood state is revoked. (6) A continuation of the detected event is assumed if a new flood is detected within three days after. In a nutshell, we now have detected events based on a regional tweet density and threshold. We use mostly county-wide identified events in the analysis, as preliminary results showed that most tweets were only geotagged at a country level. The validation of the FEDA showed that 75 %-90% of events were correctly identified with the sensitive and strict configuration respectively (de Bruijn et al., 2019).

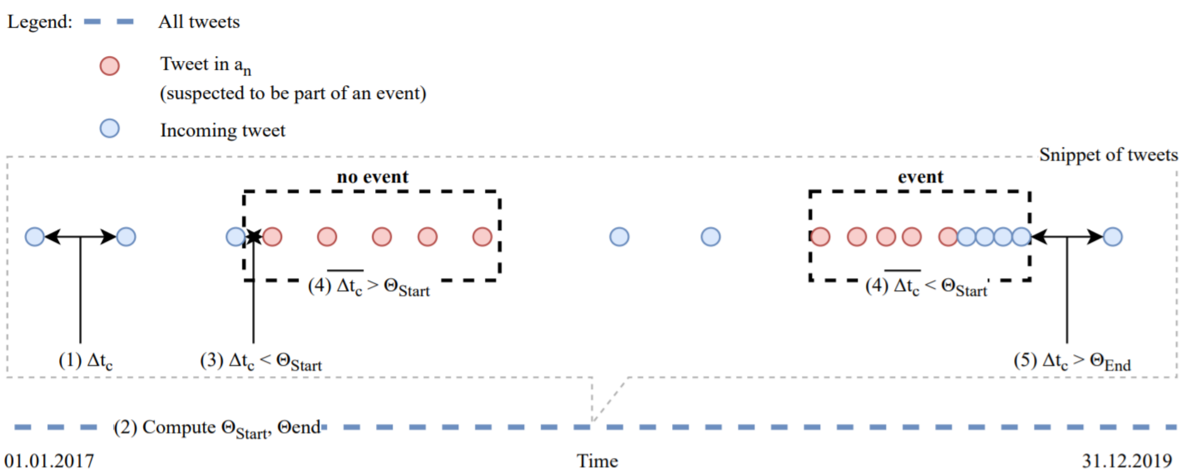


Figure 3-7: Flood event detection algorithm (FEDA) with steps (1)-(5) indicated, adapted from de Bruijn et al., 2019. Abbreviations: Δt_c – Corrected time between tweets, Θ_{Start}, Θ_{End} – Thresholds for the start and end of an event, a_n – array of tweets suspected to be flood-related.

Data pre-processing of FANFAR forecasts, tweets, and rainfall

Firstly, we selected countries to analyze and lumped the FANFAR forecasts to facilitate a comparison with the events identified from tweets. Secondly, the tweets extracted by the algorithm needed to be assigned a unique location and relevancy. Lastly, we extracted rainfall exceedances to assess the influence rainfall has on Twitter activity.

Forecasts: Firstly, we selected 14 countries for this study: Benin, Burkina Faso, Gambia, Ghana, Guinea, Ivory Coast, Liberia, Mali, Mauritania, Niger, Nigeria, Senegal, Sierra Leone, and Togo. These countries are completely covered by the basin delineation of HYPE (unlike e.g. Algeria, see Figure 3-8) and account for different languages in West Africa (English, French). Secondly, the resolution of the FANFAR forecasts is approximately 100 times higher than the resolution of the events detected from tweets at a country level (i.e. every country contains around 100 sub-basins). Therefore, the forecasts had to be lumped on a country scale for every day. The HYPE SC contained in a specific level 0 or level 1 zone (Figure 3-8) were selected with the help of the Python function sjoin included in the geopandas library. Subsequently, we developed a flood risk indicator FRI at a basin level (Equation 4). The FRI sums up the weighted number of forecasted streamflow return periods, divided by the total number of catchments within a country. For example, if a country *i* contains 100 basins, and during day *k* four catchments were predicted to have a streamflow return period of 5 years and one catchment of 30 years, then $FRI_{k,i} = \frac{4*5+1*30}{100} = 0.5$.

$$FRI_{k,i} = \frac{green * 0\ years + yellow * 2\ years + orange * 5\ years + red * 30\ years}{\sum_{b=1}^{n} basin_{b,i}}$$

Equation 4: Calculation of lumped forecast intensity $FRI_{k,i}$ during day k in country i. Green, yellow, orange, and red denote the number of forecasts that have been issued with predicted return periods of R = 0 or R > 2, 5, 30 respectively. $basin_{b,i}$ denotes the b^{th} basin in country i.

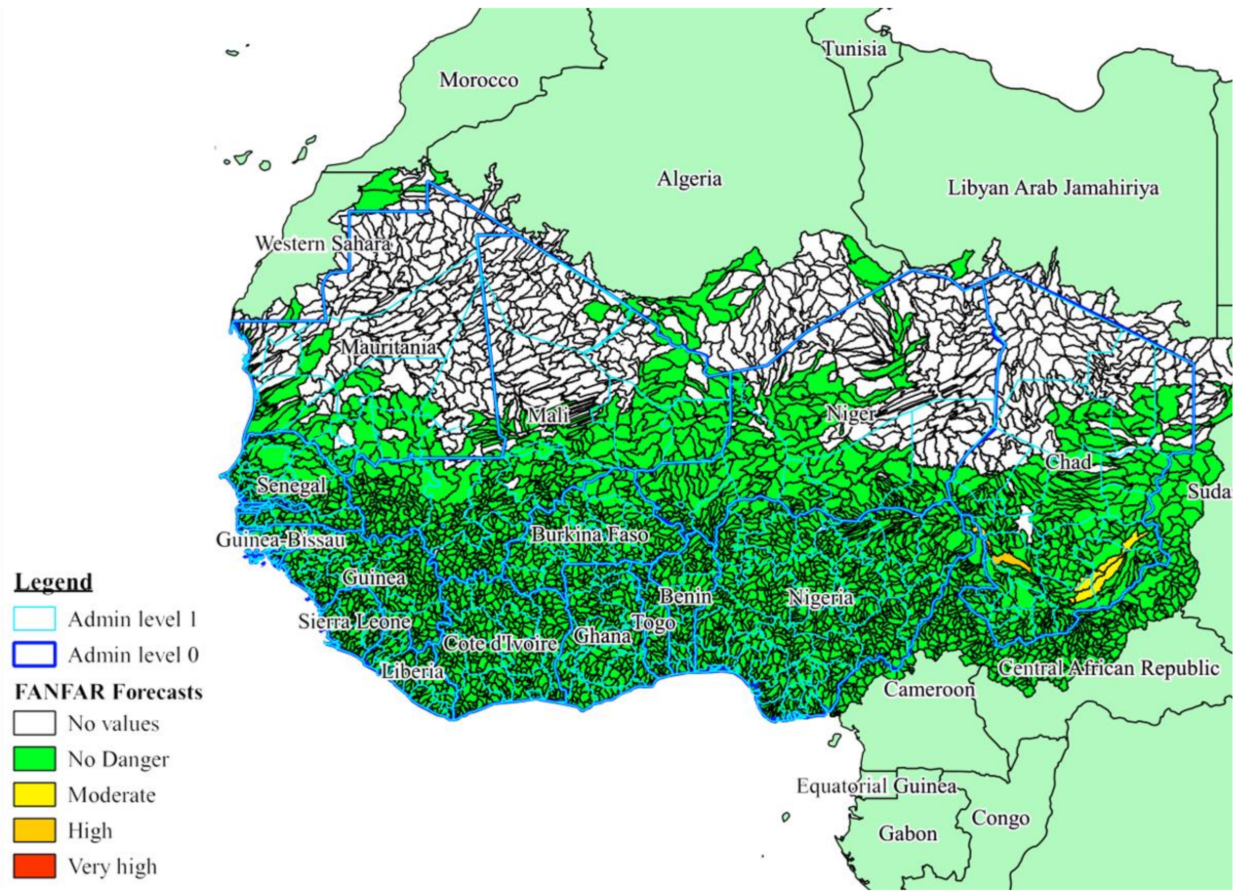


Figure 3-8: FANFAR forecasts, World Wide Hype (WWH) basin delineation, and subset level 0 (country, dark-blue) and level 1 (region, light-blue) borders for West Africa for the 19th of January 2017. The colors indicate the forecasted streamflow return period. No values can be mostly found in the desert area. No danger (green) signifies no increased streamflow. Moderate (yellow), high (orange), and very high (red) indicate a return period of 2, 5, and 30 years respectively.

Tweets: Firstly, we extracted only the most probably candidate location (location with the highest score) out of all the possible candidate locations determined by the TAGGS. We then filtered the tweets according to the country of interest, e.g. Ghana or Niger. Secondly, we calculated the average number of tweets per day of every country, disregarding outliers. We refer to this number as the base activity of Twitter users (ε). ε_{FR} (Equation 5) and ε_{NFR} (Equation 6) are indicators of the number of flood-related and flood-unrelated tweets that would usually occur in a country and account for the difference in social media use across West Africa. For every date, the number of flood-related or flood-unrelated tweets posted were counted. Subsequently, we collected only the days $D_{non\text{-}zero}$ where flood-related or flood-unrelated tweets were posted (e.g. a date without flood-related tweets was not considered). Next, we disregarded the days where the number of tweets fell below the 10 % quantile and exceeded the 90 % quantile during the whole timeframe from 2017-2019, yielding $T_{i,flood\text{-}related,0.1\text{-}0.9}$ and $T_{i,flood\text{-}unrelated,0.1\text{-}0.9}$ respectively. Subsequently, the average number of posted Tweets on the remaining days were calculated. This procedure was carried out for every country in West Africa.

$$\varepsilon_{i,FR} = \frac{\sum_{2017}^{2019} T_{i,flood-related,0.1-0.9}}{D_{i,non-zero}}$$

Equation 5: Calculation of the base activity of Twitter users of flood-related tweets in country i ($\varepsilon_{i,FR}$). $T_{i,flood\text{-}related,0.1\text{-}0.9}$ denotes the number of flood-related tweets between the 10% and 90% quantile, $D_{i,non\text{-}zero}$ denotes the number of days where flood-related tweets were posted.

$$\varepsilon_{i,NFR} = \frac{\sum_{2017}^{2019} T_{i,flood-unrelated,0.1-0.9}}{D_{i,non-zero}}$$

Equation 6: Calculation of the base activity of Twitter users of flood-unrelated tweets in country i ($\varepsilon_{i,NFR}$). $T_{i,flood\text{-}unrelated,0.1\text{-}0.9}$ denotes the number of flood-unrelated tweets between the 10% and 90% quantile, $D_{i,non\text{-}zero}$ denotes the number of days where flood-unrelated tweets were posted.

Rainfall: We counted for every day between the years 2017 and 2019 (inclusive), how many basins within a country or region experienced rainfall larger than 20 mm/day (Equation 7). We assumed that this threshold is adequate for medium to high-intensity events over all West African countries. We then divided by the number of basins included in the WWH domain within one region (e.g. Ghana or Niger) to receive the normalized rainfall exceedance (NRE).

$$NRE_{>20mm,k,i} = \frac{\sum Catchments\ with\ rainfall > 20mm}{\sum_{b=1}^{n} basin_{b,i}}$$

Equation 7:Calculation of rainfall exceedances above 20mm in country i during day k, normalized by the number of basins within a region.

3.3.2 Data Analysis

Visual analysis of time series during the wet season

We visually investigated the EMDAT and FEDA events, the FRI, and rainfall exceedances to assess the performance of FANFAR during the wet season (May-October). We focused on the wet season since model performance is arguably the most important during that period. We present Ghana and Niger in detail since preliminary results showed that they have the most data available and to consider anglophone and francophone countries.

EMDAT and FEDA: Firstly, we investigated if the events registered by EMDAT overlap with the events detected by FEDA. This analysis sheds light on the accuracy of the detection algorithm and allows us to conclude whether we can assume that FEDA correctly indicates events in West Africa.

FRI and FEDA: Secondly, under the assumption, that the FEDA correctly registers events, we identified true positives (TP), false positives (FP), and false negatives FN (Table 3-5). We then calculated the POD and FAR and compared them to the values obtained using streamflow data. We counted a TP, FP, and FN if an increased FRI and FEDA detected event are not (TP), respectively are (FP and FN) more than five days apart, to account for potential lags between the two time-series. Lags could arise due to people tweeting during rainfall (FEDA detected event precedes and increased FRI, see below) or during the after-effects of a flood (FEDA detected event follows up on an increased FRI). Therefore, we also checked if an increased FRI usually follows up on or precedes a FEDA detected event.

Table 3-5: Confusion matrix, showing how we define true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). FEDA detected events are assumed to be actual events.

		FEDA actual event	
		Yes	*No*
FANFAR prediction	*Yes*	TP	FP
	No	FN	TN

Rainfall and FEDA: Finally, we checked if an increase of rainfall exceedances P > 20 mm/day coincides with FEDA detected events. Since FANFAR predicts streamflow, tweets that are issued during rainfall cannot be directly used to assess the performance of FANFAR but would need additional criteria (e.g. time of regional runoff concentration).

Correlation analysis of FRI and tweets

We went beyond the visual analysis and investigated correlation coefficients between the variables *number of tweets*, the *FRI* and the *NRE* . All coefficients were calculated with the scipy package (version 1.5) in Python.

Pearson and Spearman: We compared the Spearman (r_S) and the Pearson (r_P) correlation coefficients. We used two types of coefficients because a significant r_S can imply a significant or non-significant r_P on the same dataset (Hauke and Kossowski, 2011). Both values range between -1 and 1, where values close to -1 indicate an inverse correlation and values close to 1 a correlation. r_p measures the linear relationship between two variables, assuming that the variables correlate linearly (Hauke and Kossowski, 2011). On the contrary, r_s measures how well a monotonic function can describe the relationship between two variables, without assuming a frequency distribution of the variables (Hauke and Kossowski, 2011). Preliminary results showed that r_s is better suited for our research.

All tweets and FEDA: We compared all flood-related tweets with the FRI and calculated r_s to find out how well FANFAR predicts periods where floods or no floods are happening. Since most data points are not inside a FEDA detected event, FNs largely influence correlation. We then compared the sensitive, balanced, and strict FEDA configurations to calculate the event-related r_s to investigate if FANFAR manages to capture flood events.

Regional: We chose the biggest city in every country (one basin from the FANFAR basin delineation) and extracted its FRI. We aim to find if the match of forecasts and tweets improves in cities under the assumption that all tweets tagged at a country-scale emerged from that city.

Seasonal: We calculated the wet-season (May-October) r_s for the whole country to investigate if FANFAR performs differently compared to the whole year. This analysis also allows us to find if flood-related tweets are falsely placed in the dry season.

Lag: Finally, we extracted the lagged r_s. to see whether an increased FRI comes before or after an increased Twitter activity. The time series of tweets were shifted by -5 to 5 days relative to the FRI time-series (Figure 3-9). For example, Lag 3 calculates r_s between day 1 of the tweet time series and day 3 of the FRI time series. This analysis allows us to detect systematic shifts between the two time-series and to conclude if the lag must be considered as an additional criterion in assessing the performance of FANFAR.

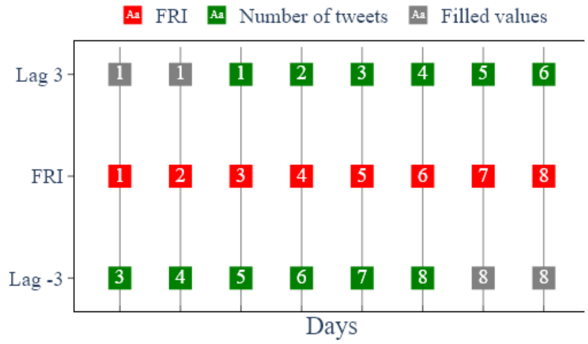


Figure 3-9: Lagged correlation between FRI (red) and number of tweets (green) for the lags -3 and 3 days. The filled values (grey) are equal to the first respectively last value of the shifted time series.

NRE: We also calculated r_s between the NRE and tweets, to assess the influence of rainfall on tweets.

t-Test of FRI during and outside of FEDA

We used the Welch's test to find if the sample *FRI during events* detected by the FEDA (sensitive configuration) significantly differs from the sample *FRI when no event* was predicted by the algorithm. A systematic difference would suggest that the forecasts manage to capture flood events under the assumption that FEDA correctly detects flood events. To find whether the events detected by FEDA generally precede, coincide with, or follow up on an increased FRI, we used three different group separations (Figure 3-10) and calculated the p-values p_{pre}, p, and p_{sub}. p_{pre} includes the FRI during and five additional days before the detected event, to find whether an increased FRI occurs before and during a FEDA detected event. p is calculated based on the FRI occurring only during a detected event. Finally, p_{post} includes the FRI during and five days after a FEDA detected event. The Welch's test is based on the Student's t-test, but only assumes normality and not equal variance of the two samples (Ahad and Yahaya, 2014). For small sample sizes, increasing variance with decreasing sample size and skewness of the data, the false positives of the test increases (Ahad and Yahaya, 2014). We checked for the skewness of the data with the Shapiro-Wilk test (normality test) to conclude how reliable the results of the Welch's test are. The Shapiro-Wilk test is powerful to assess the normality of sets containing more than 30 samples (Mohd Razali and Yap, 2011). The Welch's and Shapiro-Wilk tests were both calculated with the Python package scipy (version 1.5).

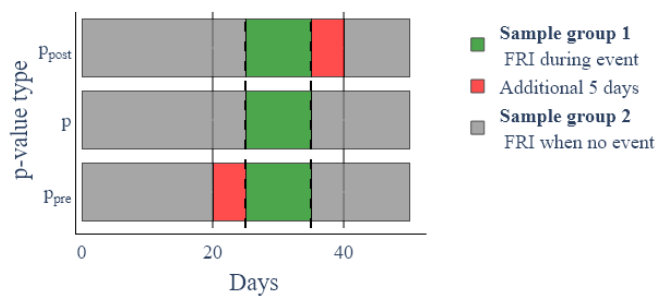


Figure 3-10: Example of t-test sample group separations for the p-value types p_{pre}, p, and p_{post}, accounting for potential lags between the FRI and events detected by the Twitter algorithm.

4. Chapter: Results and Discussion

4.1 Flood Warning Dissemination in West Africa

The newspapers involved in warning dissemination vary between Ghana and Nigeria (Table 4-1). The number of early flood warnings issued in the investigated countries varies between 0 and 22 (Table 4-2). In Ghana, we cannot see a trend regarding flood early warnings posted between flood events where many people are affected compared to flood events where fewer people are affected. During every event, a different warning administration issued the warning, which was then disseminated by the newspaper. NEMA and NIHSA are the most prominent agencies that are mentioned in newspapers. In Nigeria, larger events seem to coincide with a larger number of early flood warnings. However, early and late warnings do not seem to depend on the number of people that were affected by a flood event. Furthermore, more warning administrations are mentioned in news articles compared to Ghana. The number of tweets containing early flood warnings ranges between 6 and 21. In Ivory Coast and Burkina Faso, we could not find any flood early warnings issued between 2017-2019 (Table 4-2).

The results show that in the countries investigated, no overall trend can be established regarding event size and number of warnings disseminated. We assume that in the other countries in West Africa, the information about warning dissemination is scarcer, since Ghana, Nigeria, Ivory Coast, and Burkina Faso have the most inhabitants among countries in West Africa. Furthermore, the analysis of early warning is difficult when based on EMDAT events, since the EMDAT start date does not necessarily correctly register the actual start of a flood. Besides, NexisUni only offers a narrow sample of available newspapers (Macnamara, 2005). A more in-depth analysis of local newspapers for every country should be conducted. Furthermore, online issues might not include all supplementary material that printed versions might have (Macnamara, 2005). Finally, a more in-depth analysis of the links provided in tweets might give insight into additional early warnings issued. Most of the events have around 90 % of tweets with links attached (Table 4-2).

These findings indicate that every country should be investigated individually regarding flood early warning dissemination, as we could not find a database containing newspapers, TV, and radio of all West African countries. Furthermore, the investigation of warning dissemination in TV and radio must likely be conducted “live” (i.e. when an event is predicted, the researcher should scout live TV and radio programs).

Table 4-1: Countries and news publishers involved in warning dissemination.

Country	**Publishers**
Ghana	Ghana News Agency
Nigeria	The Sun (Nigeria), The Nation (Nigeria), Weekly Trust, PM News, Vanguard (Lagos), Leadership (Abuja), Daily Trust (Abuja), Nigerian Tribune, Business Day, This Day (Lagos)
Ivory Coast	No warnings found
Burkina Faso	No warnings found

Table 4-2 Flood event media analysis based on EMDAT. Total warnings denote the number of news or tweets that are categorized as warning-related. Early warnings are issued latest one day before an event registered by EMDAT, and late warnings are issued at the beginning or during the registered EDMAT flood event. Warning administration displays which administrations are mentioned in the news article. Contain links shows the percentage of tweets containing links that potentially point to more information. Administration abbreviations: NADMO - National Disaster Management Organisation, GMA – Ghana Meteorological Agency, VRA – Volta River Authority, SEMA – State Emergency Management Agency, NIHSA – National Hydrological Services Agency, Gov. – Government, NBA – Niger Basin Authority, Shiroro Hydroelectric Power Station, NOA - National Orientation Agency, KSG – Kwara State Government, FCT – Federal Capital Territory, DSG – Delta State Government. Note that for Burkina Faso and Ivory Coast, we do not distinguish between early and late warnings, as we do not have a flood event that we base the query on.

EMDAT				**Newspaper**				**Tweets**			
Start search	**Start event**	**End search**	**People affected**	**Total news**	**Total warnings** *early*	*late*	**Warning administration**	**Total tweets**	**Total warnings** *early*	*late*	**Contain links**
				Ghana							
15/03/2017	01/04/2017	21/07/2017	1’000’000	20	1	0	NADMO	40	0	0	83%
15/08/2018	31/08/2018	02/10/2018	100’000	8	1	1	VRA	585	0	0	98%
20/05/2019	04/06/2019	04/06/2019	-	10	0	1	GMA	133	0	0	86%
22/09/2019	06/10/2019	06/10/2019	26’000	1	0	0	N/A	1	0	0	100%
				Nigeria							
24/06/2017	08/07/2017	09/07/2017	500	13	1	3	SEMA, NIHSA, Gov.	191	2	2	89%
01/08/2018	15/08/2017	19/09/2017	10’000	106	0	13	NBA, Gov., SHPS	1473	0	50	87%
26/06/2018	13/07/2018	16/07/2018	16’000	13	1	0	NEMA	111	0	0	93%
06/09/2018	20/09/2018	02/10/2018	2’000’000	134	22	1	NEMA, NIHSA, NOA	2113	13	0	95%
06/08/2019	20/08/2019	25/08/2019	50’000	81	14	3	NIHSA, KSG, FCT, DSG	179	1	5	58%
				Ivory Coast							
01/01/2017	N/A	31/12/2019	N/A	3	0		N/A	373		21	30%
				Burkina Faso							
01/01/2017	N/A	31/12/2019	N/A	28	0		N/A	46		6	87%

4.2 Assessing FANFAR with Twitter

4.2.1 The Statistics of EMDAT, the FRI, Rainfall and tweets

FEDA and EMDAT Events

Around 100 FEDA (sensitive) and 28 EMDAT events are registered across all 14 countries. The event detection algorithm based on tweets captures most of the events registered by EMDAT (Figure 4-1). Also, the event detection algorithm indicates events that have not been captured by EMDAT (Figure 4-1). If tweets reliably register flooding, this result implies that the detection algorithm is better suited than EMDAT to assess the performance of FANFAR. Furthermore, Twitter bears the potential for additional information, such as photos or videos, which is information usually not found in event disaster databases.

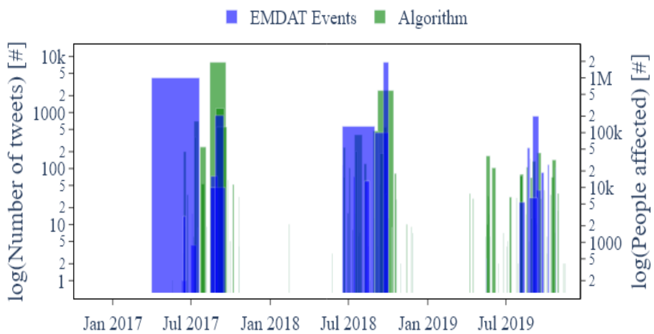


Figure 4-1: Events detected by the sensitive algorithm (Algorithm, green) on the left and Events registered by EMDAT (EMDAT Events, blue) on the right for all 14 countries. Note that both y-axes are logarithmic.

Pre-processed FANFAR forecasts, rainfall and tweets

FANFAR forecasts: Benin, Gambia, and Togo show the highest maximum FRI (Table 4-3). The mean of the FRI in all countries is approximately between 0 and 0.1. The maximum is between 20 times and 300 times higher than the mean. Also, the positive skewness and low variance indicate that the FRIs are zero-inflated, implying that the majority of statistical tests are unsuited (Sheard, 2018).

Table 4-3:Mean, maximum, variance, and skewness of the Flood Risk Indicator (FRI) for every country.

Country	*Benin*	*Burkina Faso*	*The Gambia*	*Ghana*	*Guinea*	*Ivory Coast*	*Liberia*
Mean	0.040	0.005	0.009	0.108	0.003	0.024	0.003
Maximum	5.636	1.579	4.412	2.182	0.296	3.061	0.526
Variance	0.136	0.004	0.026	0.016	0.001	0.034	0.001
Skewness	11.612	19.131	22.769	11.04	8.402	11.479	13.012
Country	*Mali*	*Mauritania*	*Niger*	*Nigeria*	*Senegal*	*Sierra Leone*	*Togo*
Mean	0.024	0.003	0.053	0.006	0.004	0.004	0.041
Maximum	1.678	0.19	1.583	0.337	0.878	0.947	7.500
Variance	0.008	0.0	0.015	0.001	0.002	0.003	0.148
Skewness	12.827	7.242	5.689	6.594	14.814	16.236	13.049

Rainfall: The mean NREs vary between approx. 0 and 0.09 (Table 4-4) and the maximum lies between approx. 0.1 and 1.0. A value of 1.0 indicates that all basins were experiencing daily rainfall > 20 mm. The positive skewness and low variance indicate that the normalized rainfall exceedances are zero-inflated, implying that the majority of statistical tests are unsuited (Sheard, 2018).

Table 4-4: Mean, maximum, variance, and skewness of number of daily rainfall exceedances > 20mm (NRE), normalized by the number of World Wide HYPE delineated basins, for every country.

Country	*Benin*	*Burkina Faso*	*The Gambia*	*Ghana*	*Guinea*	*Ivory Coast*	*Liberia*
Mean	0.034	0.026	0.025	0.021	0.066	0.035	0.084
Maximum	0.727	0.684	1.0	0.673	0.901	0.564	1.0
Variance	0.008	0.006	0.013	0.004	0.017	0.005	0.025
Skewness	0.034	0.026	0.025	0.021	0.066	0.035	0.084
Country	*Mali*	*Mauritania*	*Niger*	*Nigeria*	*Senegal*	*Sierra Leone*	*Togo*
Mean	0.018	0.002	0.009	0.047	0.026	0.088	0.029
Maximum	0.502	0.127	0.294	0.684	0.717	1.0	0.69
Variance	0.002	0.0	0.001	0.006	0.008	0.035	0.006
Skewness	0.018	0.002	0.009	0.047	0.026	0.088	0.029

Tweets: Overall, we use approximately 63’000 tweets, of which around 26’000 are classified as flood-related (Table 4-5). For both flood-related and flood-unrelated tweets, the tagging algorithm locates most of them on a country scale and approximately 4 % on a regional scale. This implies that the tweets are useful to assess FANFAR countrywide. Furthermore, the tweets are written mostly in English, followed by Spanish (Figure 4-2). Surprisingly, French is under-represented, as 9 out of the 14 countries investigated are francophone (Gut, 2010). This shows that the classification algorithm either cannot properly identify tweets written in French (e.g. due to colloquial semantics) or fewer people tweet about flooding in French. This means that the assessment of FANFAR in francophone countries might be more difficult. On a side note, the surprisingly large amount of Spanish tweets seems to come from Alerta Geo, a website aiming to convey natural events to the world (Lopez, n.d.). Other languages, such as Indonesian, could indicate that the TAGGS misplaced a tweet in West Africa, that was issued in Indonesia e.g. by a news agency.

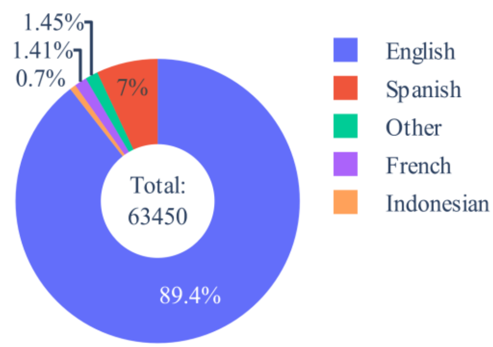


Figure 4-2: Distribution of languages of all tweets, regardless of relatedness to flooding.

Table 4-5: The total number of tweets and the percentage of tweets classified as flood-related and not related. Furthermore, the levels at which the tweets are geo-located (country or regional) are indicated. Note that most tweets that are geo-located at a regional level are also located at a country level.

Total = 63’450 Tweets					
Class		% of total	Class	Amount	% of total
<u>Flood-related</u>	26’423	41.6	<u>Not related</u>	37’027	58.4
Country	26’081	41.1	Country	36’617	57.7
Region	2’893	4.5	Region	2’599	4.1

Our analysis shows that Ghana, Niger, Nigeria, and Sierra Leone display the highest base activity of Twitter users (ε_{FR}, ε_{NFR}) and the highest number of days over which this activity was computed (d_{FR}, d_{NFR}), for flood-related (FR) and flood-unrelated (NFR) tweets respectively (Table 4-6). On the one hand, the ε_{FR} in Sierra Leone is with 11 tweets/day around 4 times higher than in Ghana. This suggests that users in Sierra Leone tweet more than in any other West-African country investigated during a flood. However, a relatively low d_{FR} of around 50 days compared to Ghana, Niger, and Nigeria could indicate that this peak of tweets only occurred on a special occasion and does not indicate a generally high tweet activity. Figure B-13 in the Appendix shows that the peak of tweets occurred in September 2017. A brief check of floodlist.com[1] shows, that during that time, indeed a mudslide ravaged Freetown. This is the first indicator that tweets indicate flooding or problems related to flooding. However, we can see that FANFAR would likely not predict such an event that is not directly related to streamflow. On the other hand, ε_{NFR} is around 11 tweets/day in Nigeria, implying that most tweets containing the word flood* are, according to BERT, not related to flooding. A high d_{NFR} implies that tweet activity in the country is generally high, also when tweets are not related to flooding (but still contain the word flood*). In a nutshell, Ghana, Niger, and Nigeria show the overall highest tweet activity.

Table 4-6: Base activity ε_{FR} of flood-related tweets and ε_{NFR} of tweets containing the word flooding, but classified as not flood-related according to the natural language processing algorithm (BERT) over the years 2017-2020. ε_{FR} and ε_{NFR} are the average numbers of tweets inside the 10% and 90% quantile. d_{FR} and d_{NFR} indicate the number of days over which the average of tweets is calculated for ε_{FR} and ε_{NFR} respectively. Example: In Benin, on average 3.34 flood-related tweets were posted over the 32 days where the daily number of tweets did not fall below the 10 % or surpass the 90 % quantile.

Country	ε_{FR} [tweets/day]	d_{FR} [days]	ε_{NFR} [tweets/day]	d_{NFR} [days]
Benin	3.34	32	2.33	54
Burkina Faso	2.0	2	2.0	5
Gambia	2.2	5	2.26	34
Ghana	3.64	104	3.98	590
Guinea	2.75	12	2.18	34
Ivory Coast	3.56	18	2.76	17
Liberia	2.88	8	2.0	24
Mali	2.61	28	2.26	82
Mauritania	2.0	2	2.0	3
Niger	8.18	177	3.95	251
Nigeria	5.23	492	11.03	818
Senegal	2.45	20	2.0	24
Sierra Leone	11.12	48	3.41	137
Togo	2.0	5	2.0	7

[1] http://floodlist.com/?s=sierra+leone&submit=, accessed 23/11/2020

4.2.2 Data Analysis

Visual analysis of time series during the wet season

We compare EMDAT with FEDA and rainfall exceedances and illustrate the results for Ghana and Niger in detail to show how Figure B-1 to Figure B-14 in the Appendix should be read. We then discuss observations from all countries. Overall, the algorithm captures all events registered by EMDAT. In general, an improvement of the FEDA detected events and the FRI over the years 2017-2019 can be observed in Senegal, Mauritania, Mali, Nigeria, and Sierra Leone. The POD and FAR based on tweets and streamflow yield similar results in the Upper Niger basin. However, the results differ for the Lower Niger basin. Further, we did not observe a correlation between rainfall and FEDA.

In Ghana, EMDAT and FEDA both indicate events in 2017 and 2018 (Figure 4-3). We observe that EMDAT registered a longer period of flooding than FEDA. FEDA detected 14 events, where 4 coincide with an increased FRI (TP) and 10 do not (FN). The FRI is increased on 10 occasions without a FEDA detected event (FP). We did not account for the generally increased FRI from June 2017 to October 2019, but only included sudden changes in the FRI to identify TP, FP, and FN. Furthermore, tweets do not seem to be systematically influenced by rainfall.

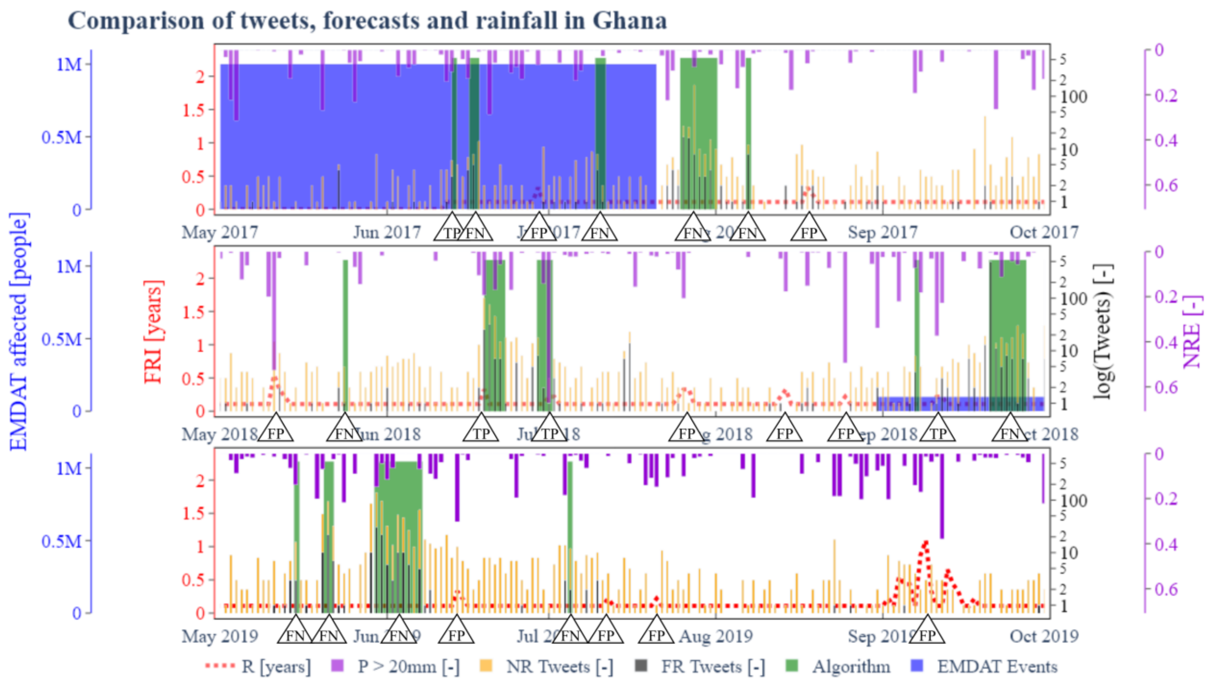


Figure 4-3: Time series of Ghana for EMDAT events (blue bar), FRI (red, dashed), flood-related and -unrelated tweets (black and yellow bar), and precipitation exceedances (purple bar). The yellow bars display the tweets not related to flooding, and the black bars show the tweets that have been identified as flood-related by the classification algorithm. Furthermore, the duration of events identified by the detection algorithm is displayed in green. The intensity of the event can be deduced from the number of black bars inside the green area. True positives (TP), false positives (FP), and false negatives (FN) are indicated as triangles.

In Niger, EMDAT and FEDA both indicate events in all three years (Figure 4-4). We again observe that the EMDAT registered a longer period of flooding than FEDA. FEDA detected 18 events, where 8 coincide with an increased FRI (TP) and 10 do not (FN). The FRI is increased on 2 occasions without a FEDA detected event (FP). Furthermore, tweets do not seem to be systematically influenced by rainfall.

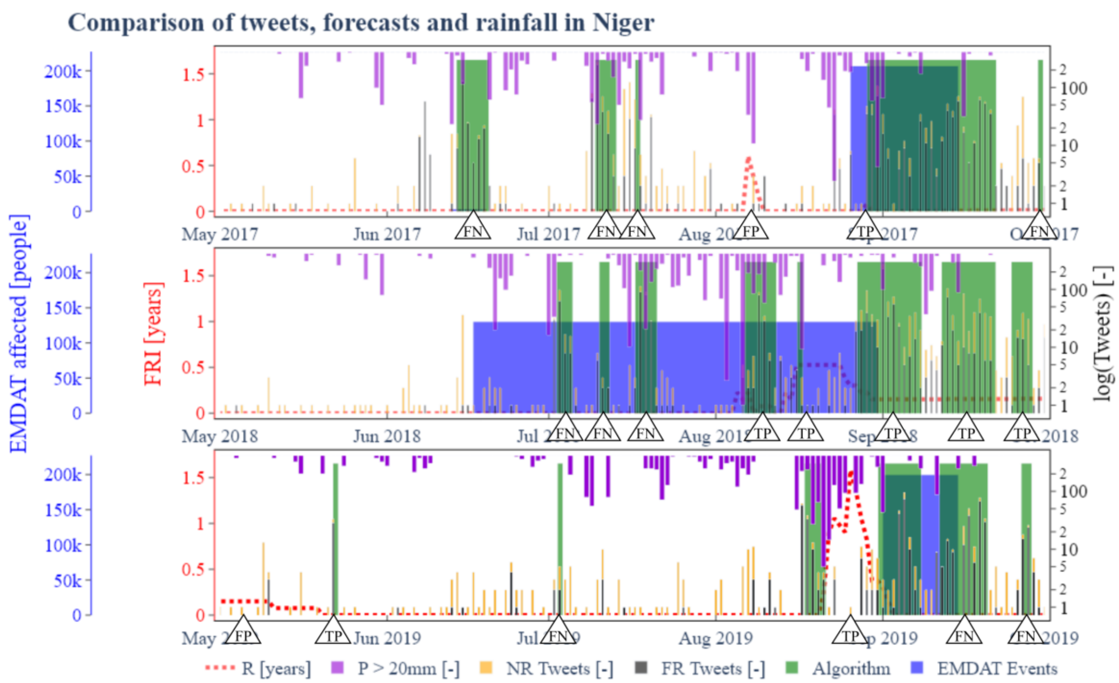


Figure 4-4: Time series of Niger for EMDAT events (blue bar), FRI (red, dashed), flood-related and -unrelated tweets (black and yellow bar), and precipitation exceedances (purple bar). The yellow bars display the tweets not related to flooding, and the black bars show the tweets that have been identified as flood-related by the classification algorithm. Furthermore, the duration of events identified by the detection algorithm is displayed in green. The intensity of the event can be deduced from the number of black bars inside the green area. True positives (TP), false positives (FP), and false negatives (FN) are indicated as triangles.

We did not define an FRI threshold for when to identify a TP, FP, or FN. For example, in July 2019, we identified two FPs in Ghana, where the FRI merely increased around 0.1 (Figure 4-3), which suggests that our assessment is strict. Conversely, in Niger in September 2017, we count a TP where the FRI is increased around 0.05 (Figure 4-4). This suggests a favorable assessment of the FANFAR forecast.

EMDAT and FEDA: In general, the events registered by EMDAT and FEDA overlap, except e.g. in Benin, Burkina Faso, Gambia, and Guinea (see Section B.2.1 in the Appendix). However, the detection algorithm indicates more events than EMDAT in all countries except Togo. Furthermore, EMDAT normally shows events that last longer than the events detected by FEDA, e.g. in Ghana, Mali, Niger, and Nigeria (see Section B.2.1 in the Appendix). On the one hand, this could imply that FEDA detects events at a higher temporal resolution than EMDAT. On the other hand, people tweet less towards the end of an event (de Bruijn et al., 2019) and this might lead to a prematurely ended event. Overall, the events detected by FEDA appear to indicate actual flooding (per EMDAT). However, a comparison to other disaster databases should be undertaken to back this claim.

FRI and FEDA: Overall, we identified 25 TP, 53 FP, and 41 FN (Table 4-7). In Niger and Nigeria, we registered the most TP. In the Gambia, Liberia, Sierra Leone, and Togo, no TP were counted. Also, we observe the largest number of FP in Ghana and Ivory Coast and the lowest number in Mali, Niger, and Sierra Leone. Finally, FNs are most prominent in Ghana and Niger, while in Ivory Coast, Mauritania and Togo we observe no FNs. The resulting POD and FAR imply that in Gambia, Liberia, and Sierra Leone, FANFAR does not capture flood events during the wet season. In the Ivory Coast and Mauritania, FANFAR forecasts floods accurately, but also tends to forecast floods when none occur. Overall, we do not observe a trend regarding the timing of increased FRI and FEDA detected events. In Nigeria for example, an increased FRI occurs between FEDA detected events, while in the Ivory Coast, an increased FRI comes after a FEDA detected event. This implies that in the wet season, no systematic influence (e.g. rainfall or aftermath of a flood) drives a lag between an increased FRI and FEDA detected event.

Table 4-7: True positives (TP), false positives (FP), and false negatives (FN). A TP is noted when an event detected by the algorithm and an increased FRI occur together. An FP is noted when an increased FRI occurs without a detected event. An FN is noted when FEDA detected an event without an increased FRI. POD_{FEDA} and FAR_{FEDA} indicate the probability of detection (POD) and false alarm rate (FAR) based on tweets. POD_s and FAR_s indicate the POD and FAR based on streamflow exceednaces. In the bottom row, the sum of TP, FP, and FN and the average of POD_{FEDA} and FAR_{FEDA} are noted.

Country	**TP**	**FP**	**FN**	**POD_{FEDA}**	**FAR_{FEDA}**	**POD_s**	**FAR_s**
Benin	2	4	3	0.4	0.67	-	-
Burkina Faso	1	3	1	0.5	0.75	-	-
Gambia	0	2	2	0	1	-	-
Ghana	3	10	11	0.21	0.77	-	-
Ivory Coast	1	9	0	1	0.9	-	-
Liberia	0	5	1	0	1	-	-
Mali	2	1	2	0.5	0.33	0.5-1.0	0-0.2
Mauritania	1	3	0	1	0.75	-	-
Niger	7	1	10	0.41	0.13	-	-
Nigeria	8	5	7	0.53	0.38	0.0-0.1	0-0.2
Senegal	1	3	1	0.5	0.75	-	-
Sierra Leone	0	1	2	0	1	0.4-0.6	0-0.2
Togo	0	6	0	N/A	1	-	-
Σ or average	25	53	41	0.38	0.68	-	-

Both the POD based on tweets (POD_{FEDA}) and based on streamflow (POD_s) show, that in the Upper Niger, floods are detected well (POD close to 1, Figure 4-5). In the Lower Niger, we observe that POD_{FEDA} suggests a better flood detection of FANFAR than POD_s. We expect the POD_{FEDA} to be larger than the POD_s since we did not include an FRI threshold in the counting of a TP (i.e. we identify a TP with already small changes on FRI). This results in a favorable POD assessment of the FANFAR forecasts (Equation 2 on p. 14).

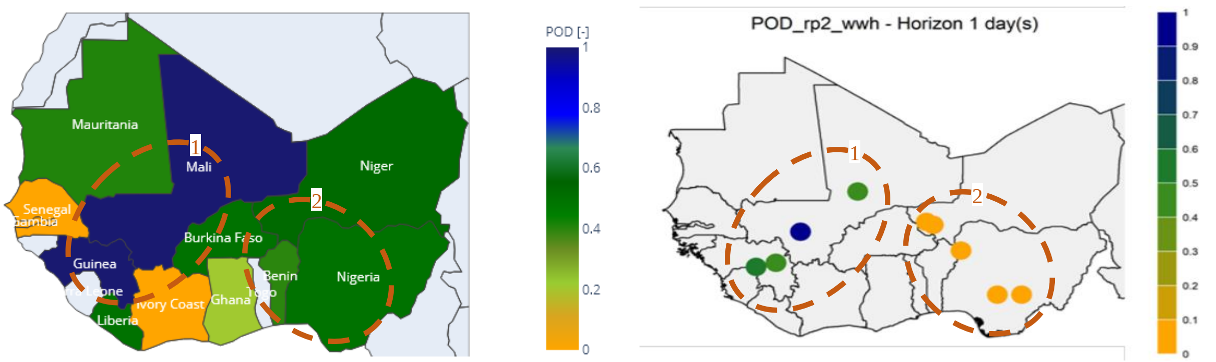


Figure 4-5: Probability of detection (POD) based on FEDA (POD_{FEDA}, left) and streamflow (right, see Section 3.1.3). The red-dashed circles indicate the Upper Niger (1) and Lower Niger (2) basin.

Both the FAR based on tweets (FAR_{FEDA}) and based on streamflow (FAR_s) show that in the Upper Niger, around half of the floods are falsely forecasted (FAR close to 0.5, Figure 4-6), with exception of one streamflow measurement in Guinea (Figure 4-6, right). In the Lower Niger, we observe a larger FAR_{FEDA} than FAR_s. We expect the FAR_{FEDA} to be larger than the FAR_s since we did not include an FRI threshold in identifying FP (i.e. we identify an FP with already small changes on FRI). This results in a strict FAR assessment of the FANFAR forecasts (Equation 3 on p. 14).

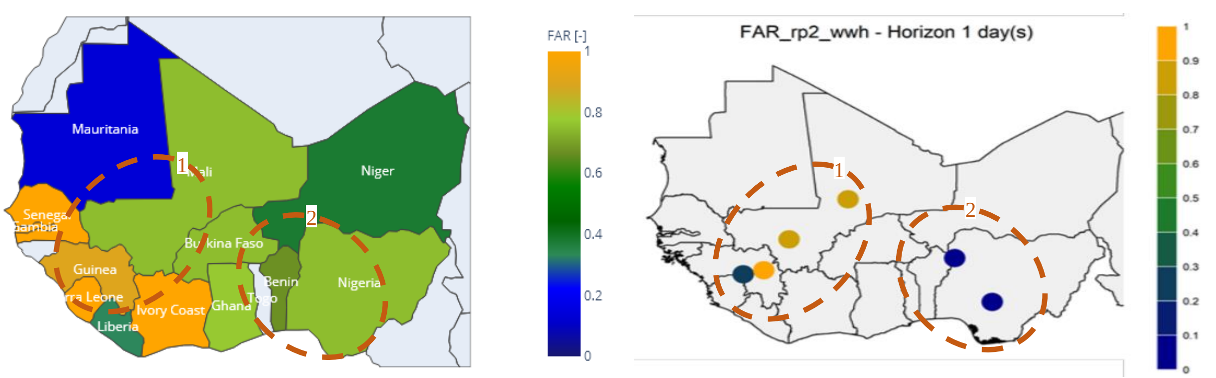


Figure 4-6: False alarm rate (FAR) based on FEDA (left) and streamflow (right, see Section 3.1.3). The red-dashed circles indicate the Upper Niger (1) and Lower Niger (2) basin.

The comparison of POD_s with POD_{FEDA} and FAR_s with FAR_{FEDA} implies that the FEDA could be used additionally to streamflow data to assess the performance of FANFAR. However, our estimate of POD and FAR in Mali and Sierra Leone could be inaccurate due to the low tweet activity (Table 4-6 on p. 29). Furthermore, we did not consider any other criteria to identify FP, like rainfall or regional flood protection measures. Since FANFAR predicts streamflow and an increased FRI does not necessarily indicate flooding (and therefore no FEDA detected event), we might have detected unjustified FP. Regional levels of the FRI that lead to flooding should be defined to identify FP better. We also did not consider the gravity of the FP and FN, i.e. we did not check how much the FRI was increased when FEDA did not detect an event, or how intense a FEDA event was when the FRI was not increased. Finally, we have to keep in mind, that POD_{FEDA} and FAR_{FEDA} are computed based on HydroGFD 1.0, while POD_s and FAR_s are computed based on HydroGFD 3.0 (see Section 3.1.3). An analysis of both scores should be undertaken, where the hydrological forcing is the same.

NRE and FEDA: As with the FRI, no trend regarding timing is observed. In all countries, the NREs sometimes occur without a FEDA event. Although this result implies that tweets are not always triggered by rainfall, we must keep in mind that a constant threshold of 20 mm/day over all countries might not be adequate. In every country, a different amount of rainfall could induce flooding, depending on e.g. drainage capacity. Furthermore, the country-wide and daily resolution of the lumped rainfall data is too coarse to consider cloud bursts that could cause flooding.

Correlation analysis of tweets with FRI

Overall, more than half of the correlation coefficients r_s are not significant at a 5 % level (Table 4-8 and Table 4-9), likely due to lack of data. Also, the majority of r_s lies between approx. 0 and 0.2. Although r_s gives insight into the performance of FANFAR, they need to be interpreted carefully. Other factors, such as commemoration events (de Bruijn et al., 2019) and rainfall, might influence tweeting activity.

Spearman and Pearson: The Pearson and Spearman correlation coefficients r_p and r_s are at most 0.1 different from each other (Table 4-8). The difference between r_p and r_s is to be expected, as the FRI in all countries and the tweets in countries with little tweet activity are zero-inflated and non-normal (Table 4-3 and Table 4-6 on p. 28 and p. 29 respectively). Therefore, r_s is likely a more reliable correlation measure than r_p (see e.g. Kuller et al., 2018).

All tweets and FEDA: Over all West African countries, the performance of FANFAR according to r_s does not vary much between countries (Figure 4-7, left). The correlation coefficient with the sensitive

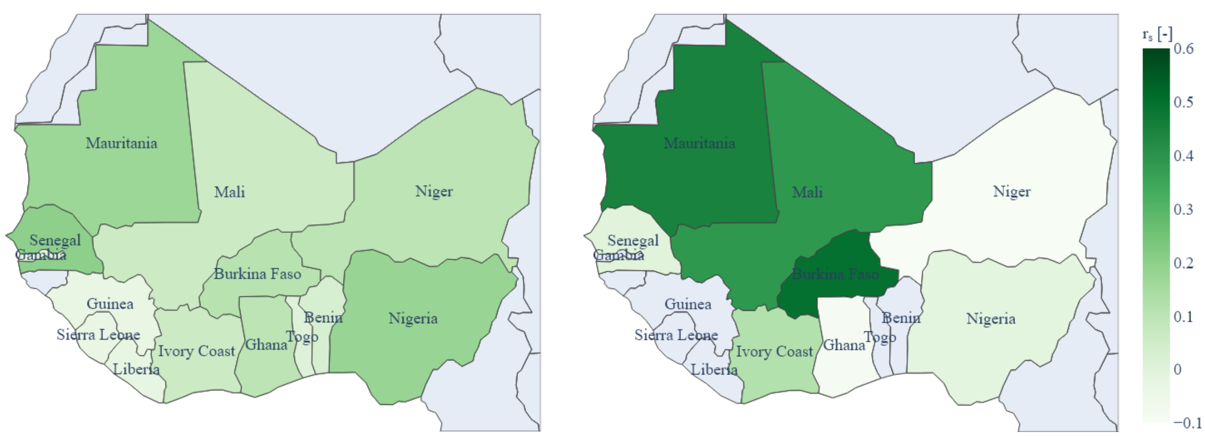


Figure 4-7: Spearman correlation coefficients between FRI and tweets for the whole time series r_s (left) and only during events predicted by FEDA with the sensitive configuration $r_{s,sen}$ (right).

FEDA configuration $r_{s,sen}$ is the highest in Burkina Faso, Mali, and Mauritania (Figure 4-7, right). However, again, this analysis should be interpreted carefully since most correlation coefficients are either close to 0 or not significant at a 5 % level (Table 4-8). The configuration of FEDA (sensitive, balanced, and strict) influences the correlation coefficients (Table 4-8). When comparing r_s with the correlation coefficients based on the sensitive, balanced, and strict settings of the FEDA $r_{s, sen}$, $r_{s, bal,}$ and $r_{s, str}$, we observe that in up to half the countries, the latter could not be calculated, and in most cases, the correlation is not significant. It is expected, that r_{strict} has more undefined values since there are fewer data points. Although in some countries FANFAR performs better with balanced or strict FEDA configuration, we suggest using the sensitive configuration to better identify FN. If an FN is detected, the FEDA event can still be confirmed or rejected based on the content of the tweets (de Bruijn et al., 2019).

Regional: In most countries $r_{s, regional}$ could not be computed (Table 4-8). Only taking FRI from one region increases its zero-inflation. We observe that the calculated $r_{s, regional}$ are usually lower than r_s, indicating that the assumption of all tweets occurring in the most populated region is not accurate. However, only one correlation coefficient is significant, pointing towards a lack of data in other countries.

Seasonal: We observe a larger r_s than $r_{s, season}$ in 10 out of 14 countries, with a maximum difference of around 0.1 (Table 4-8). We expected a higher r_s in most countries, as during the dry season, likely there will neither be an increased FRI nor increased tweet activity. However, $r_{s, season} > r_s$ could point to misclassified tweets in the dry season, e.g. due to a commemoration of a historic flood, or a pipe burst (de Bruijn et al., 2019). The use of single days of Twitter activity to conclude on the accuracy of FANFAR is therefore not adequate and FEDA should be applied for the assessment of FANFAR.

Table 4-8: Correlation coefficients between flood-related tweets and the FRI with different features. The features include type (Spearman S), Seasonal (if yes, only values from May-Oct are included), Event (if yes, only forecasts and tweets during an event registered by the algorithm are included), and Lag (shifted tweets time-series between 1-5 days). The best correlations are highlighted in bold for every country. The sum of the best correlations per column is shown in the bottom row. The underlined values indicate significance at a 5 % level. The p-values are listed in Table B-4 in the Appendix.

		$\mathbf{r_s}$	$\mathbf{r_p}$	$\mathbf{r_{s,\ sen}}$	$\mathbf{r_{s,\ bal}}$	$\mathbf{r_{s,\ str}}$	$\mathbf{r_{s,\ regional}}$	$\mathbf{r_{s,\ season}}$
Feature	*Type*	*S*	*P*	*S*	*S*	*S*	*S*	*S*
	Seasonal	*no*	*no*	*no*	*no*	*no*	*no*	*yes*
	Event	*no*	*no*	*sensitive*	*balanced*	*strict*	*sensitive*	*no*
	Regional	*no*	*no*	*no*	*no*	*no*	*yes*	*no*
	Lag [days]	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Country (Region)	*Benin (Cotonou)*	**0.03**	0.0	-	-	-	-0.01	-0.03
	Burkina Faso (Ouagadougou)	0.11	0.06	0.5	-1.0	-	-	**0.14**
	Gambia (Banjul)	**-0.01**	**-0.01**	-	-	-	**-0.01**	-0.03
	Ghana (Kumasi)	0.1	0.0	-0.09	0.19	**0.2**	0.08	0.02
	Guinea (Conakry)	-0.03	**-0.01**	-	-	-	-	-0.07
	Ivory Coast (Abidjan)	0.06	0.01	**0.12**	-0.15	-0.15	-	0.03
	Liberia (Monrovia)	-0.02	-0.01	-	-	-	-	-0.05
	Mali (Bamako)	0.06	0.04	0.39	**0.87**	0.83	-	0.18
	Mauritania (Nouackchoott)	0.17	0.19	**0.45**	**0.45**	**0.45**	-	0.16
	Niger (Niamey)	0.1	0.0	-0.14	-0.13	-0.08	-	**0.24**
	Nigeria (Lagos)	**0.18**	0.02	-0.01	0.03	0.01	0.04	0.04
	Senegal (Dakar)	0.2	**0.26**	0.0	0.13	0.13	-	0.24
	Sierra Leone (Freetown)	-0.03	**-0.01**	-	-	-	-	-0.05
	Togo (Lome)	**0.01**	-0.01	-	-	-	-0.02	-0.03
	Σ	4	4	2	2	2	1	2

Lag: The tweets and the FRI do not appear to be a certain number of days apart (Table 4-9). Comparing r_s and the lagged Spearman correlation coefficients from -5 to 5 days $r_{s,\ lag\text{-}5}$, $r_{s,\ lag\text{-}3}$, $r_{s,\ lag\ 3}$ and $r_{s,\ lag\ 5}$, we see that not a single lag is dominating the other lags in all countries. The assumption, that an increased tweet activity precedes or follows up on an increased FRI with a fixed amount of days, is therefore not valid. However, the mixed results indicate that the precedence of tweets might be depending on the country and the situation. For example, during the first rainfall of the rainy season, people might be more inclined to predict flooding and tweet about it than towards the end of the rainy season. This suggests an individual application of tweets and FEDA for every country to assess the performance of FANFAR.

Rainfall: Although we did not find an influence of rainfall on tweets in the visual analysis of the time series during the wet season, we observe that the Spearman correlation coefficient of the NREs with tweets $r_{s,rain}$ is larger than r_s, implying that tweets correlate better with rainfall than with the FRI (Table 4-9). However, $r_{s,\ rain}$ does not surpass 0.2, with the exception in Nigeria. The correlation between NREs and tweets only during FEDA detected events (sensitive configuration) $r_{s,rain,sen}$ yields overall the highest correlation coefficients, although most of them are non-significant. These results indicate that rainfall plays a role in tweets. This is not surprising, since it can cause pluvial flooding. FANFAR predicts streamflow and therefore fluvial flooding. This means that, to assess FANFAR with the help of tweets, the influence of rainfall should be accounted for, i.e. already during the classification stage of tweets (see de Bruijn et al., 2020).

Table 4-9: Correlation coefficients between flood-related tweets and the FRI with different features. The features include type (Pearson P, Spearman S), Seasonal (if yes, only values from May-Oct are included), Event (if yes, only forecasts and tweets during an event registered by the algorithm are included), and Lag (shifted tweets time-series between 1-5). The best correlation is highlighted in bold for every country (with exception of Gambia). The sum of the best correlations per column is shown in the bottom row. Underlined values are significant at a 5 % level. The p-values are listed in Table B-5 in the Appendix.

		r_s	$r_{s,\ lag\text{-}5}$	$r_{s,\ lag\text{-}3}$	$r_{s,\ lag\ 3}$	$r_{s,\ lag\ 5}$	$r_{s,\ rain}$	$r_{s,\ rain,\ sen}$
Feature	*Type*	*S*	*S*	*S*	*S*	*S*	*S*	*S*
	Seasonal	*no*	*no*	*yes*	*no*	*no*	*no*	*no*
	Event	*no*	*no*	*no*	*no*	*no*	*no*	*sensitive*
	Regional	*no*	*no*	*no*	*no*	*no*	*no*	*no*
	Lag [days]	*0*	*-5*	*-3*	*3*	*5*	*0*	*0*
Country	*Benin*	0.03	0.07	0.09	0.11	0.06	0.09	**0.13**
	Burkina Faso	0.11	-0.02	-0.02	0.05	-0.02	0.10	**0.95**
	Gambia	-0.01	**0.08**	-0.01	-0.01	-0.01	0.05	-0.54
	Ghana	0.1	0.08	0.08	0.06	0.08	**0.16**	-0.11
	Guinea	-0.03	0.02	0.05	0.08	0.08	0.14	**0.50**
	Ivory Coast	0.06	0.09	0.05	0.07	0.11	0.09	**0.33**
	Liberia	-0.02	0.04	0.03	-0.02	-0.02	**0.06**	nan
	Mali	0.06	0.04	0.05	0.07	0.09	**0.18**	-0.41
	Mauritania	0.17	0.03	0.03	0.22	0.17	0.07	**0.74**
	Niger	0.10	0.1	0.10	0.11	0.1	**0.20**	0.13
	Nigeria	0.18	0.15	0.14	0.18	0.15	**0.40**	-0.01
	Senegal	0.20	0.02	0.06	0.11	0.07	0.15	**0.24**
	Sierra Leone	-0.03	-0.03	0.01	-0.03	-0.03	0.17	**0.28**
	Togo	0.01	0.01	0.01	0.01	**0.05**	0.03	nan
	Σ	0	1	0	0	1	5	7

Drawbacks of correlation coefficients: It is debatable whether correlation coefficients are the best statistical measure to compare FEDA and FRI. In the future, a meaningful indicator for tweets, similarly to the FRI, should be developed, to use goodness-of-fit measures (GOFs) like the Root Mean Square Error (RMSE) or KGE. The development of tweets as an indicator is, however, difficult. The two time-series of tweets and FRI should be in the same order of magnitude to allow the application of GOFs and the comparison between countries. One approach would be to normalize tweets with ε, which indicates the tweet activity in every country (Table 4-6 on p. 29).

t-Test of the FRI during and outside of FEDA detected events

We observe that according to the Shapiro Wilk test, in no country both the *event* and *no event* group follow a normal distribution (Table 4-10). Mostly, the *no event* group follows a non-normal distribution (Table B-6 in the Appendix). This is due to the zero-inflation of the FRI when no events from the FEDA are detected. However, the sample sizes, with exception of Sierra Leone, are smaller than 30, rendering the Shapiro-Wilk test results unreliable (Mohd Razali and Yap, 2011). Although the Welch's test results are not reliable and must be interpreted with caution, the FRI does not seem to systematically come before or follow up on a FEDA detected event.

Table 4-10: p-values of the Welch's test p_{pre}, p, and p_{post} between tweets during a FEDA detected event and outside. p_{pre} includes the FRI 5 days before and during a detected event into the event group, p only includes the FRI during a detected event, and p_{post} includes the FRI during a detected event and 5 days after. In no case were both the event and no-event groups normally distributed according to the Shapiro-Wilk test. The results of the Shapiro-Wilk teste are displayed in Table B-6 in the Appendix.

Country	$\mathbf{p_{pre}}$	**p**	$\mathbf{p_{post}}$
Benin	0.96	0.0	0.16
Burkina Faso	0.23	0.22	0.23
Gambia	0.07	0.07	0.07
Ghana	0.0	0.04	0.01
Guinea	0.0	0.0	0.0
Ivory Coast	0.32	0.42	0.13
Liberia	0.0	-	0.0
Mali	0.05	0.47	0.04
Mauritania	0.23	0.2	0.02
Niger	0.0	0.21	0.0
Nigeria	0.15	0.22	0.0
Senegal	0.27	0.38	0.2
Sierra Leone	0.02	0.02	0.02
Togo	-	-	-

5. Chapter: Conclusion and Further Research

This thesis investigated two aspects of flood early warning systems: Firstly, we aimed to establish a baseline of current early flood warning dissemination practice in West African media. We addressed this aim by quantitatively analyzing the content of newspapers and tweets. Secondly, we aimed to assess the applicability of social media to evaluate the performance of FANFAR in West Africa. We employed a flood event detection algorithm based on Twitter and investigated correlations between the FANFAR forecasts and tweets.

Firstly, based on the newspaper and Twitter analysis, we conclude that little early warning is disseminated on these channels, although differences between countries apply. However, we found that the lack of a comprehensive media database, especially for radio and TV, complicated our investigation. We, therefore, suggest investigating every country individually regarding warning dissemination and search for local media archives. Furthermore, we suggest scrutinizing the links in tweets, as they might include further information about warning dissemination content and channel. Besides, other social media channels, such as Flickr, Facebook, or Whatsapp, should be considered. In conclusion, media, and especially social media, can be used as both a tool for research on warning dissemination and for warning dissemination itself. In the future, FEWSs, such as FANFAR, could also include tweets and other social media to raise awareness during an on-going flood event in their framework, and disseminate this information to stakeholders. This would add the “response” element to the early warning element of a FEWS.

Secondly, based on the flood detection algorithm employing Twitter, we conclude that social media can be used in the assessment of flood early warning systems. However, we found that social media alone does not suffice to assess the performance of a flood early warning system. The challenge to find the ground truth, i.e. does social media, flood event databases, newspaper archives, or other information sources identify flooding most reliably, remains. Furthermore, language detection likely limits the use of social media in flood detection. We also found that additional criteria, like rainfall or flood risk thresholds, need to be included in the forecast assessment. This is especially important if a FEWS, like FANFAR, predicts streamflow return periods and not actual flooding. In the case of FANFAR, we, therefore, suggest assessing the forecasts regionally, e.g. apply different flood risk thresholds for every region. Furthermore, rainfall (e.g. with the help of TAMSAT, see Ogbu et al., 2020) should be scrutinized, such that events that FANFAR could not predict (i.e. pluvial flooding) can be excluded from the forecast assessment. In general, we suggest investigating other social media, such as Flickr or Facebook, also regarding images or videos, against which FEWS forecasts can be evaluated. Besides, the potential of social media emergence to detect areas at risk of flooding (i.e. trace a flood through space and time) should be analyzed.

This thesis outlined the potential of media and social media in both early warning dissemination and FEWS assessment. In a nutshell, Twitter is not yet actively used in flood early warning dissemination in West Africa. Also, we could show that tweets alone are not adequate to assess the performance of FANFAR and likely other FEWSs. However, we expect an increase in social media use in the future. This could improve flood early warning dissemination and flood detection algorithms based on social media, and could even open up new possibilities to include social media beyond early warning dissemination and the assessment of a FEWS.

Bibliography

Adekola and Lamond, 2017 Olalekan Adekola and Jessica Lamond. A media framing analysis of urban flooding in Nigeria: current narratives and implications for policy. *Regional Environmental Change*, 18 (4): 1145–1159, November 2017. doi: 10.1007/s10113-017-1253-y. URL https://doi.org/-10.1007/s10113-017-1253-y.

Ahad and Yahaya, 2014 Nor Aishah Ahad and Sharipah Soaad Syed Yahaya. Sensitivity analysis of welch's t-test. 2014. doi: 10.1063/1.4887707.

Ahadzie and Proverbs, 2010 D. K. Ahadzie and D. G. Proverbs. Flooding and post flooding response strategies in ghana. In *Flood Recovery, Innovation and Response II*. WIT Press, May 2010. doi: 10.2495/friar100241. URL https://doi.org/10.2495/friar100241.

Ahadzie and Proverbs, 2011 D.K. Ahadzie and D.G. Proverbs. Emerging issues in the management of floods in ghana. *International Journal of Safety and Security Engineering*, 1 (2): 182–192, August 2011. doi: 10.2495/safe-v1-n2-182-192. URL https://doi.org/10.2495/safe-v1-n2-182-192.

Al-Rfou et al., 2015 Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. POLYGLOT-NER: massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM international conference on data mining*, pages 586–594, 2015. doi: https://doi.org/10.1137/1.9781611974010.66.

Amitay et al., 2004 Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-Where: Geotagging Web Content. In *Proceedings of the 27th annual international conference on research and development in information retrieval - SIGIR '04*, pages 273–280, 2004. doi: https://doi.org/10.1145/1008992.1009040.

Andersson, 2020a Jafet Andersson. Report documenting and explaining the hydrological models. Report: Deliverable D3.2, FANFAR, 2020. URL https://fanfar.eu/wp-content/uploads/sites/4/2020/-05/FANFAR-D3.2-Hydrological-Models.pdf.

Andersson, 2020b Jafet Andersson. EN: FANFAR Forecast service, February 2020. URL https://-knowledge.terradue.com/display/FANFAR/EN: FANFAR Forecast service.

Baudoin et al., 2014 Marie-Ange Baudoin, Sarah Henly-Shepard, Nishara Fernando, Asha Sitati, and Zinta Zommers. Early warning systems and livelihood resilience: Exploring opportunities of community participation. UNU-EHS Working Paper Series No. 11, Bonn: United Nations University Institute of Environment and Human Security (UNU-EHS), 2014.

Berg et al., 2020 Peter Berg, Fredrik Almén, and Denica Bozhinova. HydroGFD3.0: a 250.167emkm global near real-time updated precipitation and temperature data set. sep 2020. doi: https://doi.org/10.5194/essd-2020-236.

Berger, 1998 Arthur Berger. *Media research techniques*. SAGE Publications Inc, Thousand Oaks, Calif, 1998. ISBN 9781452251264.

Besada and Sewankambo, 2009 Hany Besada and Nelson K. Sewankambo. Climate change in africa: Adaptation, mitigation and governance challenges. Special report, The Centre for International Governance Innovation, Ontario, Canada, 2009.

Castillo, 2016 Carlos Castillo. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press, Cambridge, first edition, 2016. ISBN 978-1107135765.

Cools et al., 2016 Jan Cools, Demetrio Innocenti, and Sarah O'Brien. Lessons from flood early warning systems. *Environmental Science & Policy*, 58: 117–122, April 2016. doi: 10.1016/j.envsci.2016.01.006. URL https://doi.org/10.1016/j.envsci.2016.01.006.

CRED, 2020 CRED. FAQs. https://www.emdat.be/frequently-asked-questions, 2020. Accessed: 2020-07-22.

de Albuquerque et al., 2015 João Porto de Albuquerque, Benjamin Herfort, Alexander Brenning, and Alexander Zipf. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29 (4): 667–689, February 2015. doi: 10.1080/13658816.2014.996567. URL https://doi.org/10.1080/13658816.2014.996567.

de Bruijn et al., 2017 Jens A. de Bruijn, Hans de Moel, Brenden Jongman, Jurjen Wagemaker, and Jeroen C. J. H. Aerts. TAGGS: Grouping Tweets to Improve Global Geoparsingfor Disaster Response. *Journal of Geovisualization and Spatial Analysis*, 2 (2), 2017. doi: https://doi.org/10.1007/s41651-017-0010-6.

de Bruijn et al., 2019 Jens A. de Bruijn, Hans de Moel, Brenden Jongman, Marleen C. de Ruiter, Jurjen Wagemaker, and Jeroen C. J. H. Aerts. A global database of historic and real-time flood events based on social media. *Scientific Data*, 6 (1), December 2019. doi: 10.1038/s41597-019-0326-9. URL https://doi.org/10.1038/s41597-019-0326-9.

de Bruijn et al., 2020 Jens A. de Bruijn, Hans de Moel, Albrecht H. Weerts, Marleen C. de Ruiter, Erkan Basar, Dirk Eilander, and Jeroen C.J.H. Aerts. Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. *Computers & Geosciences*, 140, July 2020. doi: 10.1016/j.cageo.2020.104485. URL https://doi.org/10.1016/j.cageo.2020.104485.

Dessalegn and Akalu, 2015 Obsi Gemeda Dessalegn and Dafisa Sima Akalu. The impacts of climate change on African continent and the way forward. *Journal of Ecology and The Natural Environment*, 7 (10): 256–262, October 2015. doi: 10.5897/jene2015.0533. URL https://doi.org/10.5897/-jene2015.0533.

EM-DAT Public, 2020 EM-DAT Public. Disaster Data. https://public.emdat.be/data, 2020. Accessed: 2020-07-16.

Emerton et al., 2016 Rebecca E. Emerton, Elisabeth M. Stephens, Florian Pappenberger, Thomas C. Pagano, Albrecht H. Weerts, Andy W. Wood, Peter Salamon, James D. Brown, Niclas Hjerdt, Chantal Donnelly, Calum A. Baugh, and Hannah L. Cloke. Continental and global scale flood forecasting systems. *WIREs Water*, 3 (3): 391–418, feb 2016. doi: 10.1002/wat2.1137.

ETH Library, 2020 ETH Library. Nexis uni. https://library.ethz.ch/en/locations-and-media/media-types/databases-standards-patents/nexis-uni.html, 2020. Accessed: 27/11/2020.

Fuchs et al., 2013 Georg Fuchs, Natalia Andrienko, Gennady Andrienko, Sebastian Bothe, and Hendrik Stange. Tracing the German centennial flood in the stream of tweets. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD 13*. ACM Press, 2013. doi: 10.1145/2534732.2534741. URL https://-doi.org/10.1145/2534732.2534741.

Ganiyu et al., 2016 Mutiu Adekunle Ganiyu, Mohammed Rosli, and Ismail Awan. Evaluating the Participation of Nigerian Newspapers in Flood Disaster Risk Reduction. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, 30 (4): 95–108, 2016.

Ganiyu et al., 2017a Mutiu Adekunle Ganiyu, Mohammed Rosli, and Ismail Awan. DISASTER RISK REDUCTION: FRAMING OF FLOOD DISASTER IN NIGERIAN NEWSPAPERS. *Journal of Social Sciences and Humanities*, 12 (1): 151–164, 2017.

Ganiyu et al., 2017b Mutiu Adekunle Ganiyu, Mohammed Rosli, and Ismail Awan. Effective Media Involvement in Flood Disaster Management in Nigeria: Pressing Problems and Recommendations. *Asia Pacific Journal of Education, Arts and Sciences*, 4 (1): 120–131, January 2017.

Gut, 2010 Ulrike Gut. Standards of English in West Africa. pages 213–228, 2010. doi: 10.1017/CBO9781139023832.012.

Gwimbi, 2007 Patrick Gwimbi. The effectiveness of early warning systems for the reduction of flood disasters: Some experiences from cyclone induced floods in Zimbabwe. 9, 01 2007.

Hauke and Kossowski, 2011 Jan Hauke and Tomasz Kossowski. Comparison of Values of Pearsons and Spearmans Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, 30 (2): 87–93, jun 2011. doi: 10.2478/v10117-011-0021-1.

Höppner et al., 2010 Corina Höppner, Matthias Buchecker, and Michael Bründl. *Risk Communication and Natural Hazards*. CapHaz-Net, WP5. Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) and WSL Institute for Snow and Avalanche Research SLF – Birmensdorf & Davos, Birmensdorf, Switzerland, November 2010.

Horita et al., 2017 Flávio E.A. Horita, João Porto de Albuquerque, Victor Marchezini, and Eduardo M. Mendiondo. Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in brazil. *Decision Support Systems*, 97: 12–22, May 2017. doi: 10.1016/j.dss.2017.03.001. URL https://doi.org/10.1016/-j.dss.2017.03.001.

Imran et al., 2015 Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computer Surveys*, 47 (67), 2015. doi: https://doi.org/10.1145/2771588.

Keijsers, 2010 N.L.W. Keijsers. Neural Networks. In *Encyclopedia of Movement Disorders*, pages 257–259. Elsevier, 2010. doi: 10.1016/b978-0-12-374105-9.00493-7.

Knoben et al., 2019 Wouter J. M. Knoben, Jim E. Freer, and Ross A. Woods. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23 (10): 4323–4331, October 2019. doi: 10.5194/hess-23-4323-2019. URL https://doi.org/10.5194/hess-23-4323-2019.

Kuller et al., 2018 Martijn Kuller, Peter M. Bach, Diego Ramirez-Lovering, and Ana Deletic. What drives the location choice for water sensitive infrastructure in Melbourne, Australia? *Landscape and Urban Planning*, 175: 92–101, jul 2018. doi: https://doi.org/10.1016/j.landurbplan.2018.03.018.

Leetaru et al., 2013 Kalev Leetaru, Shaowen Wang, Anand Padmanabhan, and Eric Shook. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18 (5), May 2013. doi: 10.5210/fm.v18i5.4366. URL https://doi.org/10.5210/fm.v18i5.4366.

Li et al., 2012 C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: Named entity recognition in targeted Twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 721–730. ACM, 2012.

Lindström, 2017 Göran Lindström. Introduction to the HYPE model. https://www.youtube.com/-watch?v=JZBZdHuwS5s, Nov 2017.

Ling Chen, 2009 Abhishek Roy Ling Chen. Detection from Flickr Data through Wavelet-based Spatial Analysis. 2009.

Lopez, n.d. Rodrigo Contreras Lopez. Alerta geo. https://www.alertageo.org/, n.d. Accessed: 1/12/2020.

Lorini et al., 2019 V. Lorini, C. Castillo, F. Dottori, M. Kalas, D. Nappo, and P. Salamon. Integrating Social Media into a Pan-European Flood Awareness System: A Multilingual Approach. *CoRe Paper – Social Media in Crisis and Conflicts*, 2019.

Lucas et al., 2014 Ji Lucas, Muhammad Imran, Carlos Castillo, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. *International World Wide Web Conference Committee*, 2014. doi: http://dx.doi.org/10.1145/2567948.2577034.

Macnamara, 2005 Jim Macnamara. Media content analysis: Its uses, benefits and best practice methodology. *Asia-Pacific Public Relations Journal*, 6, 01 2005.

Madamombe, 2004 EK Madamombe. CASE STUDY ZIMBABWE: FLOOD MANAGEMENT PRACTICES – SELECTED FLOOD PRONE AREAS ZAMBEZI BASIN. Report, World Meteorological Organization and Global Water Partnership, 2004.

Mazzoleni et al., 2017 Maurizio Mazzoleni, Martin Verlaan, Leonardo Alfonso, Martina Monego, Daniele Norbiato, Miche Ferri, and Dimitri P. Solomatine. Can assimilation of crowdsourced data in hydrological modelling improve flood prediction? *Hydrology and Earth System Sciences*, 21 (2): 839–861, February 2017. doi: 10.5194/hess-21-839-2017. URL https://doi.org/10.5194/hess-21-839-2017.

Mohd Razali and Yap, 2011 Nornadiah Mohd Razali and Bee Yap. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *J. Stat. Model. Analytics*, 2, January 2011.

Musaev et al., 2015 Aibek Musaev, De Wang, and Calton Pu. Litmus: A multi-service composition system for landslide detection. *IEEE Transactions on Services Computing*, 8 (5): 715–726, 2015.

Naoko Nitta and Babaguchi, 2014 Tomochika Kato Naoko Nitta, Yusuke Kumihashi and Noboru Babaguchi. *MultiMedia Modeling*, chapter Real-World Event Detection Using Flickr Images, pages 307–314. Springer, Dublin, Ireland, January 2014.

Neuendorf, 2002 Kimberly A. Neuendorf. *The Content Analysis Guidebook*. SAGE Publications, Inc., Thousand Oaks, California 91320, 2 edition, 2002. ISBN 9781412979474.

Neuman, 2013 W. Neuman. *Social Research Methods: Pearson New International Edition*. Pearson Education Limited, 2013. ISBN 1292020237. URL https://www.ebook.de/de/product/30394935/-w_neuman_social_research_methods_pearson_new_international_edition.html.

Newbold et al., 2002 Chris Newbold, Oliver Boyd-Barret, and Hilde Van den Bulck. *The media book*. Arnold Oxford University Press, London New York, 2002. ISBN 0340740485.

Ogbu et al., 2020 Kingsley N. Ogbu, Nina Rholan Houngué, Imoleayo E. Gbode, and Bernhard Tischbein. Performance evaluation of satellite-based rainfall products over nigeria. *Climate*, 8 (10): 103, sep 2020. doi: doi:10.3390/cli8100103.

Palen and Anderson, 2016 Leysia Palen and Kenneth M. Anderson. Crisis informatics – new data for extraordinary times. *Science AAAS*, 353 (6296): 224–225, 2016.

Perera et al., 2019 Duminda Perera, Ousmane Seidou, Jetal Agnihotri, Mohamed Rasmy, Vladimir Smakhtin, Paulin Coulibaly, and Hamid Mehmood. *Flood Early Warning Systems: A Review Of Benefits, Challenges And Prospects*. UNU-INWEH Report Series, Issue 08. United Nations University Institute for Water, Environment and Health, Hamilton, Canada, 2019.

Ping et al., 2016Neoh Siew Ping, Uta Wehn, Chris Zevenbergen, and Pieter van der Zaag. Towards two-way flood risk communication: current practice in a community in the UK. *Journal of Water and Climate Change*, 7 (4): 651–664, October 2016. doi: 10.2166/wcc.2016.015. URL https://doi.org/-10.2166/wcc.2016.015.

Power et al., 2014 Robert Power, Bella Robinson, John Colton, and Mark Cameron. Emergency situation awareness:twitter case studies. In F. BÂ´enaben C. Hanachi and F. Charoy, editors, *Information Systems forCrisis Response and Managementin Mediterranean Countries*, pages 218–231, Toulouse, France, October 2014. ISCRAM-med, Springer.

Restrepo-Estrada et al., 2018 Camilo Restrepo-Estrada, Sidgley Camargo de Andrade, Narumi Abe, Maria Clara Fava, Eduardo Mario Mendiondo, and João Porto de Albuquerque. Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring. *Computers & Geosciences*, 111: 148–158, February 2018. doi: 10.1016/j.cageo.2017.10.010. URL https://doi.org/10.1016/j.cageo.2017.10.010.

Riley, 2008 W J Riley. Algorithms for frequency jump detection. *Metrologia*, 45 (6): S154–S161, dec 2008. doi: 10.1088/0026-1394/45/6/s21.

Rollason et al., 2018 E. Rollason, L. J. Bracken, R. J. Hardy, and A. R. G. Large. Rethinking flood risk communication. *Natural Hazards*, (92): 1665–1686, 2018.

Rossi et al., 2018 C. Rossi, F.S. Acerbo, K. Ylinen, I. Juga, P. Nurmi, A. Bosca, F. Tarasconi, M. Cristoforetti, and A. Alikadic. Early detection and information extraction for weather-induced floods using social media streams. *International Journal of Disaster Risk Reduction*, 30: 145–157, September 2018. doi: 10.1016/j.ijdrr.2018.03.002. URL https://doi.org/10.1016/j.ijdrr.2018.03.002.

Ruggiero and Vos, 2014 Aino Ruggiero and Marita Vos. Communication challenges in CBRN terrorism crises: Expert perceptions. *Journal of Contingencies and Crisis Management*, 23 (3): 138–148, September 2014. doi: 10.1111/1468-5973.12065. URL https://doi.org/10.1111/1468-5973.12065.

Schulz et al., 2013 Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. A Multi-Indicator Approach for Geolocalization of Tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

Seeger et al., 2018 Matthew W. Seeger, Laura E. Pechta, Simani M. Price, Keri M. Lubell, Dale A. Rose, Saloni Sapru, Melanie C. Chansky, and Belinda J. Smith. A conceptual model for evaluating emergency risk communication in public health. *Health Security*, 16 (3): 193–203, June 2018. doi: 10.1089/hs.2018.0020. URL https://doi.org/10.1089/hs.2018.0020.

Sheard, 2018 Judithe Sheard. Quantitative data analysis. pages 429–452, 2018. doi: http://dx.doi.org/10.1016/B978-0-08-102220-7.00018-2.

Simon et al., 2015 Tomer Simon, Avishay Goldberg, and Bruria Adini. Socializing in emergenciesa review of the use of social media in emergency situations. *International Journal of Information Management*, 35 (5): 609–619, October 2015. doi: 10.1016/j.ijinfomgt.2015.07.001. URL https://-doi.org/10.1016/j.ijinfomgt.2015.07.001.

SMHI, 2018 SMHI. *The code of HYSS and HYPE overview*. Swedish Meteorological and Hydrological Institute (SMHI), 2018. Retrieved from: https://sourceforge.net/projects/hype/files/-latest_docs/, Accessed: 08-07-2020.

SMHI, 2019 SMHI. What is HydroGFD2.0, 2019. URL http://hypeweb.smhi.se/wp-content/uploads/-2019/04/What-is-HydroGFD2.pdf. Accessed: 2020-07-09.

SMHI, 2020a SMHI. FANFAR: Reinforced cooperation to provide operational flood forecasting and alerts in West Africa. https://fanfar.eu, 2020. URL https://fanfar.eu. Accessed: 2020-07-08.

SMHI, 2020b SMHI. FANFAR: Reinforced cooperation to provide operational flood forecasting and alerts in West Africa. https://fanfar.eu/about, 2020. URL https://fanfar.eu/about. Accessed: 2020-03-07.

SMHI, 2020c SMHI. *HYPE model description*. Swedish Meteorological and Hydrological Institute (SMHI), 2020. Retrieved from: https://sourceforge.net/projects/hype/files/latest_docs/, Accessed: 08-07-2020.

SMHI, 2020d SMHI. *HYPE version 5.9.0*. Swedish Meteorological and Hydrological Institute (SMHI), 2020. Retrieved from: https://sourceforge.net/projects/hype/files/latest_docs/, Accessed: 08-07-2020.

Smith et al., 2015 L. Smith, Q. Liang, P. James, and W. Lin. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management*, 10 (3): 370–380, mar 2015. doi: 10.1111/jfr3.12154. URL https://doi.org/10.1111/-jfr3.12154.

Statcounter, 2020 Statcounter. Social Media Stats Africa. https://gs.statcounter.com/social-media-stats/all/africa/, August 2020. URL https://gs.statcounter.com/social-media-stats/all/africa/2017. Accessed: 07-08-2020.

Twitter, 2009 Twitter. Location, location, location, August 2009. URL http://blog.twitter.com/2009/08/-location-location-location.html. Accessed: 13-8-2020.

Twitter, 2020a Twitter. Tweet Location FAQs, 2020. URL https://help.twitter.com/en/safety-and-security/tweet-location-settings. Accessed: 13-8-2020.

Twitter, 2020b Twitter. How to add your location to a Tweet, 2020. URL https://help.twitter.com/en/-using-twitter/tweet-location. Accesserd: 13-8-2020.

Twitter, n.d. Twitter. Getting started: Using the twitter api. https://developer.twitter.com/en/docs/-twitter-api/getting-started/guide, n.d. Accessed: 12/3/2020.

UN-SPIDER, 2020 UN-SPIDER. AGRHYMET Regional Centre, 2020. URL http://www.un-spider.org/links-and-resources/institutions/agrhymet-regional-centre. Accessed: 15-08-2020.

UNDRR, 2006 UNDRR. Global Survey of Early Warning Systems: An assessment of capacities, gaps and opportunities towards building a comprehensive global early warning system for all natural hazards. Technical report, United Nations Office for Disaster Risk Reduction (UNDRR), 2006.

UNEP, 2012 UNEP. Early warning systems: A state of the art analysis and future directions. Division of early warning and assessment (DEWA), United Nations Environment Programme (UNEP), Nairobi, 2012.

UNISDR, 2009 UNISDR. Terminology on disaster risk reduction. Report, United Nations International Strategy for Disaster Reduction (UNISDR), Geneva, Switzerland, 2009.

UNISDR, 2017 UNISDR. Report of the open ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction. Technical report, United Nations Office of Disaster Risk Reduction, 2017. URL https://www.preventionweb.net/files/-50683_oiewgreportenglish.pdf. United Nations General Assembly. Agenda Item: 19 (C).

van Niekerk and Nemakonde, 2017 Dewald van Niekerk and Livhuwani David Nemakonde. *Natural Hazards and Their Governance in Sub-Saharan Africa*. Oxford University Press, Sep 2017. ISBN 9780199389407. doi: 10.1093/acrefore/9780199389407.013.230. URL http://-naturalhazardscience.oxfordre.com/view/10.1093/acrefore/9780199389407.001.0001/acrefore-9780199389407-e-230.

Vieweg, 2012 Sarah E. Vieweg. *Situational Awareness in Mass Emergency:A Behavioral and Linguistic Analysis of Microblogged Communications*. phdthesis, University of Colorado, Colorado, United States, 2012.

Weber, 1990 Robert Weber. *Basic Content Analysis*. SAGE Publications, Inc., 1990. doi: 10.4135/9781412983488.

WMO, 2011 WMO. Manual on flood forecasting and warning. Technical report, World Meteorological Organization, 2011. URL http://www.wmo.int/pages/prog/hwrp/publications/-flood_forecasting_warning/WMO 1072_en.pdf.

Wohlin, 2014 Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on*

Evaluation and Assessment in Software Engineering - EASE 14. ACM Press, 2014. doi: 10.1145/2601248.2601268.

Yin et al., 2012 Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using Social Media to Enhance Emergency Situation Awareness. *IEEE Computer Society*, November 2012.

Yin et al., 2015 Jie Yin, Sarvnaz Karimi, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using Social Media to Enhance Emergency Situation Awareness: Extended Abstract. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015.

A. Appendix: Methods

A.1 Case Study: FANFAR, a FEWS in West Africa

A.1.1 FANFAR

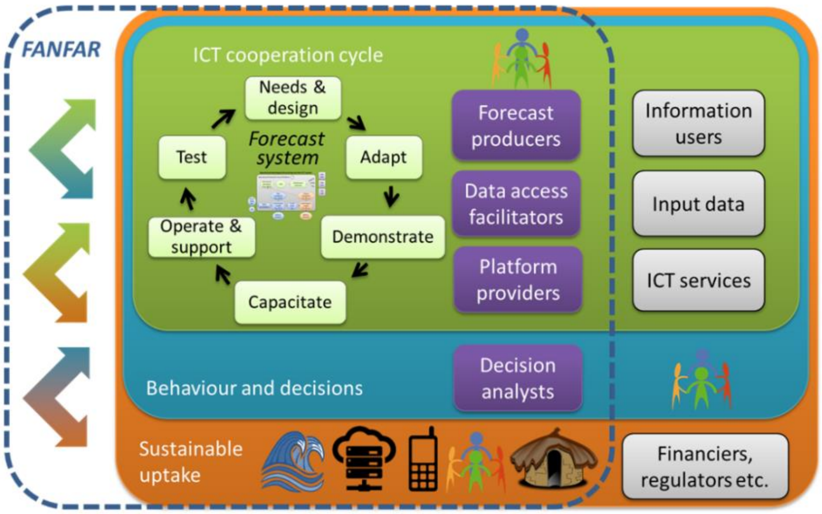


Figure A-1: The dimensions of FANFAR. The green dimension shows the information and communication technology ICT, the blue dimension displays behavior and decisions and the orange dimension is concerned with a sustainable implementation of the project. Taken from SMHI, 2020b.

A.1.2 The HYPE Model

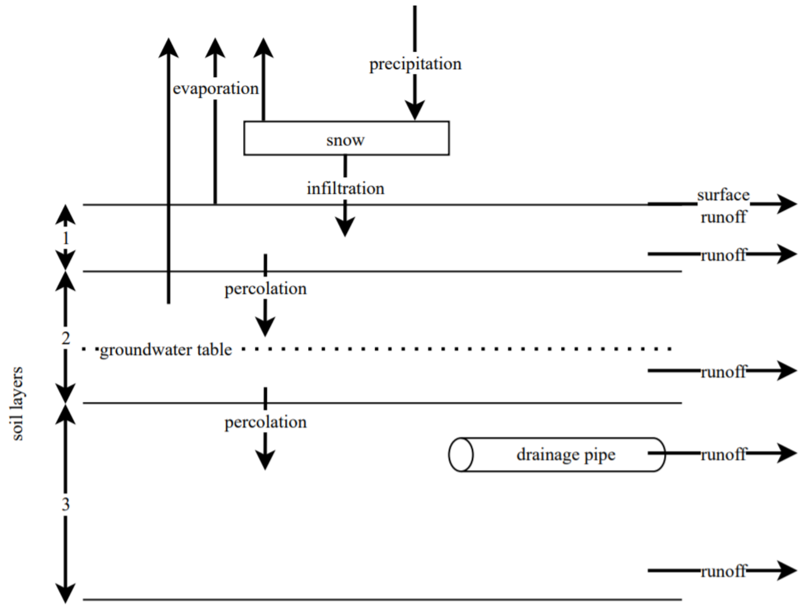


Figure A-2: Water flows in the vertical component (HRU) in HYPE. The numbers 1-3 indicate the soil layer. Adapted from SMHI, 2020d.

A.2 Twitter and FANFAR

A.2.1 Data Collection

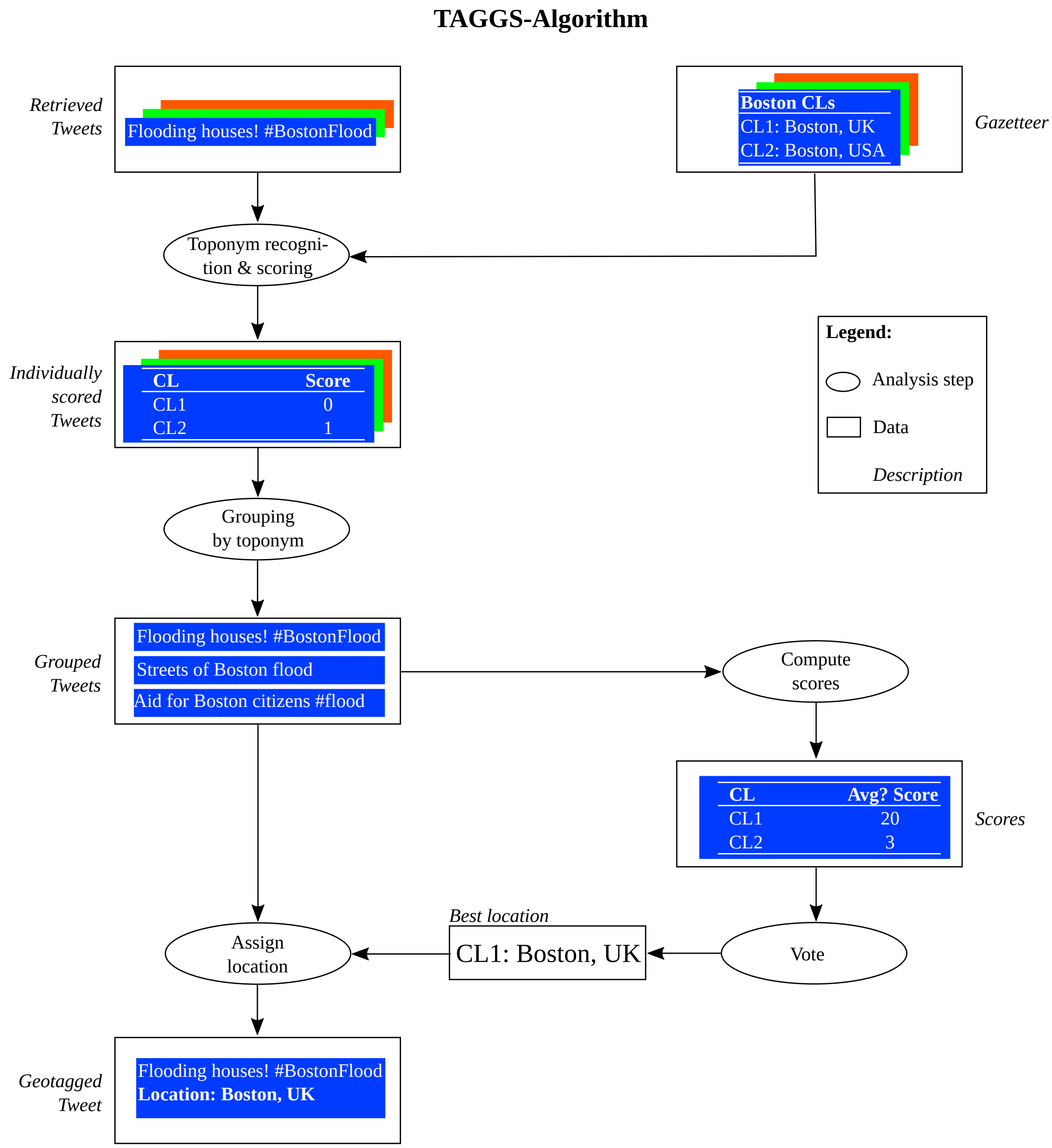


Figure A-3: Overview of the TAGGS to geotag Tweets. (1) In the toponym (i.e. name of a location) recognition, contiguous words are compared to a gazetteer (dataset containing known location names). For every toponym mentioned in the text, one or more candidate locations CL are extracted. (2) A scoring system is then applied to choose the most probable CL, based on additional info Twitter users can include in their profile or tweet (time zone, user hometown, other locations). (3) The tweets are then grouped under the assumption, that tweets containing identical toponyms in a certain timeframe refer to the same. (4) Subsequently, within every group, the average score of every candidate location is computed. For example, if 5 Tweets are in a group with three candidate locations, the 5 individual scores (calculated in the scoring step) of every Tweet are averaged for every candidate location. This yields 3 averaged values (one for every candidate location. (5) Finally, the CL with the highest score is assigned to the tweet. Adapted from de Bruijn et al., 2017.

B. Appendix: Results and Discussion

B.1 Flood Warning Dissemination in West Africa

B.1.1 Data Collection

Table B-1: Meta-sources ordered alphabetically, that show lists of newspaper, TV or radio sources, and the corresponding weblinks.

Name of source	Weblink
Geopoll	https://www.geopoll.com/blog/ghana-media-measurement-report-top-tv-radio-print-outlets-2017/
ilissafrica	https://ilissafrica.wordpress.com/2012/06/21/guide-african-newspapers/
Onlinenewspapers	http://www.onlinenewspapers.com/africa-newspapers.htm
Stanford Library	https://library.stanford.edu/areas/african-collections/journal-and-newspaper-articles-databases

Table B-2: News sources ordered alphabetically and the corresponding weblinks.

Name of source	Link
Abidjan news	https://news.abidjan.net/
Afrol News	http://www.afrol.com/html/archive/archive_countries.htm
All Africa	https://allafrica.com/list/publisher/editorial/editorial/type/pub.html
Archive.org	https://archive.org/
BBC Africa	https://www.bbc.com/news/world/africa
Bing news search API	https://docs.microsoft.com/en-us/rest/api/cognitiveservices-bingsearch/bing-news-api-v7-reference
Center of Research Libraries	https://www.crl.edu/focus/article/6694
News API	https://newsapi.org/
Newsbank	https://www.newsbank.com/
NexisUni	https://www.lexisnexis.com/en-us/gateway.page
Nordiska Afrikainstitutet	https://nai.uu.se/library/about/contact-the-library.html
Panapress	https://www.panapress.com/archive-lang2.html
Radio Nigeria	https://www.radionigeria.gov.ng/tag/weather/
ReliefWeb	https://reliefweb.int/updates?advanced-search=%28F12.F12570%29&search=ghana+flood
Reuters	https://af.reuters.com/news/archive/worldNews
Stanford Library xSearch	https://xsearch.stanford.edu/search/desktop/en/search/viewId:Karen%20Fung-African%20Studies/

B.2 Assessing FANFAR with Twitter

B.2.1 Data Analysis

Table B-3: Correlation coefficients between flood-related tweets and the FRI with different features. The features include type (Spearman S), Seasonal (if yes, only values from May-Oct are included), Event (if yes, only forecasts and tweets during an event registered by the algorithm are included), and Lag (shifted tweets time-series between -5 and 5 days).

		r_s	p_s	r_p	p_p	$r_{s,\,sen}$	$p_{s,sen}$	$r_{s,\,bal}$	$p_{s,bal}$	$r_{s,\,str}$	$p_{s,str}$	$r_{s,\,reg}$	$p_{s,reg}$	$r_{s,\,seas}$	$p_{s,seas}$
Feature	*Type*	*S*		*P*		*S*		*S*		*S*		*S*		*S*	
	Seasonal	*no*		*no*		*no*		*no*		*no*		*no*		*yes*	
	Event	*no*		*no*		*sensitive*		*balanced*		*strict*		*sensitive*		*no*	
	Regional	*no*		*no*		*no*		*no*		*no*		*yes*		*no*	
	Lag [days]	*0*		*0*		*0*		*0*		*0*		*0*		*0*	
Country (Region)	*Benin (Cotonou)*	0.03	0.26	0.0	0.91	-	-	-	-	-	-	-0.01	0.64	-0.03	0.63
	Burkina Faso (Ouagadougou)	0.11	0.12	0.06	0.07	0.5	0.5	-1.0	-	-	-	-	-	0.14	0.01
	Gambia (Banjul)	-0.01	0.73	-0.01	0.82	-	-	-	-	-	-	-0.01	0.73	-0.03	0.61
	Ghana (Kumasi)	0.1	0.0	0.0	0.88	-0.09	0.42	0.19	0.16	0.2	0.2	0.08	0.01	0.02	0.64
	Guinea (Conakry)	-0.03	0.26	-0.01	0.63	-	-	-	-	-	-	-	-	-0.07	0.18
	Ivory Coast (Abidjan)	0.06	0.24	0.01	0.77	0.12	0.8	-0.15	0.77	-0.15	0.77	-	-	0.03	0.52
	Liberia (Monrovia)	-0.02	0.55	-0.01	0.77	-	-	-	-	-	-	-	-	-0.05	0.38

Table B-4: Correlation coefficients of flood-related tweets with the FRI with different features. The features include type (Spearman S), Seasonal (if yes, only values from May-Oct are included), Event (if yes, only forecasts and tweets during an event registered by the algorithm are included), and Lag (shifted tweets time-series between 1-5). The best correlation is highlighted in bold for every country.

		r_s	p_s	r_p	p_p	$r_{s,\,sen}$	$p_{s,sen}$	$r_{s,\,bal}$	$p_{s,bal}$	$r_{s,\,str}$	$p_{s,str}$	$r_{s,\,reg}$	$p_{s,reg}$	$r_{s,\,seas}$	$p_{s,seas}$
Feature	*Type*	*S*		*P*		*S*		*S*		*S*		*S*		*S*	
	Seasonal	*no*		*no*		*no*		*no*		*no*		*no*		*yes*	
	Event	*no*		*no*		*sensitive*		*balanced*		*strict*		*sensitive*		*no*	
	Regional	*no*		*no*		*no*		*no*		*no*		*yes*		*no*	
	Lag [days]	*0*		*0*		*0*		*0*		*0*		*0*		*0*	
Country (Region)	*Mali (Bamako)*	0.06	0.08	0.04	0.19	0.39	0.22	0.87	0.0	0.83	0.04	-	-	0.18	0.0
	Mauritania (Nouackchoott)	0.17	0.0	0.19	0.0	0.45	0.55	0.45	0.55	0.45	0.55	-	-	0.16	0.0
	Niger (Niamey)	0.1	0.0	0.0	0.87	-0.14	0.11	-0.13	0.17	-0.08	0.41	-	-	0.24	0.0
	Nigeria (Lagos)	0.18	0.0	0.02	0.58	-0.01	0.93	0.03	0.79	0.01	0.94	0.04	0.16	0.04	0.43
	Senegal (Dakar)	0.2	0.0	0.26	0.0	0.0	1.0	0.13	0.8	0.13	0.8	-	-	0.24	0.0
	Sierra Leone (Freetown)	-0.03	0.35	-0.01	0.87	-	nan	-	nan	-	nan	-	nan	-0.05	0.31
	Togo (Lome)	0.01	0.39	-0.01	0.68	-	nan	-	nan	-	nan	-0.02	0.51	-0.03	0.54

Table B-5: Correlation coefficients of flood-related tweets and corresponding p values with different features. The features include type (Pearson P, Spearman S), Seasonal (if yes, only values from May-Oct are included), Event (if yes, only forecasts and tweets during a FEDA detected event are included), and Lag (shifted tweets time series between -5-5). Underlined p-values are significant at a 5 % level.

		$r_{s, lag-5}$	$p_{s, lag-5}$	$r_{s, lag-3}$	$p_{s, lag-3}$	r_s	p_s	$r_{s, lag 3}$	$p_{s, lag 3}$	$r_{s, lag 5}$	$p_{s, lag 5}$	$r_{s, rain}$	$p_{s, rain}$	$r_{s,rain,sen}$	$p_{s,rain,sen}$
Feature	*Type*	*S*		*S*		*S*		*S*		*S*		*S*		*S*	
	Seasonal	*no*		*no*		*yes*		*no*		*no*		*no*		*no*	
	Event	*no*		*no*		*no*		*no*		*no*		*no*		*sensitive*	
	Regional	*no*		*no*		*no*		*no*		*no*		*no*		*no*	
	Lag [days]	*0*		*-5*		*-3*		*3*		*5*		*5*		*no*	
Country	Benin	0.07	<u>0.01</u>	0.09	<u>0.0</u>	0.03	0.26	0.11	<u>0.0</u>	0.06	<u>0.05</u>	0.09	0.0	0.13	0.67
	Burkina Faso	-0.02	0.61	-0.02	0.61	0.05	0.12	0.05	0.13	-0.02	0.61	0.1	0.0	0.95	0.05
	Gambia	0.08	<u>0.01</u>	-0.01	0.73	-0.01	0.73	-0.01	0.73	-0.01	0.73	0.05	0.11	-0.54	0.34
	Ghana	0.08	<u>0.0</u>	0.08	<u>0.01</u>	0.09	<u>0.0</u>	0.06	<u>0.03</u>	0.08	<u>0.01</u>	0.16	0.0	-0.11	0.32
	Guinea	0.02	0.41	0.05	0.07	-0.03	0.26	0.08	<u>0.01</u>	0.08	<u>0.01</u>	0.14	0.0	0.5	0.39
	Ivory Coast	0.09	<u>0.0</u>	0.05	0.08	0.04	0.24	0.07	<u>0.02</u>	0.11	<u>0.0</u>	0.09	0.0	0.33	0.46
	Liberia	0.04	0.24	0.03	0.26	-0.02	0.55	-0.02	0.55	-0.02	0.55	0.06	0.05	-	-
	Mali	0.04	0.15	0.05	0.13	0.05	0.08	0.07	<u>0.02</u>	0.09	<u>0.0</u>	0.18	0.0	-0.41	0.18
	Mauritania	0.03	0.36	0.03	0.39	0.12	<u>0.0</u>	0.22	<u>0.0</u>	0.17	<u>0.0</u>	0.07	0.03	0.74	0.26
	Niger	0.1	<u>0.0</u>	0.1	<u>0.0</u>	0.09	<u>0.0</u>	0.11	<u>0.0</u>	0.1	<u>0.0</u>	0.2	0.0	0.13	0.15
	Nigeria	0.15	<u>0.0</u>	0.14	<u>0.0</u>	0.19	<u>0.0</u>	0.18	<u>0.0</u>	0.15	<u>0.0</u>	0.4	0.0	-0.01	0.86
	Senegal	0.02	0.48	0.06	<u>0.04</u>	0.15	<u>0.0</u>	0.11	<u>0.0</u>	0.07	<u>0.03</u>	0.15	0.0	0.24	0.61
	Sierra Leone	-0.03	0.35	0.01	0.83	-0.03	0.35	-0.03	0.35	-0.03	0.35	0.17	0.0	0.28	0.06
	Togo	0.01	0.75	0.01	0.74	-0.03	0.39	0.01	0.74	0.05	0.12	0.03	0.31	-	-

Table B-6: p-values of the Welch's test p_{pre}, p, and p_{post}. p_{pre} includes FRI 5 days before and during a detected event into the event group, p only includes FRI during a detected event, and p_{post} includes the FRI during a detected event and 5 days after. $Event_{pre}$, Event, $Event_{post}$ and No $event_{pre}$, No event, No $event_{post}$ signify the event and no event group respectively. s and N show the sample size and normality of the group, derived from the Shapiro-Wilk test. A value below 0.05 indicates a rejection of the null hypothesis that the sample is normally distributed.

Country	$\mathbf{p_{pre}}$	$\mathbf{Event_{pre}}$		**No** $\mathbf{event_{pre}}$		**p**	**Event**		**No event**		$\mathbf{p_{post}}$	$\mathbf{Event_{post}}$		**No** $\mathbf{event_{post}}$	
		s	*N*	*s*	*N*		*s*	*N*	*s*	*N*		*s*	*N*	*s*	*N*
Benin	0.96	49	0.0	1047	0.0	0.0	14	1.0	1082	0.0	0.16	49	0.0	1047	0.0
Burkina Faso	0.23	14	0.0	1082	0.0	0.22	4	0.27	1092	0.0	0.23	14	0.0	1082	0.0
Gambia	0.07	15	1.0	1081	0.0	0.07	5	1.0	1091	0.0	0.07	15	1.0	1081	0.0
Ghana	0.0	202	0.0	906	0.0	0.04	77	0.0	1019	0.0	0.01	202	0.0	906	0.0
Guinea	0.0	15	1.0	1083	0.0	0.0	5	1.0	1091	0.0	0.0	15	1.0	1083	0.0
Ivory Coast	0.32	22	0.0	1074	0.0	0.42	7	0.02	1089	0.0	0.13	22	0.0	1074	0.0
Liberia	0.0	6	1.0	1090	0.0	-	1	-	1095	0.0	0.0	6	1.0	1090	0.0
Mali	0.05	32	0.0	1064	0.0	0.47	12	0.0	1084	0.0	0.04	32	0.0	1064	0.0
Mauritania	0.23	9	0.01	1087	0.0	0.2	4	0.02	1092	0.0	0.02	9	0.0	1087	0.0
Niger	0.0	267	0.0	841	0.0	0.21	132	0.0	964	0.0	0.0	267	0.0	841	0.0
Nigeria	0.15	232	0.0	867	0.0	0.22	142	0.0	954	0.0	0.0	232	0.0	867	0.0
Senegal	0.27	22	0.0	1074	0.0	0.38	7	0.0	1089	0.0	0.2	22	0.0	1074	0.0
Sierra Leone	0.02	62	1.0	1034	0.0	0.02	47	1.0	1049	0.0	0.02	62	1.0	1034	0.0
Togo	-	0	-	1096	0.0	-	0	-	1096	0.0	-	0	-	1096	0.0

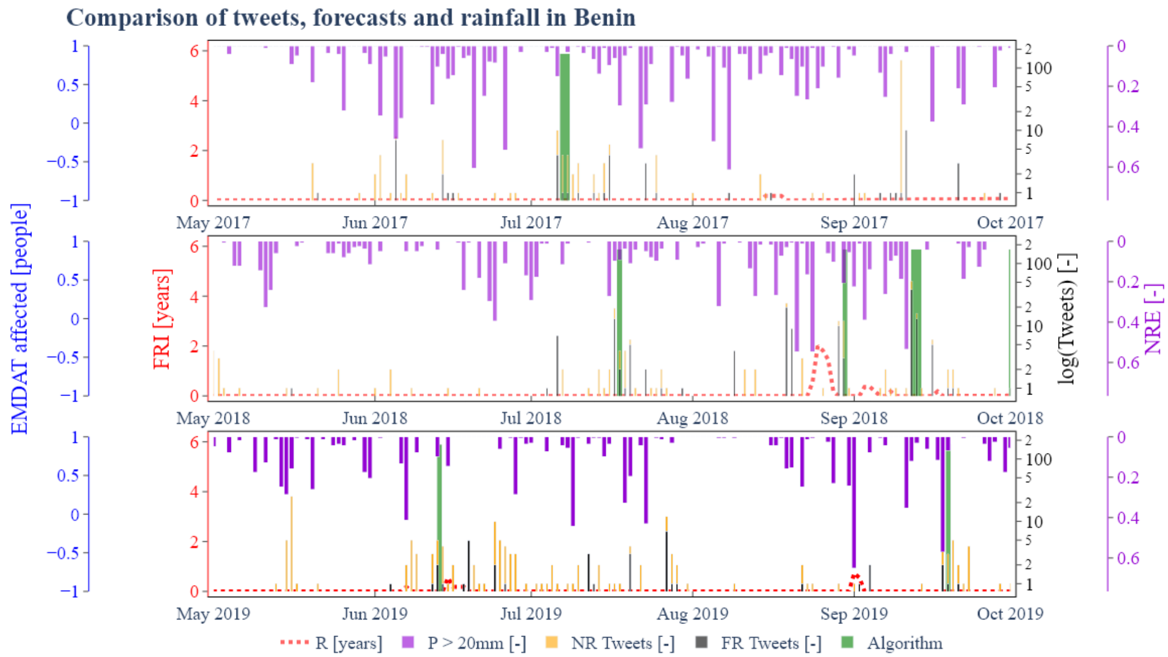


Figure B-1: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Benin. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

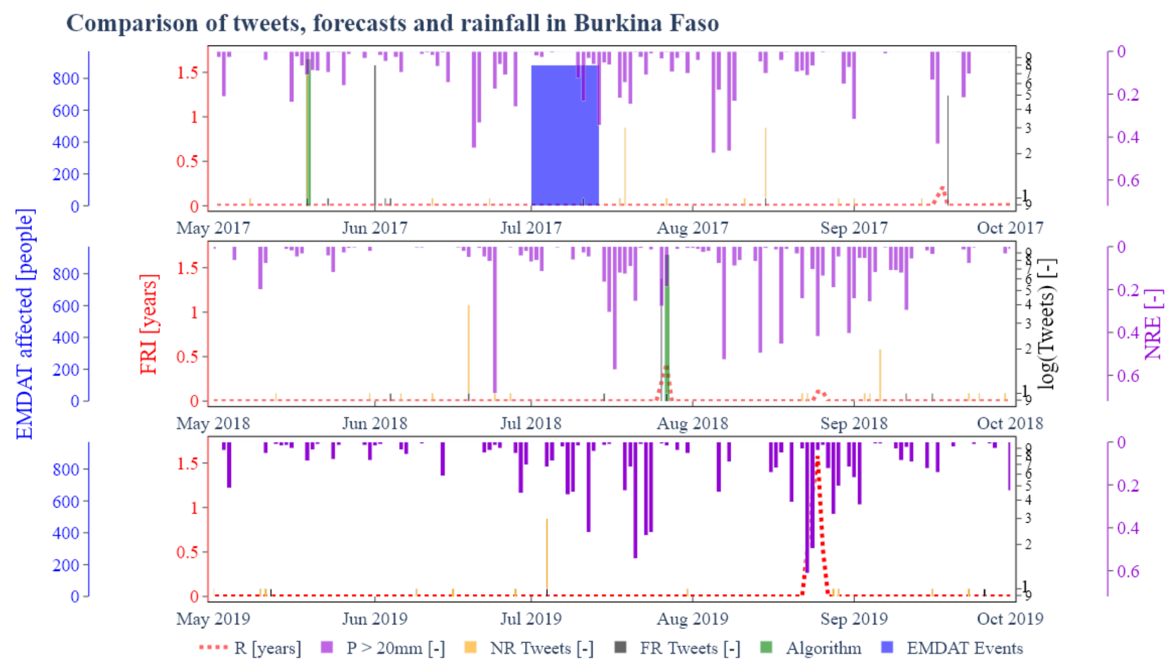


Figure B-2: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Burkina Faso. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

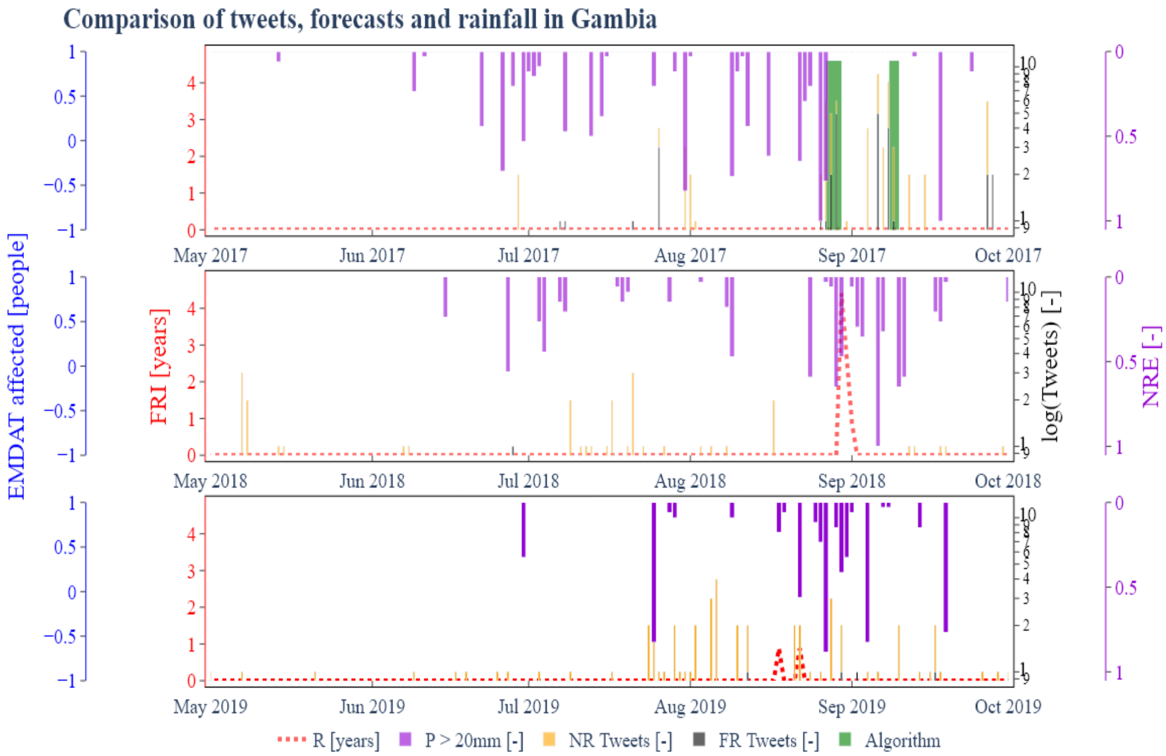


Figure B-3: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in the Gambia. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

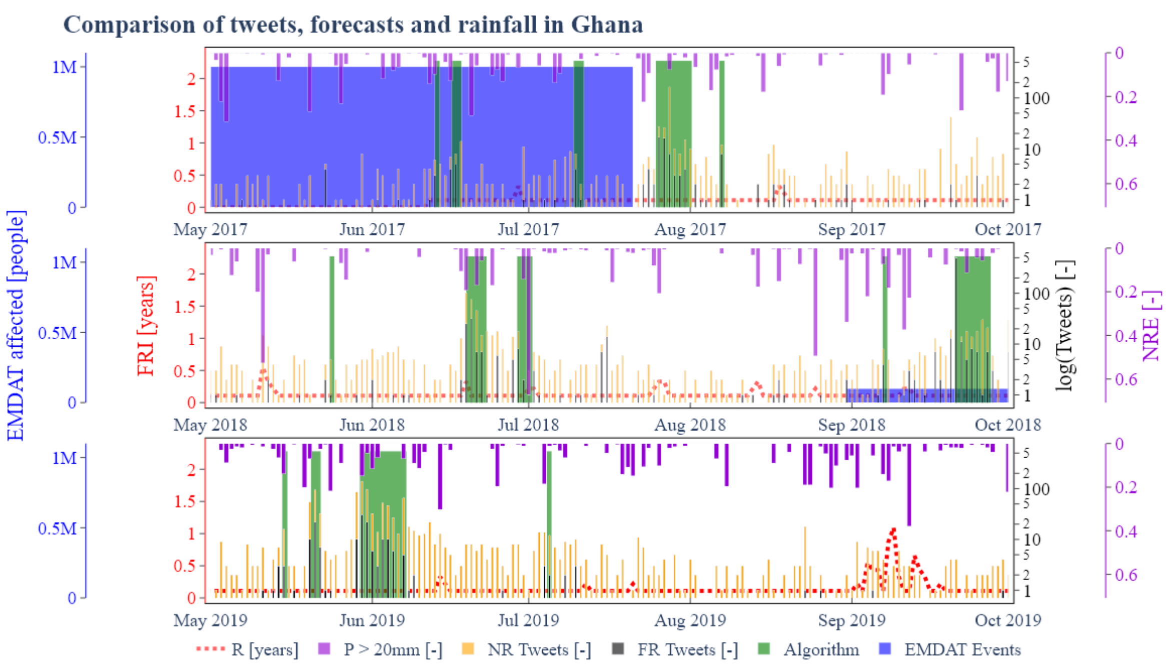


Figure B-4:: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Ghana. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

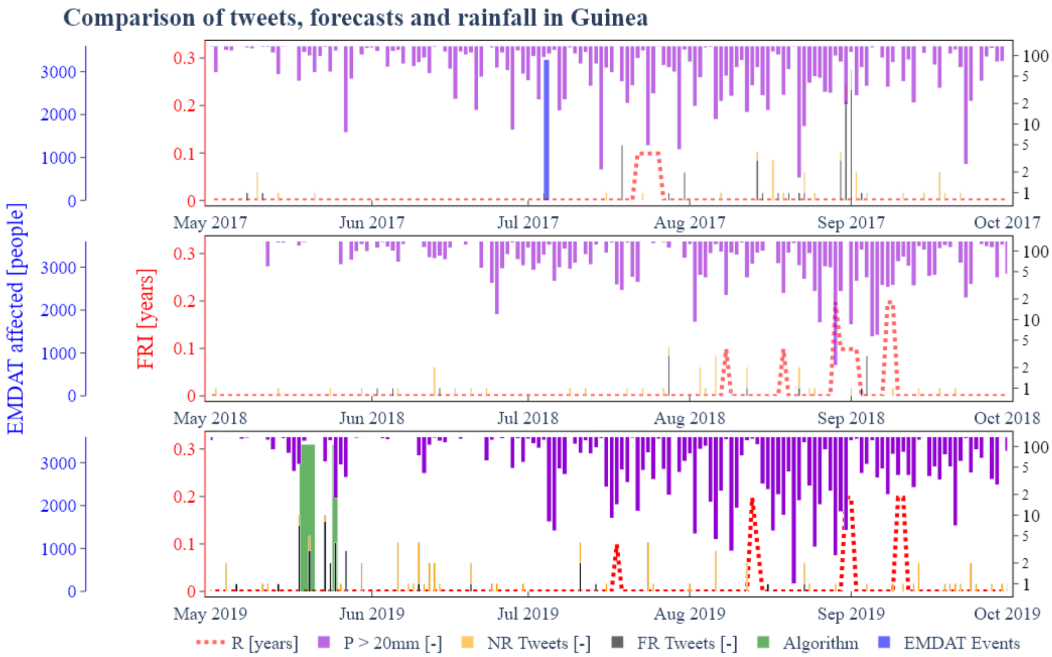


Figure B-5: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Guinea. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

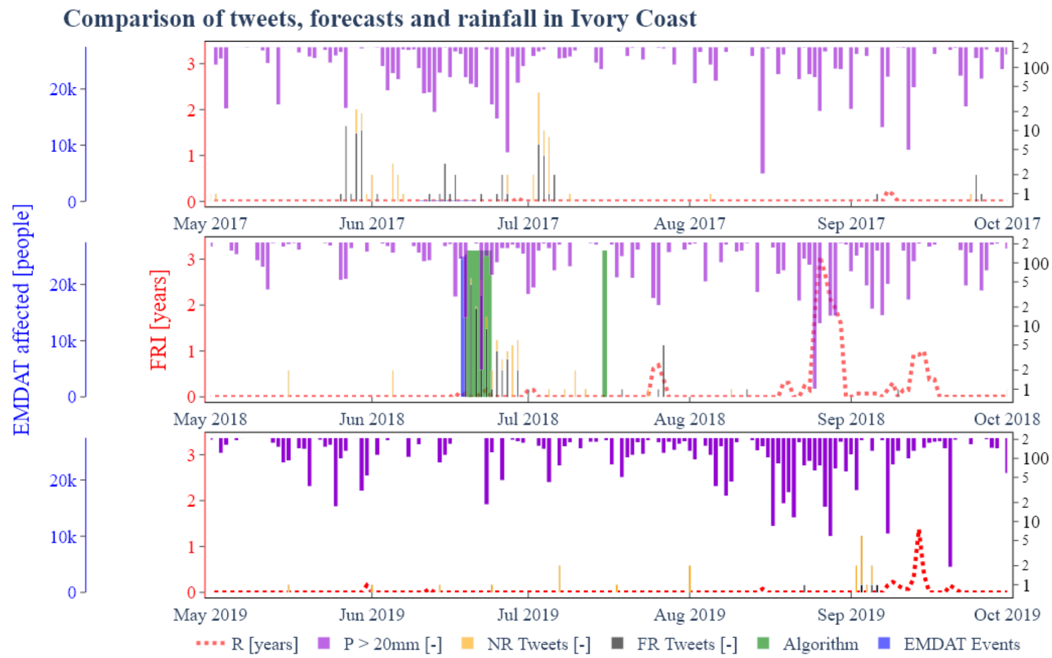


Figure B-6: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Ivory Coast. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

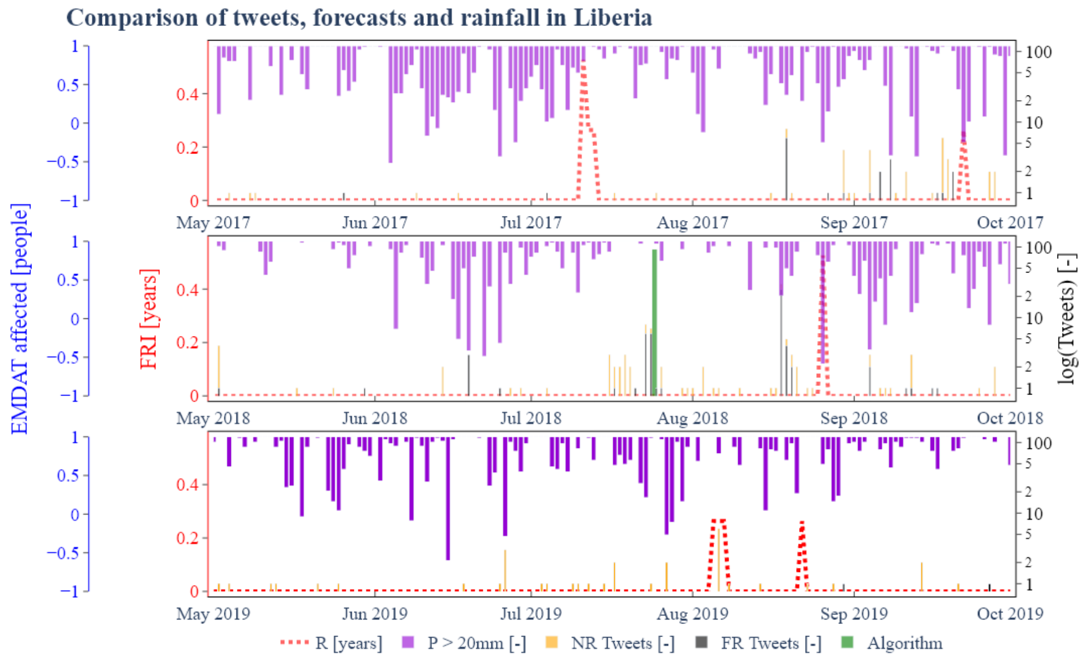


Figure B-7: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Liberia. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

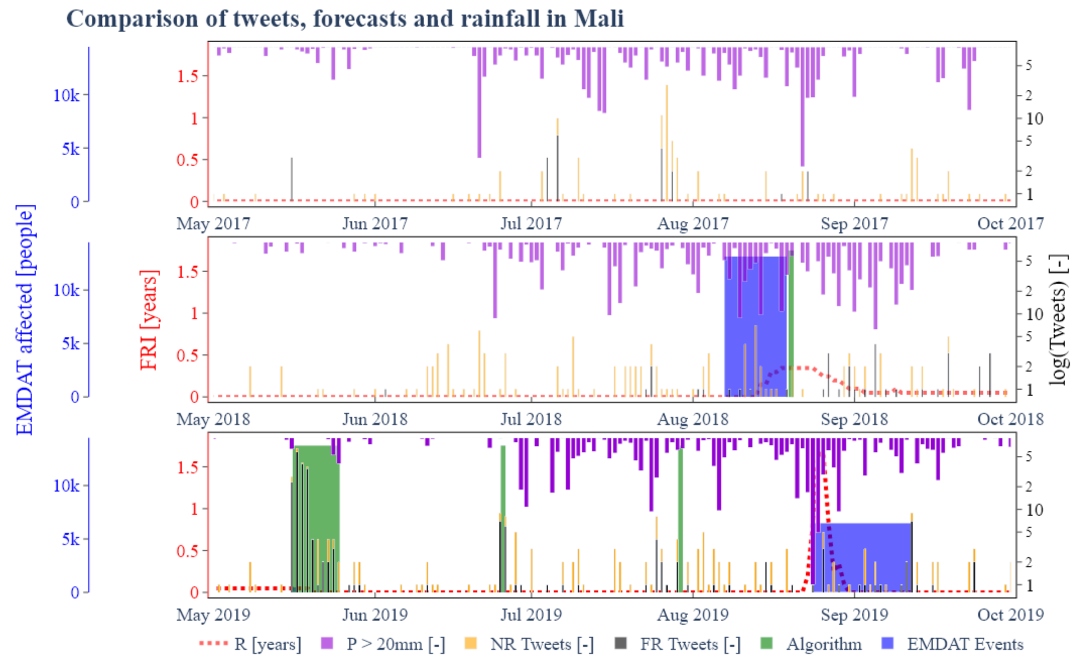


Figure B-8: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Mali. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

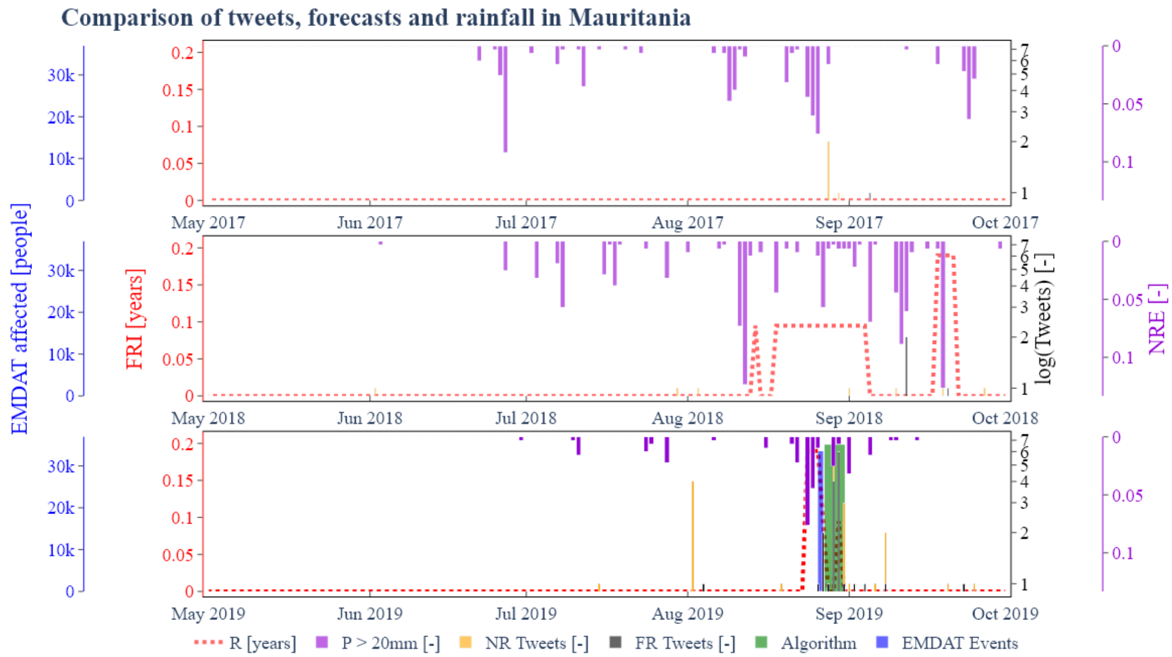


Figure B-9: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Mauritania. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

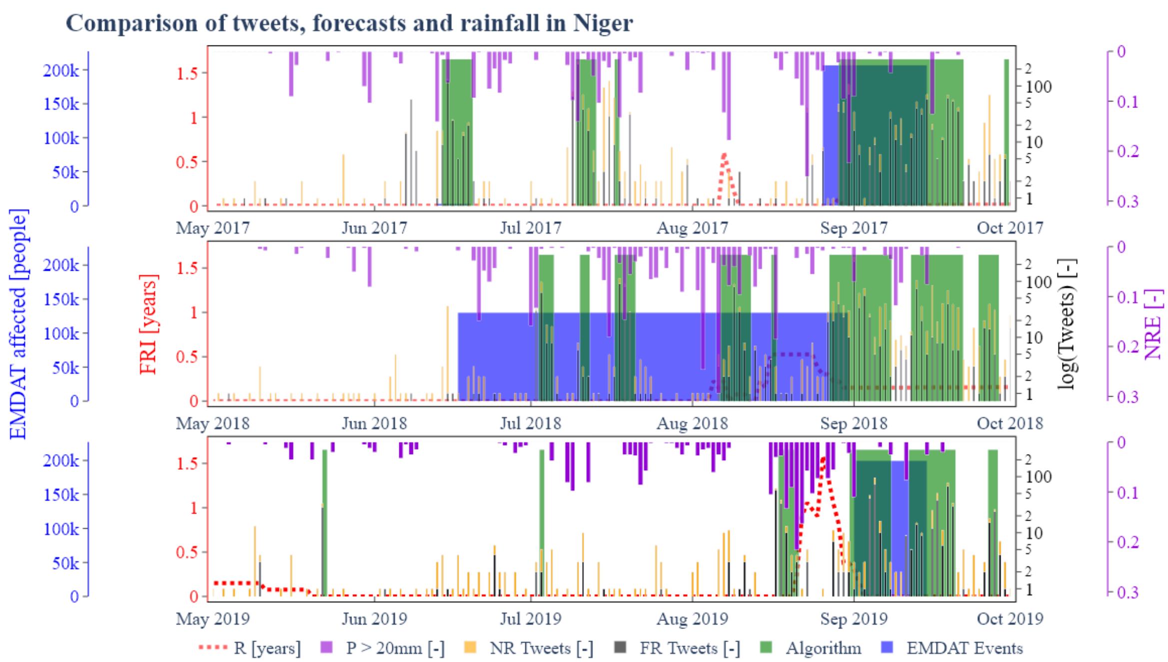


Figure B-10: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Niger. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

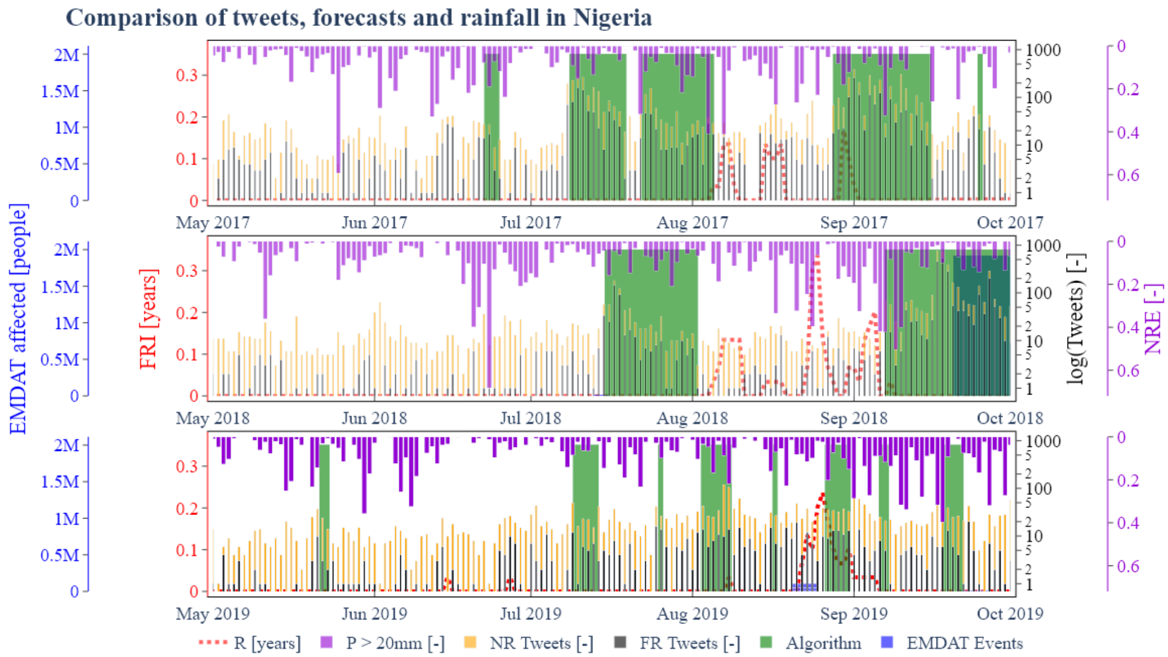


Figure B-11: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Nigeria. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

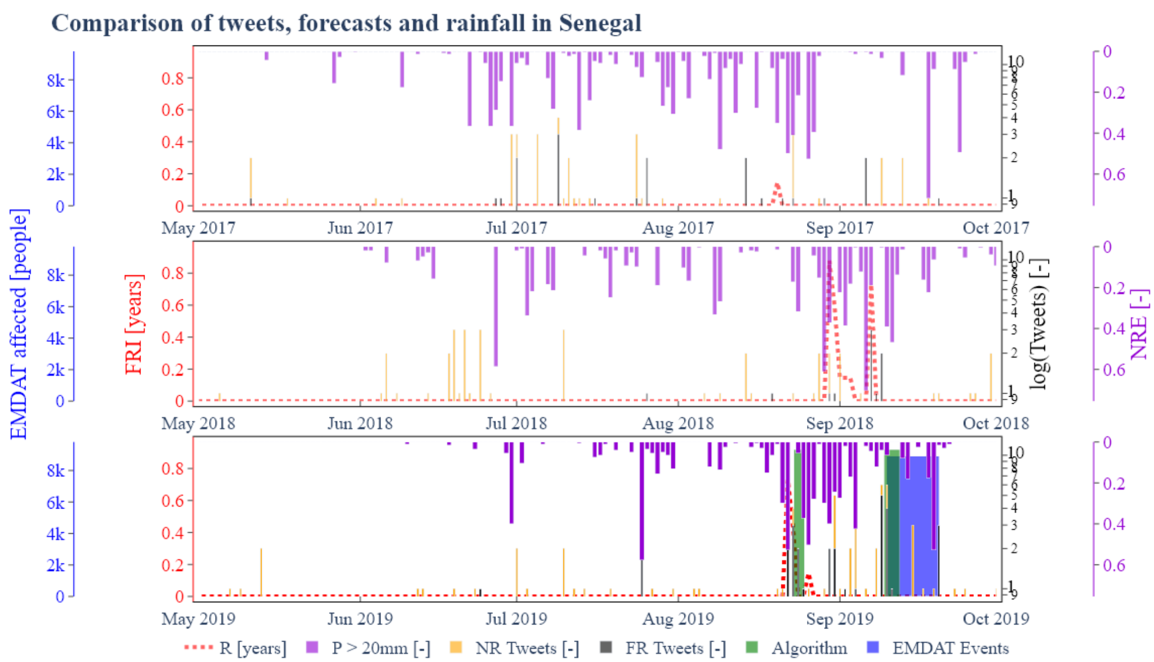


Figure B-12: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Senegal. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

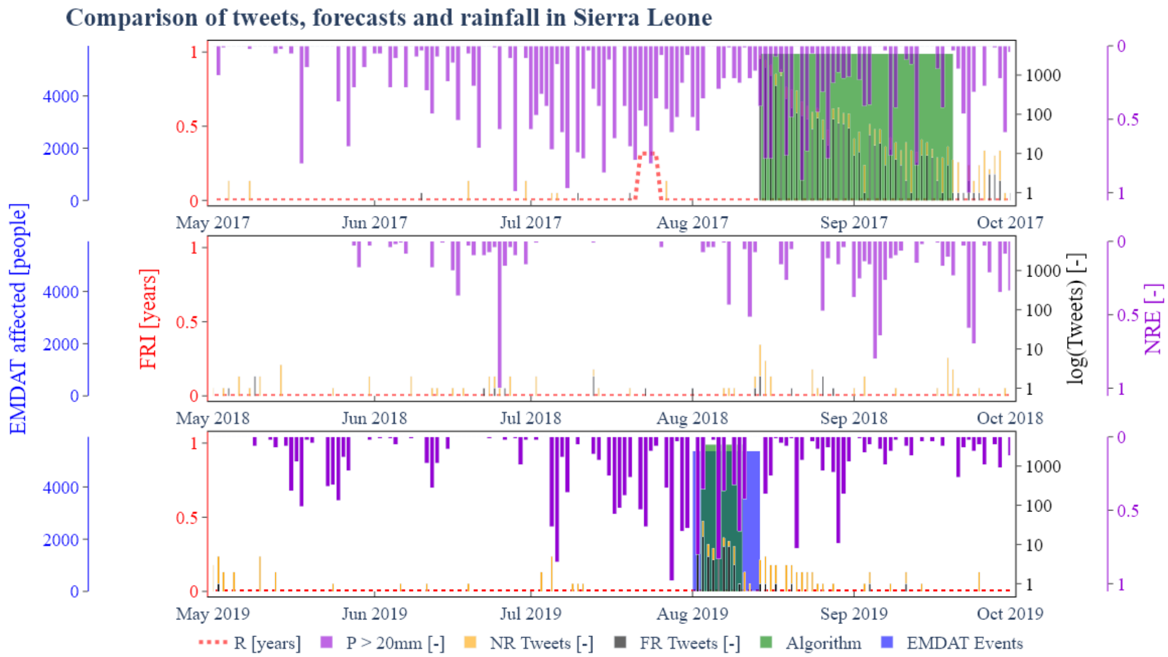


Figure B-13: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Sierra Leone. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.

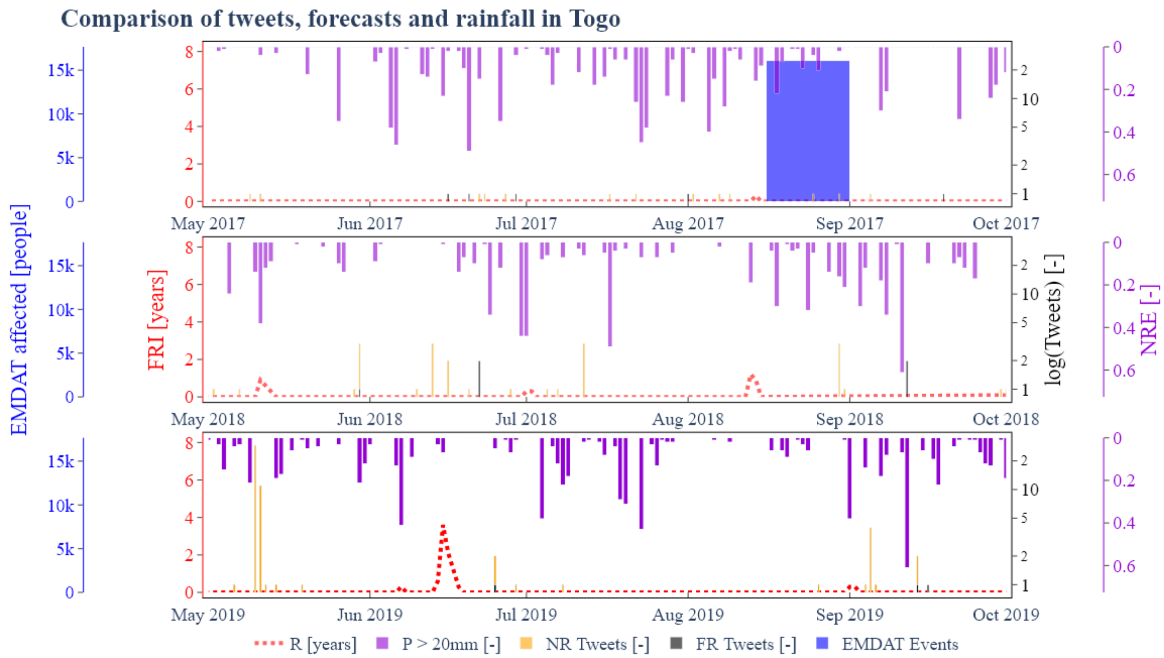


Figure B-14: Comparison of tweets, forecasts, rainfall exceedances, and EMDAT events during the rainy season for the years 2017-2019 in Togo. Note that the y-axis of the Tweets is logarithmic. P: Rainfall, R: Return period, NR: Tweets not related to flooding, FR: Tweets related to flooding.