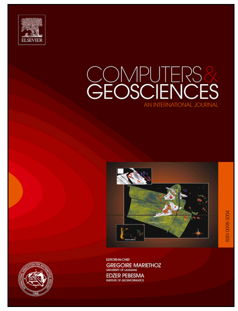
Accepted Manuscript



Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring

Camilo Restrepo-Estrada, Sidgley Camargo de Andrade, Narumi Abe, Maria Clara Fava, Eduardo Mario Mendiondo, João Porto de Albuquerque

PII: S0098-3004(17)30611-8

DOI: 10.1016/j.cageo.2017.10.010

Reference: CAGEO 4038

To appear in: *Computers and Geosciences*

Received Date: 1 June 2017

Revised Date: 18 September 2017

Accepted Date: 22 October 2017

Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring

Camilo Restrepo-Estrada[a,*], Sidgley Camargo de Andrade[b,c], Narumi Abe[a], Maria Clara Fava[a], Eduardo Mario Mendiondo[a], João Porto de Albuquerque[d,c]

[a] *Sao Carlos School of Engineering, University of Sao Paulo, Sao Carlos, Brazil*
[b] *Federal University of Technology – Parana, Toledo, Brazil*
[c] *Institute of Mathematical and Computing Sciences, University of Sao Paulo, Sao Carlos, Brazil*
[d] *Centre for Interdisciplinary Methodologies, University of Warwick, Coventry, UK*

Abstract

Floods are one of the most devastating types of worldwide disasters in terms of human, economic, and social losses. If authoritative data is scarce, or unavailable for some periods, other sources of information are required to improve streamflow estimation and early flood warnings. Georeferenced social media messages are increasingly being regarded as an alternative source of information for coping with flood risks. However, existing studies have mostly concentrated on the links between geo-social media activity and flooded areas. Thus, there is still a gap in research with regard to the use of social media as a proxy for rainfall-runoff estimations and flood forecasting. To address this, we propose using a transformation function that creates a proxy variable for rainfall by analysing geo-social media messages and rainfall measurements from authoritative sources, which are later incorporated within a hydrological model for streamflow estimation. We found that the combined use of official rainfall values with the social media proxy variable as input for the Probability Distributed Model (PDM), improved streamflow simulations for flood monitoring. The com-

*Corresponding author
Email addresses: camilo.restrepo@udea.edu.co (Camilo Restrepo-Estrada), sidgleyandrade@utfpr.edu.br (Sidgley Camargo de Andrade), narumi@usp.br (Narumi Abe), mclarafava@usp.br (Maria Clara Fava), emm@cemaden.gov.br (Eduardo Mario Mendiondo), J.Porto@warwick.ac.uk (João Porto de Albuquerque)

bination of authoritative sources and transformed geo-social media data during flood events achieved a 71% degree of accuracy and a 29% underestimation rate in a comparison made with real streamflow measurements. This is a significant improvement on the respective values of 39% and 58%, achieved when only authoritative data were used for the modelling. This result is clear evidence of the potential use of derived geo-social media data as a proxy for environmental variables for improving flood early-warning systems.

Keywords: social media, hydrological modelling, estreamflow estimation, flood monitoring

---

1. Introduction

Floods have been gradually increasing throughout the world, and causing serious levels of human, economic and social losses. For this reason, forecasting and monitoring have attracted a great deal of attention as a means of improving early warning systems (Patankar and Patwardhan, 2016; Crochemore et al., 2016). Flood forecasting and monitoring are being increasingly characterised as a problem of "big data", since there are different data sources that can be used to support decision making, such as satellites, radar systems, rainfall gauges and hydrological networks (Horita et al., 2017). However, in situations of crisis management, the apparent overabundance of data is often accompanied by a simultaneous "information dearth": a lack of information may arise because sensors are not available for certain regions or the number of available sensors is not enough to cover the territory with a suitable resolution. In hydrology, this problem is attributed to the so-called "ungauged" or "poorly gauged" catchments (Sivapalan et al., 2003). In response, big data sources are emerging that provide important information and can supplement traditional sensors. These sources include data provided by people directly linked to affected areas or flood-prone areas, which can be used in many natural disaster risk scenarios and assist in water resources management (Fraternali et al., 2012).

Over the last few years, there has been a growing interest in using georeferenced social media to support urban resilience to flooding. The advance of mobile telecommunications and the widespread use of smartphones and tablets allow people to act as human sensors, and generate volunteered geographic information (Goodchild, 2007). Moreover, they have been increasingly recognised and used as an important resource to support disaster management (Goodchild and Glennon, 2010; Horita et al., 2015). This spatial information is produced by ordinary people through different collaborative activities, such as exchanging information through geotagged social media messages (de Albuquerque et al., 2017).

One of the advantages of using social media for monitoring flood events is the extensive spatial coverage of the measurements. These make it possible to obtain useful information at different points of river catchment areas and cities where the local inhabitants are able to supplement the static sensors of the hydrometeorological networks. However, even today there are still multiple challenges that have to be faced; these, include finding the best way to extract relevant information from social media and the difficulty of integrating this information with data from other sources to achieve greater reliability. Furthermore, an additional challenge is to ensure that these new information sources can be used to assist the hydrological models to support decision-making with regard to the early warning system (Mazzoleni et al., 2017; Horita et al., 2015).

Most of the previous work in this area has concentrated on using social media data either for flood mapping or exploring spatiotemporal patterns (Smith et al., 2015; Weng and Lee, 2011; Tkachenko et al., 2017). In our previous work, we found there were close spatiotemporal links between social media activity and flood-related events (de Albuquerque et al., 2015), as well as social media activity and rainfall (de Andrade et al., 2017). However, to the best of our knowledge, so far no scientific work has used social media data quantitatively to estimate hydrological models for flood monitoring. This paper differs from our previous studies (de Andrade et al., 2017) by going one step further than simply establishing a correlation between social media activity and rainfall: it now examines the frequency of rainfall-related messages to define a data series of non-authoritative rainfall. This data series can then be used as input to enable a hydrological model to predict streamflow.

Our approach is based on the hypothesis that it is possible to use indicators derived from social media activity for flood monitoring and/or forecasting, in conjunction with data from hydrometeorological sensors in streamflow modelling, to make further improvements to early warning systems. In this paper, we seek to transform Twitter data into a proxy variable for precipitation. Transforming this data requires a function that converts Twitter messages into rainfall values. When setting up the transformation function, it is assumed that there is a direct relationship between the intensity of rainfall and the rainfall-related activity of geo-social media in a given geographical area. We can thus use the rainfall proxy variable in a rainfall-runoff model to estimate the streamflow.

This paper is structured as follows. Section 2 introduces a discussion of related works. Section 3 describe the case study and data. Section 4 describes the methodology. Section 5 and 6 examine the main results that have been achieved and include a discussion of the work. Finally, Section 7 summarizes the general conclusions and makes recommendations for future work.

2. Related work

Modelling urban catchment behaviour requires high-resolution rainfall and detailed physical characteristics owing to the fast hydrologic response of the catchment (Hapuarachchi et al., 2011; Ochoa-Rodriguez et al., 2015; Wang et al., 2015). Rainfall data is the main input in rainfall-driven hydrological models for flood modelling and forecasting. Several approaches have been tested for different situations to highlight the use of remote sensing for rainfall-driven flood forecasting (Skinner et al., 2015; Li et al., 2016) as an alternative to the traditional use of in-situ measurements. Boni et al. (2016) implemented a near real-time flood-mapping algorithm using Synthetic Aperture Radar (SAR) together with a satellite, coupled to a hydraulic model. Tiesi et al. (2016) used surface network data, radio-sounding profiles, radar and satellite (SEVIRI/MSG) data for quantitative precipitation forecasting and found they had a positive effect on the intensity and distribution of the simulated rainfall. Studies such as Wang et al. (2015) and Chen et al. (2016) showed that although radar-based precipitation measurements have the advantage of being able to reproduce the spatial structure of rainfall fields and their variation in time with regard to ground-based measurements, they still cannot achieve the accuracy and resolution required for urban hydrology.

However, it is not always possible to have information from rain gauges, or radar and meteorological satellites. Thus, it is necessary to explore other alternatives for forecasting and monitoring that can mitigate the effects of flooding. In response to this need, a new field has emerged to explore how social data can be combined with remote sensing information to improve flood forecasting in ungauged or poorly gauged catchments (Sivapalan et al., 2003).

The use of geo-social media in disaster management has been explored in the literature for various types of hazards such as earthquakes (Crooks et al., 2013; Sakaki et al., 2010), forest fires (Crooks et al., 2013; Sakaki et al., 2010), hurricanes (Huang and Xiao, 2015), tsunamis (Mersham, 2010), agricultural droughts (Enenkel et al., 2015), and floods (Smith et al., 2015; Weng and Lee, 2011; Tkachenko et al., 2017). In the particular area of flood management, scientific work has focused on using social media data for two requirements - flood mapping and exploring spatiotemporal patterns.

Tweets have been quantitatively used in both forecasting and mapping. Schnebele et al. (2014) concluded that a fusion of multiple non-authoritative data sources helps to fill in gaps in the spatial and temporal coverage of authoritative data. They used aerial photos, Youtube videos, Twitter and Google photos to create maps of the damage caused by Hurricane Sandy. Brouwer et al. (2017) harvested 8000 flood-related tweets from York in England and used this information to create a probabilistic flood extent map. Patel et al. (2017) used tweets to produce population maps. Rathore et al. (2017) devised a system that uses geo-social media to harvest, process, and analyse a large amount of data at high-speed from Twitter and make decisions in real time. Li et al.

(2017) collected tweets during a period of 18 days in South Carolina, USA, which involved filtering by means of flood-related keywords, and found 4,268 flood-related tweets. Based on this information, and using temporal granularity on a daily basis, they found a close correlation between stream gauge levels and the absolute frequency of flood-related tweets. In these studies, tweets were a weighting factor for creating inundation maps.

There are other studies that are confined to demonstrating the relationship between flood-related messages and flood events. Weng and Lee (2011) collected tweets for a month in June 2010 to detect events in Singapore, and based on this information, they built the signal events that were reported on Twitter automatically, by means of a wavelet transform. However, in this period, they only detected a single flood event. Smith et al. (2015) used tweets to improve and extrapolate data from hydraulic modelling to assess flooding. This was carried out through two events that occurred in the city of Newcastle. Tkachenko et al. (2017) also used flood-related geo-tagged messages from Flickr to detect floods in England.

Going one step further towards achieving a quantitative integration of social media activities into flood forecasting models, is of value as a supplementary resource for monitoring catchments, given the fact that sometimes the rain gauges that are usually used for this activity, are not available or fail for various reasons, such as a lack of maintenance.

3. Case study and Data

This section describes the data that will be used, both authoritative and social media data, and conducts an exploratory analysis of spatial data.

3.1. The Aricanduva Catchment

The Aricanduva catchment (Fig. 1) is located in the city of Sao Paulo, Brazil, a metropolitan region with more than 20 million inhabitants, with the largest population density in Brazil. Aricanduva is a tributary of the Tiete River, the main river of the city, and has a total drainage area of 100 km^2. In this study we selected a sub-catchment of 88 km^2, where the Sao Paulo Flood Warning System (SAISP)[1] - the organization responsible for measuring water levels – has three water level sensors, of which one was selected because is close to a risk-prone area subject to frequent flash flooding (see Fig. 2). Water level sensor measurements are provided every 10 min by SAISP. The precipitation data is also provided every 10 min by the National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN)[2].

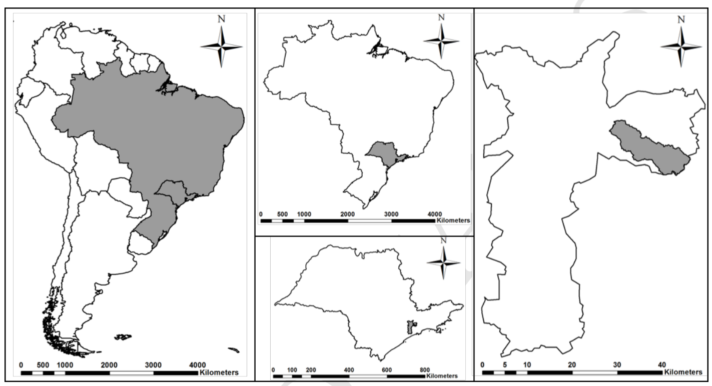


Figure 1: Aricanduva watershed, Sao Paulo Metropolitan Region, selected for this study.

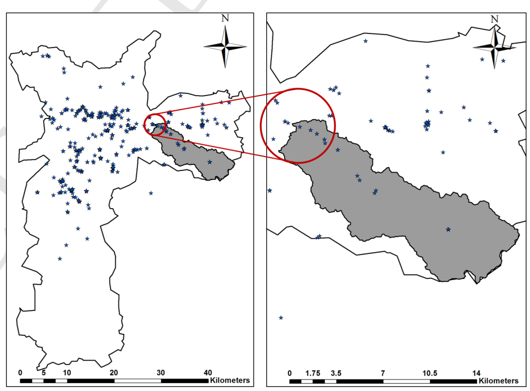


Figure 2: SAISP reported flood points.

[1] `https://www.saisp.br/estaticos/sitenovo/home.xmlt`
[2] `http://www.cemaden.gov.br/`

3.2. Social media data

The social media data used in this study were gathered from the Twitter platform using the public streaming Application Programming Interface (API) to obtain georeferenced tweets within a bounding box that encompasses the city of Sao Paulo. The total number of tweets collected was 15,883,710. The georeferenced tweets (1,631,329) were then filtered by means of keywords (21,804). From the 1st to 30th January 2016 and from 8th November 2016, to 28th February 2016, we found 6,651 geotagged tweets related to floods within the city of Sao Paulo. As in the case of our previous study (de Andrade et al., 2017), we filtered the messages to find words related to rain (chuva in Portuguese), intense rainfall and rainbows, but excluded common unrelated expressions (Fig. 3). Some examples for related tweets can be found in Table 1. Figure 4 shows the spatial distribution of the rainfall-related tweets in the city of Sao Paulo during this period.

(a) Related words (b) Unrelated words

Figure 3: Frequently-related and unrelated words. All the keywords are in unicode standard.

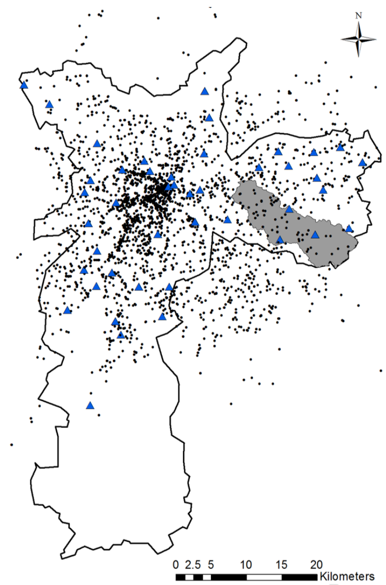


Figure 4: City of Sao Paulo during the analysed period, with related tweets as black points, rainfall gauges as blue triangles and the Aricanduva catchment shaded in gray.

The geo-located tweets containing the keywords were collected and assigned to temporal bins of 10 minutes in a variable called “absolute frequency of real-time messages” f_{kw}. Other variables obtained from the related tweets are the cumulative frequencies of every Δt min.

Table 1: Some related tweet messages collected in this study.

Date/Time	**Portuguese version**	**Translated version**
2016-11-09 20:34:23	“EM MINHA DEFESA......que fique claro que vim por causa da chuva impraticável e só tomando uma coca (@Hooters) https://t.co/KEFYXy8YM4”	“IN MY DEFENSE ...... that it is clear that I came because of the impractical rain and only drinking a coke (@Hooters) https://t.co/KEFYXy8YM4”
2016-12-03 21:43:25	“Início da noite sede sábado, com chuva... que lindo presente de Deus! (Sem filtros) https://t.co/Js7kmDrOZY”	“Early Saturday night, with rain ... what a beautiful gift from God! (No filters) https://t.co/Js7kmDrOZY”
2016-12-11 18:35:23	“Muita chuva ..... já vi que vou ganhar chá de cadeira ...... partiu casa carioca ..... https://t.co/E1q4rM5ivE”	“A lot of rain ..... I’ve already seen that I’m going to get a long wait ...... I left carioca house ..... https://t.co/E1q4rM5ivE”
2017-02-27 0:38:15	“Chuva, chuva, chuva e mais chuva ... https://t.co/wH2GOnqz80”	“Rain, rain, rain and more rain ... https://t.co/wH2GOnqz80”

3.3. Authoritative data

Rainfall data were collected from CEMADEN with the aid of an API Application. The data is updated at intervals of 10 min when the cumulative volume in the period is higher than 0.2 mm. However, if no rainfall is recorded, the data are available every hour. Thus, since our modelling is aimed at providing a tool to predict floods, the rainfall-runoff calibration is carried out for some previous rainfall events, when there is a total precipitation greater than 10 mm. This meant that 30 rainfall events greater than 10 mm were chosen for model calibration (from 2015-04-06 to 2015-12-29 and 2016-02-05 to 2016-10-14) and another 15 were chosen for validation (from 2016-01-01 to 2016-01-30 and 2016-11-09 to 2017-02-27). The quality and consistency of the available rain gauge information were assessed by comparing it with the information gathered by the University of Sao Paulo (USP), Sao Paulo, and its observatory, which calculates the monthly rainfall rate[3].This information allowed us to validate the accumulated magnitudes of the rainfall stations. As a result, we decided to use three sensors that showed values that were consistent with both sources.

Figure 5 shows an example of the difficulties that a situation room, (such as the one in CEMADEN), may face when there are problems with authoritative data. The image was taken from the official interactive map on February 2nd 2017[4]. It can be seen that on this date, there were some sensors that did not report data at all (black points), as well as apparent inconsistencies in the measurements made by some sensors, concerning the amount of rainfall that fell on the city of Sao Paulo. These situations provide a further reason for using alternative information sources to assist flood monitoring and early warning systems.

[3] http://www.estacao.iag.usp.br/

[4] http://www.cemaden.gov.br/mapainterativo/

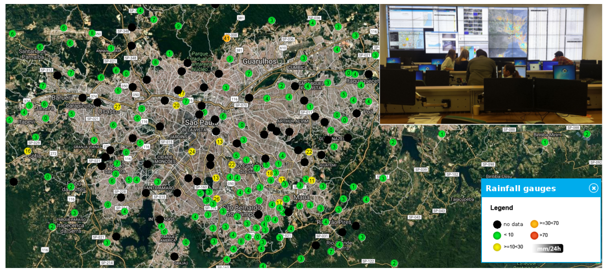


Figure 5: Problems with authoritative data, February 2^{nd}, 2017.

3.4. Exploratory data analysis

An initial exploratory data analysis is displayed in Fig. 6, which summarizes the absolute frequency of two time-series. One is carried out for the key words of Twitter phrases related to rainfall processes and collected at the same time. The other one corresponds to the rainfall depths measured by the authoritative sensors. Evidence obtained from plotting the two time series, reveal a time-dependent significant relationship between the frequency of the tweets and rainfall depths.

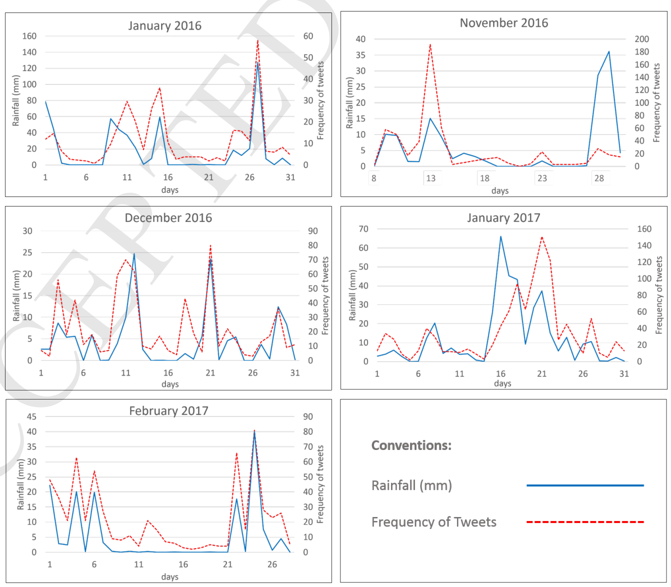


Figure 6: Time series of rainfall depths (left) with frequency of tweets (right) for the period of study January 2016 and from November 8^{th}, 2016 to February 28^{th}, 2017.

As shown in Fig. 6, in some events the two series did not follow the same behaviour or have the same relative magnitude. For instance, on November 12^{th}, 2016, there was a peak in the frequency of tweets, which coincided with a live performance of Guns and Roses, an American hard rock band. Those who attended the concert filled Twitter with images and messages in Portuguese and English referring to "November Rain", a well-known song played by this band. This reaction seems to have been heightened by the fact that it was sung while it was raining in the city. One example of how false positives can occur in detections is illustrated by the following tweet: "luizh.ap: November Rain com direito a chuva e balões vermelhos #GunsNRoses #gunsnrosesreunion #Axl #Slash #Duff #GNR" which can be translated as "November Rain with the right to rain and red balloons!!". These constraints call for a methodology for refining geotagged data related to rainfall, as explained in the following section.

4. Methodology

Figure 7 displays the methodological structure adopted to transform data from social media into a hydrometeorological proxy variable. The methodology is divided into four stages: (a) hydrological data (calibration and rainfall-streamflow modelling) (b) social media data (fitting the transformation function proxy) (c) social media data (transformation of social media signal into hydrometeorological data) (d) comparison with real data. In each stage, a series of activities is carried out. Each of these processes are in turn explained in the next sections.

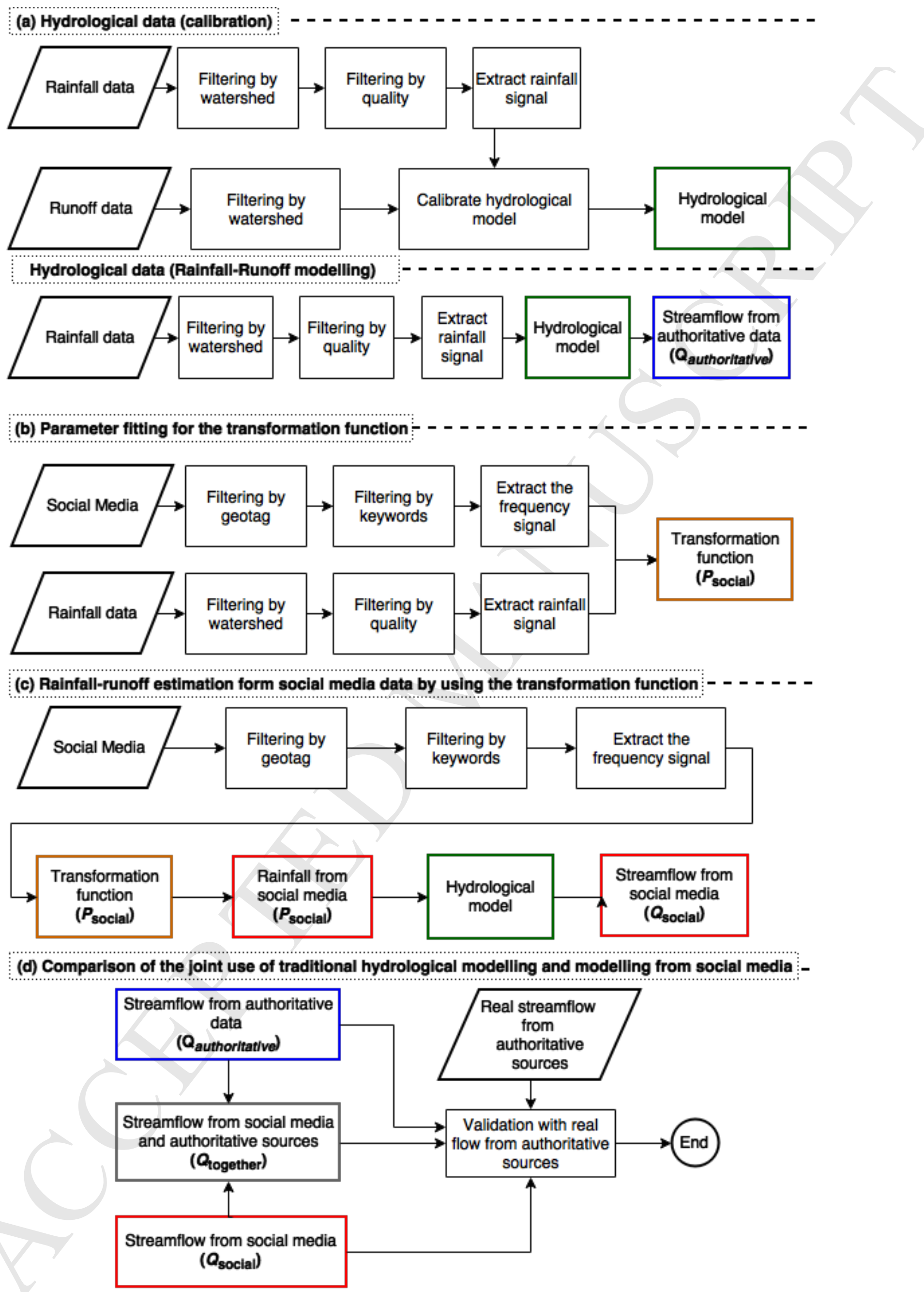


Figure 7: Methodological structure to transform authoritative and social media information to improve flood monitoring.

4.1. Hydrologic data

The first methodological procedure carried out was the calibration of the hydrological model that was used to obtain a transformation of authoritative and social media rainfall values into streamflow. This is a classic procedure in hydrology where some hydrometeorological variables such as rainfall and streamflow are used to calibrate the model (Muleta, 2011). In view of the fact that the methodology is designed to be used in ungauged and poorly gauged catchments or when there are sensors subject to failures, simple modelling seems to be more appropriate (Sivapalan et al., 2003).

The Probability Distributed Model (PDM) and similar models derived from it, are conceptual rainfall-runoff models that are widely used in research and hydrological applications (Alvarez-Garreton et al., 2014), such as parameter prediction updating, flood forecasting, and the regionalization of parameters using the Kalman filter, (Lamb, 1999; Moradkhani et al., 2005; Kay et al., 2009). PDM transforms rainfall and the estimation of the evapotranspiration time series of a catchment into streamflow at the outlet of the catchment. Moore (2007) provides a detailed description of the process modelled, parameters and model formulation. PDM has been chosen in preference to distributed and physically-based hydrological models because it requires a reasonable number of hydrometeorological variables (i.e. rainfall, potential evapotranspiration and streamflow), and is a spatially-lumped, parsimonious and user-friendly model, which reduces the modelling time. In contrast, distributed and physically-based hydrological models involve high computational requirements for simulating spatio-temporal processes in multiple control sections through non-linear equations.

In this paper, the PDM has been calibrated and validated with time-steps of 10 min, that take account of the available 10-min rainfall data and the rapid response time, (ca. 30min) of the studied catchment. Based on ArcGIS and ASTER GDEM, the catchment area was estimated to be 88 km^2. An optimization protocol was developed to calibrate the parameters of the PDM using Python 3.x language and DEAP (Distributed Evolutionary Algorithms in Python) Library. The PDM parameters were calibrated using Nash-Sutcliffe

Efficiency (NSE) as an objective function (Muleta, 2011; Nash and Sutcliffe, 1970). Details of the model parameters have already been described in Moore (2007).

The streamflow was calculated from both three rain gauges of the CEMADEN official network, and two other approximations: the maximum inter-station rainfall depth every 10 min, and the spatially-estimated mean precipitation depth, which were estimated by means of the Inverse Distance Weighting (IDW) method. Table 2 summarises the NSE values for the calibration and validation of the PDM model.

Table 2: NSE performance.

Sensor name	**NSE value (calibration)**	**NSE value (validation)**
Burgo Paulista	0.37	0.11
Cidade Tiradentes	0.39	-0.03
Boa Esperana	0.59	0.30
Max values	0.63	0.40
IDW	0.51	0.21

Transformation of authoritative rainfall data in streamflow depends on the calibration performed. In this case, the rainfall from authoritative gauges is used to model the streamflow in the same period of social media harvesting. The simulated streamflow will be later compared with the one obtained from the social media modelling and the real values from authoritative sources. Low performance in calibration and validation is probably due to problems in the rain gauges, as already mentioned.

4.2. Parameter fitting for the transformation function

To create the transformation function, three properties from people's behaviour in social media were assumed: proportionality, randomness and semantic singularity. First, it is supposed that people use more social media when discussing a phenomenon of great significance. In this case, the number of people talking about it will depend on how they were affected and thus, the intensity of the phenomenon might be directly proportional to the number of related tweets. This behaviour can be measured using bins of cumulative tweets over a certain period, depending on the duration of the phenomenon. Second, people do not "speak" in a synchronous way, namely, the users randomly post messages, before, during or after the phenomenon occurs (de Andrade et al., 2017). Third, people tend to use related words when the phenomenon becomes more intense/weaker or singular/unusual, which can lead to semantic singularities. For example, other hydrometeorological phenomena could be incorporated into the tweets because their beauty or intensity make people talk more about them. This brings about an increase in posting, with phrases, photos or videos, like a rainbow immediately after a storm, or the dazzling light of lightning flashes during a thunderstorm.

We propose a linear regression model between the frequency of social media data and the rainfall authoritative data for the signal conversion function to predict a proxy variable of rainfall data, with the following functional structure:

$$p_{social} = \alpha(1 + \eta_{strong} + \eta_{soft})\frac{f_{kw}}{A_{interest}} + \sum_{i=20}^{n} \beta_i \frac{F_{kw(i)}}{A_{interest}}$$

where p_{social} is the proxy of the precipitation variable resulting from the transformation of tweets to rainfall. The variable f_{kw} represents the absolute frequency of the number of tweets and the variable $F_{kw(i)}$ represents the accumulated absolute frequency for the number of tweets for i cumulative periods (with $i = 20, 30, 40, \ldots$ min). $A_{interest}$ is the area where tweets are being harvested, i.e. the city of Sao Paulo. Furthermore, η_{strong} and η_{soft} are two dummy variables that capture the multiplicative effect, in which some tweets have words that strengthen or reduce the intensity of the rainfall respectively. An example of a strong multiplicative effect is "heavy rain", whereas a weak multiplicative effect might imply the word "rainbow".

The system collects social media data by means of an API to fitting the transformation function. Following this, the messages are filtered by geotag and keywords. As a result, the frequency of keywords is obtained and the variables are created. Then, a 5-fold cross validation procedure for the fitting of the function is applied to regress the authoritative rainfall against social media data, which encompasses the whole city. In this procedure, one month is removed from the sample and used later to validate the transformation function of the same month, and avoid any bias in the resulting function. These stages are repeated to obtain a transformation function for each month.

4.3. Rainfall-runoff estimation from social media data using the transformation function

In transforming the social media data into a rainfall proxy, data were collected inside the catchment to obtain a rainfall proxy for this place. We collected the same variables with the same temporal resolution examined in Section 4.2. Once the tweets had been collected, the frequencies of the tweets were replaced inside the function created in the past section. However, since hydrological processes, like rainfall-runoff, are only possible in systems such as catchments, where the boundaries do not necessarily match the administrative boundaries of the city, a "regionalization" of the tweets within a catchment-area is carried out by dividing the frequencies of the related tweets every 10 min within the drainage area of the catchment. Thus, this process differs from the parameter fitting process where the whole area of the city is covered. Finally, the estimated rainfall values were used as input of the PDM hydrological model to generate the streamflow data.

4.4. Comparison of the joint use of traditional hydrological modelling and modelling from social media

This step involves comparing real streamflow values (from SAISP), with estimated streamflow values calculated from social media messages (Sect. 4.2) and with authoritative rainfall (from CEMADEN)-runoff modelling (Sect. 4.1). This comparison is made by determining if the real streamflow values are found within the confidence interval of the models, or have been overestimated/underestimated instead. This assessment makes it possible to establish the accuracy of these cases when the modelling is only carried out by means of social networks data, and employing the transformation function to estimate rainfall values for the "ungauged" catchments, i.e. when we do not have to rely on authoritative sensors. Additionally, we analysed the case when the results from both models are employed, by selecting the maximum and minimum values of the confidence interval of each model and evaluating their accuracy to predict real streamflow values. This scenario is equivalent to the case of "poorly gauged" catchments, where data from both sources is available but the authoritative data are inaccurate and/or imprecise.

5. Results

We estimated several linear regression models that were robust to heteroscedasticity to create the transformation functions for each month (see Table 3). Following the 5-fold cross validation procedure, each column summarises the data for the transformation function of each month. A small coefficient indicates that for this specific month the people wrote tweets related to rain in a more synchronous way with the rainfall measurements. That is why in December all the coefficients decrease in magnitude.

Based on these results, some simulations were carried out within the Aridancuva catchment using related tweets and authoritative rainfall data; these were incorporated into the PDM rainfall-runoff model. Figure 8.a shows the period from January 25^{th} to January 31^{st}, 2016. It can be seen that for the rainfall events of January 26^{th} and 28^{th}, the proxy variable from Twitter performed better than the one with authoritative rainfall data. However, in the period after January 29^{th}, the behaviour of the variables generated by social media considerably overestimated the streamflow values.

In turn, in Fig. 8.b, it was observed that on December 10^{th}, there is a peak in the simulation carried out by the social media proxy, which was not found either in the real value or in the authoritative model. From the end of December 10^{th} until December 12^{nd}, it was observed that only the model with authoritative data followed the streamflow pattern. However, none of them provided a suitable estimate for the highest peak streamflow, (the one above 200 m^3/s).

Moreover, in the period from January 20^{th} to 28^{th}, 2017, Fig. 8.c shows how the Twitter proxy variable reacted to all the observed peaks of the time series. It was only in some cases, such as on January 25^{th}, that this reaction took place after the flood occurrence, except on January 26^{th}, when the geo-social media reacted a bit earlier. In contrast, the streamflow only estimated from the authoritative data when the modelling was conducted in a suitable way.

For the period from February 1^{st} to February 9^{th}, 2017 (Fig. 8.d), it was observed that both simulations, whether carried out with the social media proxy or with authoritative data, follow the pattern of the streamflow. However, the authoritative model did not perform well for the first peak of streamflow, (above 200 m^3/s); on the contrary, the social media-based model reacted late, although it had a suitable magnitude. Moreover, from the end of February 6^{th} until February 7^{th}, the model that was based on social media reacted better.

In Fig. 8.e, there are 5 peaks close to 100 m^3/s for the period from February, 22^{nd} to February, 28^{th}, 2017 and it can be observed that sometimes the authorized data performs better while sometimes the social media proxy data does. However, on February 25^{th} when there was a peak in the streamflow with a value greater than 700 m^3/s, the social media streamflow proxy captured it more accurately. This pattern is probably due to convective rainfall, which is concentrated in some parts of the catchment area far away from the available rainfall gauges.

Table 3: Regression coefficients for the parameter fitting of the transformation function of geo-social data.

Coefficients	January 2016	November 2016	December 2016	January 2017	February 2017
α	322.5 ± 214.4	436.0 ± 234.6	-	427.4 ± 268.0	231.3 ± 210.2
β	547.0 ± 83.2	607.5 ± 83.8	134.7 ± 23.4	558.5 ± 92.6	563.0 ± 80.2
η_{strong}	-	-	329.0 ± 251.2	-	812.8 ± 497.8
η_{soft}	-872.4 ± 385.8	-1236.0 ± 312.2	-255.5 ± 76.4	-993.7 ± 443.0	-1129.7 ± 476.2
R^2adj	0.283	0.294	0.220	0.257	0.255

A summary of the streamflow simulation is shown in Table 4. Based on the values of the proxy variable obtained from Twitter, the simulation provides correct values in 31.3% of the cases, while overestimation is found in 19.0% and underestimation in 49.5% of the cases for the entire period. In the case of modelling with authoritative rainfall gauges, the real values are in the correct range of 38.6%, while underestimation and overestimation are found in around 58.4% and 3.0% of the cases, respectively.

We also simulated a combined rainfall variable consisting of the social media proxy variable and the rainfall gauge. In this case, the accuracy of the forecasting significantly increases, since it is able to predict the value of the real streamflow correctly in about 70.9% of the cases. The underestimation is reduced to 28.6% and there is no overestimation for the period. This significant result clearly shows the potential value of using data from social media to assist in monitoring environmental problems such as floods. An example of the combined simulation for the period from January 25^{th} to January 31^{st}, 2016 is shown in Fig. 8.f.

Table 4: Percentage of correct estimates, and cases of overestimation and underestimation of the streamflow within the confidence interval, with the use of social media and authoritative data.

	Social media only	Authoritative sensor only	Composite of social media and authoritative sensors
Observations of estimates within the model's confidence interval	31.3	38.6	70.9
Observations of cases that were underestimated	49.5	58.4	28.6
Observations of cases that were overestimated	19.0	3.0	0.5

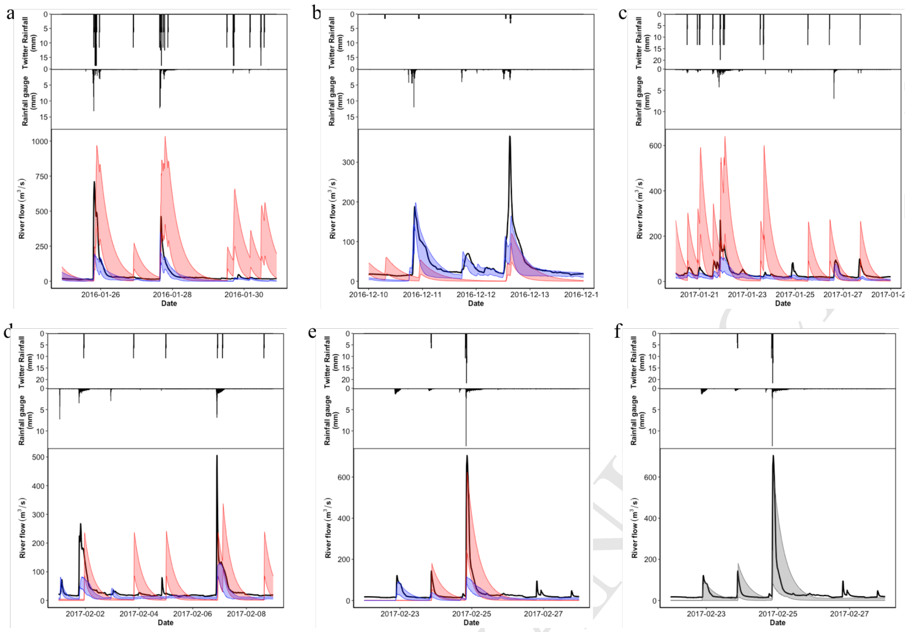


Figure 8: Examples of social media rainfall (upper, time series) and authoritative rainfall (centre, time-series), with simulated streamflow (shaded) and observed streamflow (line in bold) at the Aricanduva catchment. Streamflow simulation using only authoritative sensors are shaded in blue and simulation from social media are shaded in red.

6. Discussion

The results of this study support the use of social media information to estimate the precipitation rate or flow in poorly gauged catchments, which could help in issuing early flood warnings. In the catchments that are currently in operation, but where there are incomplete records or with sensors undergoing maintenance, the use of alternative, social media proxy variables could become even more useful. Posting and sharing information through social media where it is capable of being transformed into viable proxy variables, as an alternative monitoring data source, is a means of heightening people's awareness and is of value for fostering community resilience, especially for streamflow monitoring, and forecasting purposes. Another possible application of social media-based information lies in detecting authoritative sensors that have on-line problems, and thus require maintenance.

The results of this study complement and extend previous research in the area. For instance, Mazzoleni et al. (2017) designed a hydrological model with data collected by citizens to improve the accuracy of flood forecasts and showed that these data can reinforce the traditional monitored areas provided by static sensor networks. However, these data do not come from social media, but from citizen observatories, which are a more structured form of crowdsourced geographic data, based on dedicated data collection platforms (Degrossi et al., 2014; de Albuquerque et al., 2015), and are more difficult to disseminate than widely used social media platforms. In contrast, Rosser et al. (2017) used geo-referenced photographs from social media, optical remote sensing, and high-resolution terrain maps, to design a Bayesian statistical model that estimates the probability of floods through weight-of-evidence analysis. However, they only used these data to generate flood maps, which might detect the occurrence of floods through an ex-post evaluation, but were not able to assist forecasting impending events.

In this paper, we obtained modest values for the Adjusted Coefficient of Determination ($R^2adj < 0.30$) in the equations that transforms social media data into precipitation, a result that complements our previous results discussed in de Andrade et al. (2017). The fact that these values are low, can perhaps be attributed to problems with a) the quality of the rainfall gauge information, b) the modelling resolution and c) the different time synchronism of the sensors collected from different sources, i.e. national centers, and state agencies with the social media posts. However, this temporal resolution is crucial for timing hydrological responses like streamflows at an urban catchment. Moreover, these values could probably be improved with the aid of other social media platforms (e.g. Instagram, Flickr) or by including other variables such as information quality protocols, the spatiotemporal context, literacy and the economic circumstances of the citizens posting social media, as well as the content of information, among other factors. In addition, other methods could be tested to transform the signal by using other transformation algorithms to achieve a better performance.

It is worth noting that the messages we used here are not discriminated by the temporal context in which they were published, but only filtered by types of keywords or by their spatial location, and this might be another limitation of the model. Additional research should be carried out to review the information with regard to the type of temporal context of the messages before, during or after the rainfall events or thunderstorms. In this area, the focal point of our study has been on monitoring but future studies should take into account how a real-time environmental application can be formed.

7. Conclusion

This paper provides strong evidence that data from geo-social media can be used to derive proxy variables for rainfall and streamflow. The frequency of related messages from social media was used as a proxy for rainfall, which in turn can provide input for hydrological models to predict streamflows and flood conditions. Data from social media could be used to assist in issuing early flood warnings and to improve rainfall-runoff from observational, authoritative networks and even observed urban streamflow. Evidence showed that better results can be achieved by merging authoritative data with information from social media. The available social media data on its own should be treated with caution, because of the risk of bias and uncertainty with regard to streamflow estimation. In future research, the methods and results might be further compared with other studies, i.e. from different catchments, with several rainfall-runoff events and various time-collection periods. Despite any limitations, it is hoped that the methods employed in this paper can assist in making multiple sources of data and information more available and thus make cities more resilient to extreme events such as floods.

8. Acknowledgements

C. Restrepo-Estrada is grateful for the financial support from CAPES-PROEX. S.C. Andrade would like to thank the agencies São Paulo Research Foundation (FAPESP) grant no. #2017/15413-0, Araucária Research Foundation in Support of Scientific, and Technological Development in the State of Paraná (FAPPR), and State Secretariat of Science, Technology and Higher Education of Paraná (SETI) for their financial support. The authors would also like to thank the research funding grants provided by: the Engineering and Physical Sciences Research Council (EPSRC) through the Global Challenges Research Fund, CAPES #88887.091743/2014-01 (ProAlertas CEPED/USP), CNPq [National Council for Scientific and Technological Development] #465501/2014-1, FAPESP #2014/50848-9 & INCT–II (Climate Change, Water Security) CNPq #312056/2016-8 (EESC-USPCEMADEN/MCTIC) & CAPES PROEX (PPGSHS EESC USP).

References

de Albuquerque, J.P., Herfort, B., Brenning, A., Zipf, A., 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. International Journal of Geographical Information Science 29, 667–689. doi:10.1080/13658816.2014.996567.

de Albuquerque, J.P., Horita, F.E.A., Degrossi, L.C., Rocha, R.d.S., de Andrade, S.C., Restrepo-Estrada, C., Leyh, W., 2017. Leveraging Volunteered Geographic Information to Improve Disaster Resilience. pp. 158–184. URL: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-5225-2446-5.ch009, doi:10.4018/978-1-5225-2446-5.ch009.

Alvarez-Garreton, C., Ryu, D., Western, A., Crow, W., Robertson, D., 2014. The impacts of assimilating satellite soil moisture into a rainfall–runoff model in a semi-arid catchment. Journal of hydrology 519, 2763–2774. doi:10.1016/j.jhydrol.2014.07.041.

de Andrade, S.C., Restrepo-Estrada, C., Delbem, A.C.B., Mendiondo, E.M., de Albuquerque, J.P., 2017. Mining Rainfall Spatio-Temporal Patterns in Twitter: A Temporal Approach. Springer International Publishing, Cham. pp. 19–37. doi:10.1007/978-3-319-56759-4_2.

Boni, G., Ferraris, L., Pulvirenti, L., Squicciarino, G., Pierdicca, N., Candela, L., Pisani, A.R., Zoffoli, S., Onori, R., Proietti, C., Pagliara, P., 2016. A prototype system for flood monitoring based on flood forecast combined with cosmo-skymed and sentinel-1 data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9, 2794–2805. doi:10.1109/JSTARS.2016.2514402.

Brouwer, T., Eilander, D., van Loenen, A., Booij, M.J., Wijnberg, K.M., Verkade, J.S., Wagemaker, J., 2017. Probabilistic flood extent estimates from social media flood observations. Natural Hazards and Earth System Sciences 17, 735–747. URL: https://www.nat-hazards-earth-syst-sci.net/17/735/2017/, doi:10.5194/nhess-17-735-2017.

Chen, X., Zhang, L., Gippel, C.J., Shan, L., Chen, S., Yang, W., 2016. Uncertainty of flood forecasting based on radar rainfall data assimilation. Advances in Meteorology 2016. doi:10.1155/2016/2710457.

Crochemore, L., Ramos, M.H., Pappenberger, F., 2016. Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrology and Earth System Sciences 20, 3601–3618. URL: https://www.hydrol-earth-syst-sci.net/20/3601/2016/, doi:10.5194/hess-20-3601-2016.

Crooks, A., Croitoru, A., Stefanidis, A., Radzikowski, J., 2013. # earthquake: Twitter as a distributed sensor system. Transactions in GIS 17, 124–147. doi:10.1111/j.1467-9671.2012.01359.x.

Degrossi, L.C., de Albuquerque, J.P., Fava, M.C., Mendiondo, E.M., 2014. Flood citizen observatory: a crowdsourcing-based approach for flood risk management in brazil., in: SEKE, pp. 570–575.

Enenkel, M., See, L., Bonifacio, R., Boken, V., Chaney, N., Vinck, P., You, L., Dutra, E., Anderson, M., 2015. Drought and food security–improving decision-support via new technologies and innovative collaboration. Global Food Security 4, 51–55. doi:10.1016/j.gfs.2014.08.005.

Fraternali, P., Castelletti, A., Soncini-Sessa, R., Ruiz, C.V., Rizzoli, A., 2012. Putting humans in the loop: Social computing for water resources management. Environmental Modelling & Software 37, 68–77. URL: http://www.sciencedirect.com/science/article/pii/S1364815212000849, doi:10.1016/j.envsoft.2012.03.002.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221. doi:10.1007/s10708-007-9111-y.

Goodchild, M.F., Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. International Journal of Digital Earth 3, 231–241. doi:10.1080/17538941003759255.

Hapuarachchi, H., Wang, Q., Pagano, T., 2011. A review of advances in flash flood forecasting. Hydrological processes 25, 2771–2784. doi:10.1002/hyp.8040.

Horita, F.E., de Albuquerque, J.P., Degrossi, L.C., Mendiondo, E.M., Ueyama, J., 2015. Development of a spatial decision support system for flood risk management in brazil that combines volunteered geographic information with wireless sensor networks. Computers & Geosciences 80, 84 – 94. URL: http://www.sciencedirect.com/science/article/pii/S0098300415000746, doi:10.1016/j.cageo.2015.04.001.

Horita, F.E., de Albuquerque, J.P., Marchezini, V., Mendiondo, E.M., 2017. 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. URL: http://www.sciencedirect.com/science/article/pii/S0167923617300416, doi:10.1016/j.dss.2017.03.001.

Huang, Q., Xiao, Y., 2015. Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. ISPRS International Journal of Geo-Information 4, 1549–1568. doi:10.3390/ijgi4031549.

Kay, A., Davies, H., Bell, V., Jones, R., 2009. Comparison of uncertainty sources for climate change impacts: flood frequency in england. Climatic Change 92, 41–63. doi:10.1007/s10584-008-9471-4.

Lamb, R., 1999. Calibration of a conceptual rainfall-runoff model for flood frequency estimation by continuous simulation. Water Resources Research 35, 3103–3114. doi:10.1029/1999WR900119.

Li, Y., Grimaldi, S., Walker, J.P., Pauwels, V., 2016. Application of remote sensing data to constrain operational rainfall-driven flood forecasting: a review. Remote Sensing 8, 456. doi:10.3390/rs8060456.

Li, Z., Wang, C., Emrich, C.T., Guo, D., 2017. A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 south carolina floods. Cartography and Geographic Information Science 0, 1–14. doi:10.1080/15230406.2016.1271356.

Mazzoleni, M., Verlaan, M., Alfonso, L., Monego, M., Norbiato, D., Ferri, M., Solomatine, D.P., 2017. Can assimilation of crowdsourced data in hydrological modelling improve flood prediction? Hydrology and Earth System Sciences 21, 839. doi:10.5194/hess-21-839-2017.

Mersham, G., 2010. Social media and public information management: the september 2009 tsunami threat to new zealand. Media International Australia 137, 130–143.

Moore, R., 2007. The pdm rainfall-runoff model. Hydrology and Earth System Sciences Discussions 11, 483–499. doi:10.5194/hess-11-483-2007.

Moradkhani, H., Sorooshian, S., Gupta, H.V., Houser, P.R., 2005. Dual state–parameter estimation of hydrological models using ensemble kalman filter. Advances in water resources 28, 135–147. doi:10.1016/j.advwatres.2004.09.002.

Muleta, M.K., 2011. Model performance sensitivity to objective function during automated calibrations. Journal of hydrologic engineering 17, 756–767. doi:10.1061/(ASCE)HE.1943-5584.0000497.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part ia discussion of principles. Journal of hydrology 10, 282–290. doi:10.1016/0022-1694(70)90255-6.

Ochoa-Rodriguez, S., Wang, L.P., Gires, A., Pina, R.D., Reinoso-Rondinel, R., Bruni, G., Ichiba, A., Gaitan, S., Cristiano, E., van Assel, J., et al., 2015. Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation. Journal of Hydrology 531, 389–407. URL: https://doi.org/10.1016/j.jhydrol.2015.05.035.

Patankar, A., Patwardhan, A., 2016. Estimating the uninsured losses due to extreme weather events and implications for informal sector vulnerability: a case study of mumbai, india. Natural Hazards 80, 285–310. doi:10.1007/s11069-015-1968-3.

Patel, N.N., Stevens, F.R., Huang, Z., Gaughan, A.E., Elyazar, I., Tatem, A.J., 2017. Improving large area population mapping using geotweet densities. Transactions in GIS 21, 317–331. doi:10.1111/tgis.12214.

Rathore, M., Ahmad, A., Paul, A., Hong, W.H., Seo, H., 2017. Advanced computing model for geosocial media using big data analytics. Multimedia Tools and Applications doi:10.1007/s11042-017-4644-7.

Rosser, J.F., Leibovici, D.G., Jackson, M.J., 2017. Rapid flood inundation mapping using social media, remote sensing and topographic data. Natural Hazards , 1–18doi:10.1007/s11069-017-2755-0.

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes twitter users: Real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, ACM, New York, NY, USA. pp. 851–860. doi:10.1145/1772690.1772777.

Schnebele, E., Cervone, G., Waters, N., 2014. Road assessment after flood events using non-authoritative data. Natural Hazards and Earth System Sciences 14, 1007–1015. URL: https://www.nat-hazards-earth-syst-sci.net/14/1007/2014/, doi:10.5194/nhess-14-1007-2014.

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J., Mendiondo, E., O'connell, P., et al., 2003. Iahs decade on predictions in ungauged basins (pub), 2003-2012: Shaping an exciting future for the hydrological sciences. Hydrological sciences journal 48, 857–880. doi:10.1623/hysj.48.6.857.51421.

Skinner, C.J., Bellerby, T.J., Greatrex, H., Grimes, D.I., 2015. Hydrological modelling using ensemble satellite rainfall estimates in a sparsely gauged river basin: The need for whole-ensemble calibration. Journal of Hydrology 522, 110–122. doi:10.1016/j.jhydrol.2014.12.052.

Smith, L., Liang, Q., James, P., Lin, W., 2015. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework doi:10.1111/jfr3.12154.

Tiesi, A., Miglietta, M.M., Conte, D., Drofa, O., Davolio, S., Malguzzi, P., Buzzi, A., 2016. Heavy rain forecasting by model initialization with laps: A case study. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9, 2619–2627. doi:10.1109/JSTARS.2016.2520018.

Tkachenko, N., Jarvis, S., Procter, R., 2017. Predicting floods with flickr tags. PLOS ONE 12, 1–13. doi:10.1371/journal.pone.0172870.

Wang, L.P., Ochoa-Rodríguez, S., Van Assel, J., Pina, R.D., Pessemier, M., Kroll, S., Willems, P., Onof, C., 2015. Enhancement of radar rainfall estimates for urban hydrology through optical flow temporal interpolation and bayesian gauge-based adjustment. Journal of Hydrology 531, 408–426. doi:10.1016/j.jhydrol.2015.05.049.

Weng, J., Lee, B.S., 2011. Event detection in twitter.

- Authoritative and social media data are integrated for rainfall and flow estimation.
- New transformation function of social media posts into rainfall.
- Combined use of tweets and rainfall could be used in issuing early flood warnings.