UNIVERSITY OF SÃO PAULO
SAO CARLOS SCHOOL OF ENGINEERING

CAMILO ERNESTO RESTREPO ESTRADA

Use of social media data in flood monitoring

Corrected Final Version
São Carlos
November 2018

CAMILO ERNESTO RESTREPO ESTRADA

Use of social media data in flood monitoring

Doctoral thesis presented at *Escola de Engenharia de São Carlos* (São Carlos School of Engineering), of *Universidade de São Paulo* (University of São Paulo), in partial fulfilment of the requirements for obtaining the Degree of Doctor in Science: Hydraulics and Sanitary Engineering.

Advisor: Prof. Dr. Eduardo Mario Mendiondo

Co-Advisor: Prof. Assoc. João Porto De Albuquerque Pereira

Corrected Final Version

São Carlos

November 2018

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Restrepo-Estrada, Camilo Ernesto
R183u Uso de dados das mídias sociais no monitoramento de enchentes / Camilo Ernesto Restrepo-Estrada; orientador Eduardo Mario Mendiondo; coorientador João Porto de Albuquerque. São Carlos, 2018.

Tese (Doutorado) - Programa de Pós-Graduação em Engenharia Hidráulica e Saneamento e Área de Concentração em Hidráulica e Saneamento -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2018.

1. Mídias sociais. 2. Modelagem hidrológica. 3. Estimação de vazão. 4. Monitoramento de enchentes. 5. Assimilação de dados. 6. Ensemble Kalman Filter. 7. Fusão de dados. 8. Probability Distributed Model. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE JULGAMENTO

Candidato: Engenheiro **CAMILO ERNESTO RESTREPO ESTRADA.**

Título da tese: "Uso de dados das mídias sociais no monitoramento de enchentes".

Data da defesa: 05/11/2018.

Comissão Julgadora: **Resultado:**

Prof. Dr. **Eduardo Mario Mendiondo**
(Orientador)
(Escola de Engenharia de São Carlos/EESC) Aprovado

Prof. Dr. **Francisco Martínez-Álvarez**
(Pablo de Olavide University) Aprovado

Dra. **Luz Adriana Cuartas Pineda**
(Centro Nacional de Monitoramento e Alertas de Desastres Naturais/CEMADEN) Aprovado

Prof. Dr. **Oscar David Álvarez-Villa**
(Universidad de Antioquia) Aprovado

Prof. Titular **Alexandre Cláudio Botazzo Delbem**
(Instituto de Ciências Matemáticas e de Computação/ICMC-USP) Aprovado

Coordenador do Programa de Pós-Graduação em Engenharia Hidráulica e Saneamento:
Prof. Dr. **Eduardo Mario Mendiondo**

Presidente da Comissão de Pós-Graduação:
Prof. Associado **Luís Fernando Costa Alberto**

DEDICATORIA

A mi familia y amigos. Todos fueron parte de este viaje. Gracias también a los que alguna vez me cerraron las puertas o dijeron adiós, para todos solo afecto.

AGRADECIMIENTOS

Agradezco a mi familia. Mis padres Pedro y Marta y mi hermana Maria Isabel por apoyarme en otra etapa de formación.

A mi orientador Eduardo Mario Mendiondo por su infinita colaboración.

A mi gran amigo Sidgley, un amigo de batallas y de artículos. Una ayuda invaluable la que me brindó.

A mis de la vida, las cervezas, los lanches y las reflexiones Ana, Darío, Maria Isabel, Diego, Karol, Ramón, Mónica, Jorge, Jesús, Carolina, Leidy, Santiago, Anderson y Marco.

A mis amigos del NIBH, Diego, Guilherme, Danielle, Felipão, Marininha, Marie Claire, Naruminho, Cesarinho, Txe, Denisse, Altair, Clarissa, Caroline, Dulce por la amistad, la compañía y las horas compartiendo y apoyando.

A AGORA y el Profesor João Porto de Albuquerque por sus aportes invaluables y sus integrantes y amigos Flavio, Livia, Sidgley.

A la doctora Luz Adriana Cuartas por su apoyo académico y humano. Ayudó mucho en las horas inciertas.

A los que estuvieron y los que se fueron. A los que se me olvidaron injustamente. A todos los que han logrado que yo esté en este punto hoy. A los que dijeron no, a los que dijeron sí. A la vida.

Quiero agradecer de manera especial por el apoyo financiero de la beca CAPES-PROEX. S.C.

ABSTRACT

Restrepo-Estrada, Camilo Ernesto. Use of social media data in flood monitoring. Doctoral Thesis, São Carlos School of Engineering, University of São Paulo, São Carlos-SP, Brazil. 2018.

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. This thesis aims to show a novel methodology that shows a way to close the research gap regarding the use of social networks as a proxy for precipitation-runoff and flood forecast estimates. To address this, it is proposed to use a transformation function that creates a proxy variable for rainfall by analysing messages from geo-social media and precipitation measurements from authoritative sources, which are then incorporated into a hydrological model for the flow estimation. Then the proxy and authoritative rainfall data are merged to be used in a data assimilation scheme using the Ensemble Kalman Filter (EnKF). It is found that the combined use of authoritative rainfall values with the social media proxy variable as input to the Probability Distributed Model (PDM), improves flow simulations for flood monitoring. In addition, it is found that when these models are made under a scheme of fusion-assimilation of data, the results improve even more, becoming a tool that can help in the monitoring of "ungauged" or "poorly gauged" catchments. The main contribution of this thesis is the creation of a completely original source of rain monitoring, which had not been explored in the literature in a quantitative way. It also shows how the joint use of this source and data assimilation methodologies aid to detect flood events.

Key words: Social Media, Hydrological modelling, Streamflow estimation, Flood monitoring, Data assimilation, Ensemble Kalman Filter, Data Fusion, Probability Distributed Model.

RESUMO

Restrepo-Estrada, Camilo Ernesto. Uso de dados das mídias sociais no monitoramento de enchentes. Tese doutoral, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos-SP, Brasil. 2018.

As inundações são um dos tipos mais devastadores de desastres em todo o mundo em termos de perdas humanas, econômicas e sociais. Se os dados oficiais forem escassos ou indisponíveis por alguns períodos, outras fontes de informação são necessárias para melhorar a estimativa de vazões e antecipar avisos de inundação. Esta tese tem como objetivo mostrar uma metodologia que mostra uma maneira de fechar a lacuna de pesquisa em relação ao uso de redes sociais como uma proxy para as estimativas de precipitação e escoamento. Para resolver isso, propõe-se usar uma função de transformação que cria uma variável proxy para a precipitação, analisando mensagens de medições geo-sociais e precipitação de fontes oficiais, que são incorporadas em um modelo hidrológico para a estimativa de fluxo. Em seguida, os dados de proxy e precipitação oficial são fusionados para serem usados em um esquema de assimilação de dados usando o *Ensemble Kalman Filter* (EnKF). Descobriu-se que o uso combinado de valores oficiais de precipitação com a variável proxy das mídias sociais como entrada para o modelo distribuído de probabilidade (*Probability Distributed Model* - PDM) melhora as simulações de fluxo para o monitoramento de inundações. A principal contribuição desta tese é a criação de uma fonte completamente original de monitoramento de chuva, que não havia sido explorada na literatura de forma quantitativa.

Palavras chave: Mídias sociais, Modelagem hidrológica, Estimação de vazão, Monitoramento de enchentes, Assimilação de dados, Ensemble Kalman Filter, Fusão de dados, Probability Distributed Model.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF VARIABLES

Time of concentration

t_c: time of concentration [min]

L: mainstream longitude [km]

A: catchment area [km^2]

S: average slope of mainstream [m/m]

H: elevation difference between divide and outlet [m]

PDM variables

c: storage capacity

c_{min}: minimum value of storage capacity

c_{max}: maximum value of storage capacity

$\bar{c}$: mean storage capacity

$c^*(t)$: critical storage capacity

S_{max} : total available capacity

$S(t)$: total water in storage over the catchment

S_t: threshold storage

b: represents the degree of control of spatial variability over the catchment

E_i': actual evapotranspiration

E_i: potential evapotranspiration

b_e: exponent of the evaporation function

d_i: rate of drainage over the interval

k_g: temporary constant of drainage of groundwater

π_i: net rainfall

P_i: rainfall

V: volume

k_b: time constant of base flow

u: rate of inflow to the store per unit area

k_1 and k_2: temporary cascade constants of two linear reservoirs

q_t: surface runoff

NSE: Nash-Sutcliffe Efficiency

$\mathrm{NSE} - \log$: Nash-Sutcliffe Efficiency

$Q_{sim,i}$: model simulation discharge

$Q_{obs,i}$: observed discharge

N: number of total periods of simulation

Geo-social transformation function

p_{social}: proxy of the variable precipitation that comes from the transformation from tweets to rainfall

f_{kw}: represents the absolute frequency of the number of tweets

$F_{kw(i)}$: represents the accumulated absolute frequency for the number of tweets for a cumulative of i periods (with i = {20, 30, 40, …} minutes)

$A_{interest}$: area where tweets are being harvested

η_{strong}: dummy variable that capture the multiplicative effect, which have words that reinforce the intensity of the rainfall

η_{soft}: dummy variable that capture the multiplicative effect, which have words that diminish the intensity of the rainfall

α and β_i: regression coefficients

Data fusion and assimilation

x_t: state variables in the step of time t (e.g., soil moisture or flows)

u_t: forcing data (e.g., precipitation and evapotranspiration)

ξ_t^i error for the forcing variable

Σ_t^u: covariance for the error in the forcing variable

$f(\cdot)$: model of the system (in this case PDM)

θ: parameters of the hydrological model (in this case PDM)

$\hat{y}_t$: simulated observation

$h(\cdot)$: function that maps x_t into $\hat{y}_t$

y_t: observations (e.g., streamflow or water level)

η_t: error for observations

Σ_t^y: covariance of the measurement error

Σ_t^x: forecast state error covariance

K_t: Kalman gain in the time t

$\Sigma_t^{x\hat{y}}$: mapping error covariance

$\Sigma_t^{\hat{y}}$: covariance of the mapping error

N_{ens}: number of members in data assimilation

v_t: authoritative rainfall

Σ_t^{v}: authoritative rainfall error covariance

Σ_t^{SM}: social media rainfall error covariance

λ: fusion coefficient

ϕ_t: authoritative rainfall error

ω_t: social media rainfall error

a_h and b_h: parameters of heteroscedastic rainfall simulation

e: parameter of heteroscedastic streamflow simulation

INTRODUCTION

A. Theoretical background

Urbanization and structural measures that come with the agglomeration of people in cities, such as canalization of river courses, waterproofing of soils among others had generated a gradually increase in the floods' frequency throughout the world, causing serious levels of human, economic and social losses (Crochemore, Ramos and Pappenberger, 2016; Patankar and Patwardhan, 2016). To mitigate the associated problems with floods structural and non-structural measures are used. Structural measures such as reservoirs, dams, containment works among others are sometimes insufficient or costly and, they must be accompanied by non-structural measures such as monitoring, modelling and forecasting.

Urban hydrological modelling is a non-structural measure for monitoring that in some cases requires high-resolution rainfall and watershed information, for example, DEM, land use, among others. (Hapuarachchi, Wang and Pagano, 2011; Ochoa-Rodriguez *et al.*, 2015; Wang *et al.*, 2015). Rainfall data are the main input in rainfall-driven hydrological models for flood modelling and forecasting. Several approaches have been tested for different situations as alternative to the traditional use of *in-situ* measurements, highlighting the use of remote sensing for rainfall-driven flood forecasting (Skinner *et al.*, 2015; Li *et al.*, 2016). For example, Boni et al. (2016) implemented a near real-time flood-mapping algorithm using "Synthetic Aperture Radar" (SAR) together with satellite and coupled to a hydraulic model to forecast floods. Tiesi et al. (2016) used surface network data, radio-sounding profiles, radar and satellite (SEVIRI/MSG) data on quantitative precipitation forecasting and found a positive impact on the intensity and distribution of simulated rainfalls. In addition, studies such as the ones reported by Wang et al. (2015) and Chen et al. (2016) showed that radar-based precipitation measurements have the advantage of reproducing the spatial structure of rainfall fields and its variation in time in relation to ground-based measurements. However, these models so far have not reached the accuracy and temporal resolution required for urban hydrology.

In any rigorous hydrological modelling, it is important to characterize and/or reduce uncertainty, which can be of two types. The first one, known as random uncertainty, appears because the hydro-meteorological variables have a stochastic character. The second one is the epistemic uncertainty and it is occasioned due to our lack of

understanding of the hydrological system or our incapacity to model it. In general, uncertainty appears because of four main sources: the input and output variables, the parameters' calibration and the conceptual structure of the model, and different approaches have been developed in order to reduce uncertainty in prediction. These tools serve to improve the prediction capability of the model in light of new information. A classic way is based on the Monte Carlo approach, in which massive random sampling is applied. In other cases, mathematical tools, such as data assimilation, are also commonly used. The most widely method of data assimilation used is the Kalman Filter (KF), which provides an optimal estimation for the state variables in dynamic linear systems. However, for non-linear systems such as hydrological models, other types of filters are more adequate, such as the Ensemble Kalman Filter (EnKF), the Particle Filter, among others (Moradkhani *et al.*, 2005; Mazzoleni *et al.*, 2017; Leach, Kornelsen and Coulibaly, 2018).

Leach, Kornelsen, and Coulibaly (2018) used the Ensemble Kalman Filter (EnKF) in modelling of urban basins. They found that the best data assimilation scheme for hydrological modelling involved the use of different data sources. Mazzoleni et al. (2017) used the asynchronous EnKF filter to model jointly crowdsourcing data and data from classic static physical sensors. They found that data geared by citizens can complement traditional sensor networks. Pathiraja et al. (2016) analysed how the hydrological prediction is affected in the context of the change in land uses. For this, they used the EnKF coupled to a Probability Distributed Model (PDM) for the updating of states and parameters. They found that the technique of data assimilation is adequate and improves the predictions.

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 makings, such as satellites, radar systems, rainfall gauges and hydrological networks (Horita *et al.*, 2017). However, the apparent over-abundance of data is not always available for monitoring during crisis management and instead, this problem faces a lack of information: an "information dearth". This situation may arise because either sensors are not available in that moment 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).

Figure 1 shows an example of difficulties that a situation room, such as the one in CEMADEN, may face when there are problems with authoritative data. The image was taken on February 2, 2017 from the

official interactive map[1]. It can be seen that on this date there were some sensors that did not report data at all (black points), as well as there were apparent inconsistencies in the measurements made by other sensors, concerning the amount of rainfall that fell in the city of Sao Paulo. Such situations offer a further motivation for using alternative information sources to assist flood monitoring and early warning.

In response, some alternative sources are emerging that provide important information and can supplement the traditional approach of mainly rely on sensors. These sources include data generated by people living in 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). This spatial information is produced by ordinary people through different collaborative activities, such as exchanging information through geotagged social media messages, or platforms. Using digital technologies to employ efficiently the information that people can generate has been helping the communities' resilience. This approach has been increasingly recognised and used as an important resource to support decision making during disaster management (Goodchild and Glennon, 2010; Horita *et al.*, 2015; Andrade *et al.*, 2017; Porto de Albuquerque *et al.*, 2017; Restrepo-Estrada *et al.*, 2018).

Two types of georeferenced data sources coming from citizens can be identified. First, the so-called Volunteered Geographical Information (VGI), where people voluntarily decide to provide information about phenomena and situations that will be used by others in decision-making (Goodchild, 2007). The second form has opportunistic characteristics because people share photos, texts or audios on social media on any subject without any intention to help. However, they are unconsciously helping because it is possible through machine learning methods to extract the information and generate some benefit from it. This is possible because when a user registers in a social media for free, he is also authorizing the use of their shared data.

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

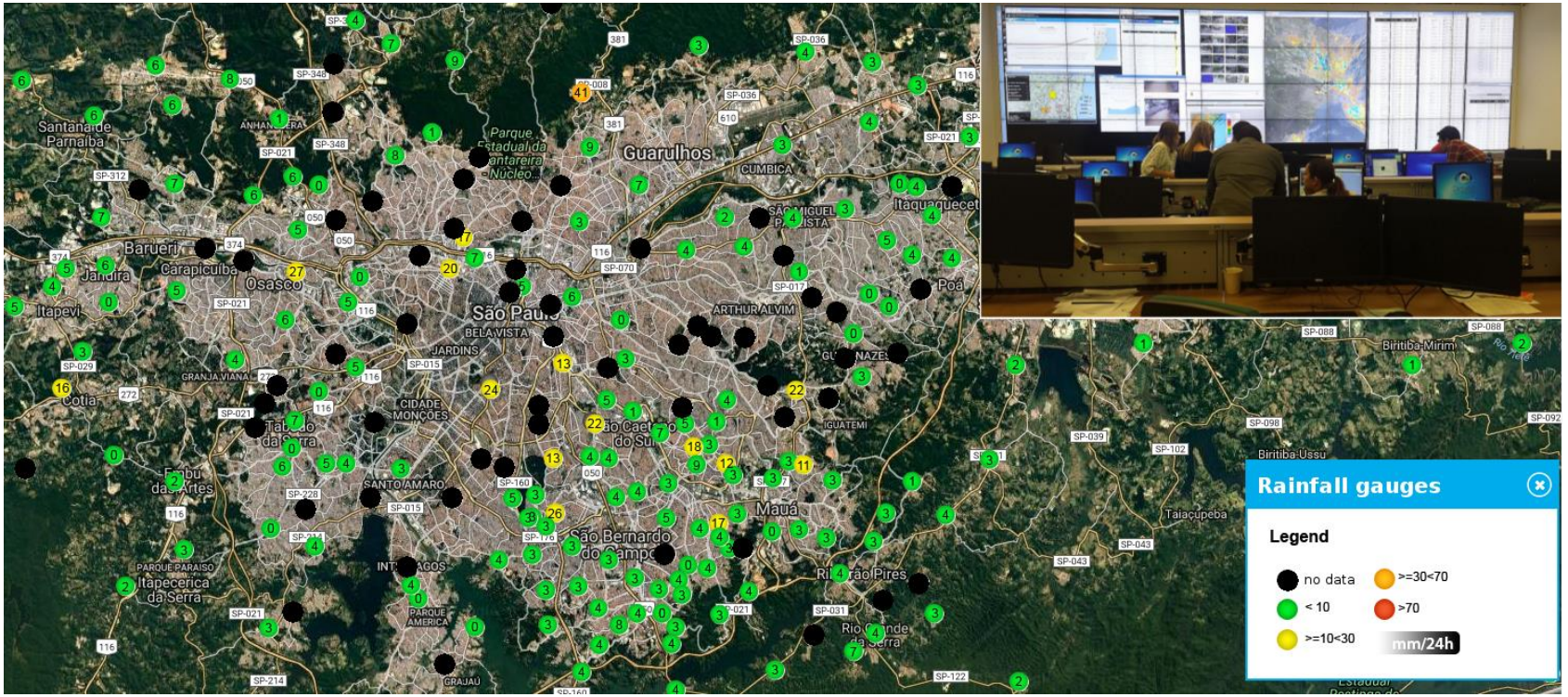


Figure 1. Problems with authoritative data, February 2, 2017.

The use of geo-social media in disaster management has been also explored in other types of hazards such as earthquakes (Sakaki and Matsuo, 2012; Crooks *et al.*, 2013), forest fires (Spinsanti and Ostermann, 2013), hurricanes (Huang and Xiao, 2015), tsunamis (Mersham, 2010), agricultural droughts (Enenkel *et al.*, 2015), floods (Weng and Lee, 2011; Smith *et al.*, 2017; Tkachenko, Jarvis and Procter, 2017), among others. In the specific context of flood management, scientific work has focused on using social media data in two directions, flood mapping and exploring spatiotemporal patterns.

Regarding water resources, tweets have been quantitatively used either in forecasting or in mapping. Schnebele et al. (2014) concluded that 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 Hurricane Sandy damage maps. Brouwer et al. (2017) harvested 8000 flood-related tweets from the York city in England and used this information to construct a probabilistic flood extent map. Patel et al. (2017) used tweets to produce population maps. Rathore et al. (2017) proposed a system that uses geo-social media to harvest, process, and analyse a large amount of data at high-speed from Twitter and makes decisions in real time. Li et al. (2018) collected tweets during 18 days in South Carolina, USA, filtering by flood-related keywords, and found 4268 flood-related tweets. With these information, and using a daily temporal granularity, they found a strong correlation between stream gauge levels and the absolute frequency of flood-related tweets. In these studies, tweets were used as a weighting factor to generate inundation maps. Other existing studies are limited in 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, with this information, built up the signal of the events reported on Twitter automatically using a wavelet transformation. In this period, they detected a single flood event. Smith et al. (2017) used tweets to improve and extrapolate data from a hydraulic flood modelling. For this purpose, they used two events that occurred in the city of Newcastle. Also, Tkachenko, Jarvis, and Procter (2017) used flood-related geo-tagged messages from Flickr to detect floods in England. Huang, Wang, and Li (2018) propose a flood reconstruction model using remote sensing images with continuous flow readings and tweets. They find that social media data improves the identification of areas of high probability of flooding during a flood event.

One of the advantages of using social media for monitoring flood events is the extensive spatial coverage of the measurements. This point makes

possible obtaining useful information at different points of river catchment areas and cities, where the local inhabitants are able to supplement static sensors of hydro-meteorological networks. Social media messages can be generated at any time, the platforms are free, and some messages are accompanied by photos or videos. However, even today there are still multiple challenges that have to be faced. These include the ambiguity of the observations, which makes them difficult to interpret and validate, the frequency of the data, which may be affected by other phenomena or events, the volume of data to process that can become high and the unstructured nature of social media messages, requiring a pre-processing that sometimes hinders the speed of assimilation of information. Besides, verifying the veracity of the information in other sources or with other users is sometimes not possible, among other problems. Finding the best way to extract relevant information from social media and integrate it with data from other sources are key requirements to increase the reliability in monitoring and forecasting. An additional challenge is to ensure that these new information sources can be used to assist hydrological models to support decision-making related to early warning systems (Horita *et al.*, 2015; Mazzoleni *et al.*, 2017).

Furthermore, social media data must be treated through computational tools and data management strategies to be evaluated, transformed and validated. Statistical and machine learning methods provide tools to extract useful information from it. Most of the previous works in this area have concentrated on using social media data either for flood mapping or exploring spatio-temporal patterns (Weng and Lee, 2011; Smith *et al.*, 2017; Tkachenko, Jarvis and Procter, 2017). In previous works, a close spatial-temporal link between social media activity and flood-related events (Albuquerque, Herfort and Brenning, 2015), and rainfall was found (Andrade *et al.*, 2017).

It is necessary, therefore, to have methodologies for extracting relevant information from social media. These methodologies should allow authoritative and social media data to work together in monitoring and forecasting environments. This joint work would help at times when sensors fail or are inexistent. In addition, taking into account the great uncertainty of these data sources, the methodologies should incorporate data assimilation techniques, in such a way that the uncertainty in the modelling can be taken into account. Thereby, to solve these problems, the hypothesis of this work is that the combined use of rainfall from authoritative and social media sources into an assimilation process can reduces the uncertainty of monitoring (see Figure 2).

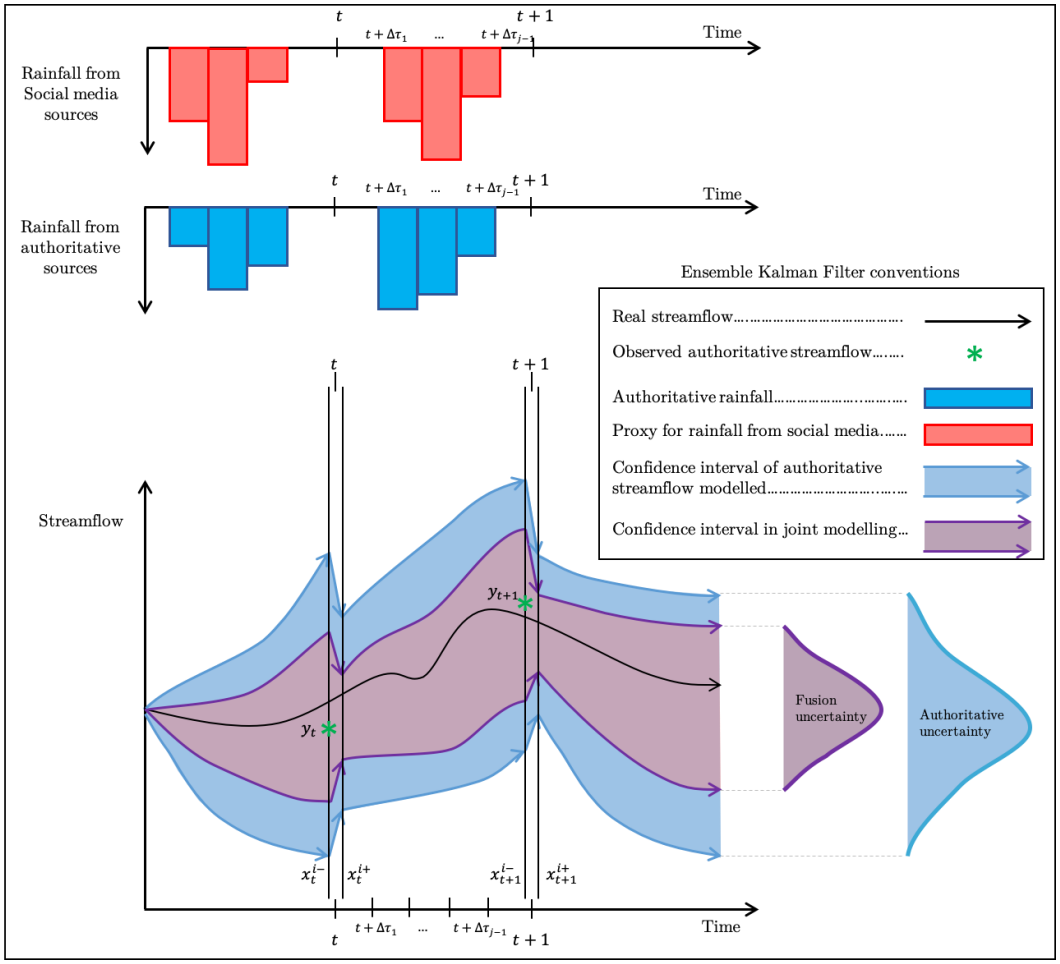


Figure 2. Hypothesis of authoritative and social media fusion-assimilation

B. Research objectives

The main objective of this thesis is to propose a methodology with the help of data assimilation techniques that allows the incorporation of alternative and complementary data sources to improve short-term streamflow monitoring in cities where sensor networks fail or are inexistent.

Specific objectives of this thesis are:

1. To propose a rainfall proxy variable from data from social media to be used in flood monitoring.
2. To implement a data assimilation technique that allows to update the state variables in each step of time and the fusion of authoritative data and social media in flood monitoring.
3. To develop guidelines to use alternative data sources in hydrology in cases of joint monitoring, for authoritative and alternative data, in “poorly gauged” or "ungauged" catchments.

C. Outline of the thesis

This thesis is divided into three chapters. A schematic brief overview of the structure of this thesis is given in Figure 3.

Chapter 1 describes the case study, the hydrological authoritative data and a brief description of the Probability Distributed Model (PDM).

Chapter 2 presents the hypothesis to create the proxy that transforms data from social media into rainfall, show a proposal for a transformation proxy function and shows how these data can be used in the monitoring of watersheds.

Chapter 3 explores the Ensemble Kalman Filter for data assimilation with authoritative and geo-social media proxy for streamflow monitoring.

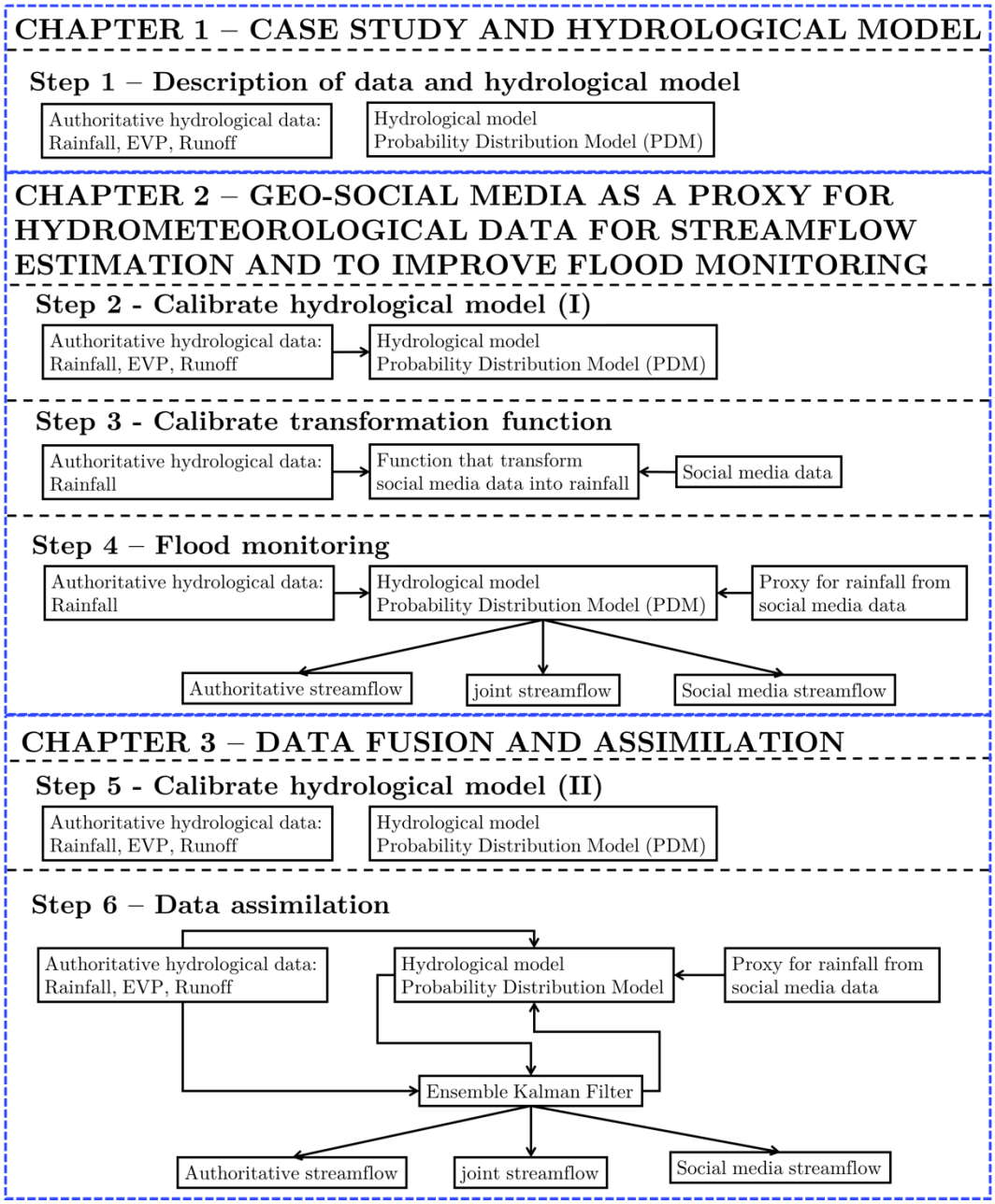


Figure 3. Graphical outline of the thesis.

1

CASE STUDY AND HYDROLOGICAL MODEL

This chapter describes the case study, the authoritative data and the hydrological model, which are going to be the same throughout the work.

1.1. Aricanduva catchment

The Aricanduva catchment (Figure 4) is located in the city of Sao Paulo, Brazil, a metropolitan region with more than 20 million inhabitants, the biggest human density in Brazil. Aricanduva is tributary of Tiete River, the main river of the city, and has a total drainage area of 100 km^2. In this study, a sub-catchment of 88 km^2 was selected, where the Sao Paulo Flood Warning System (SAISP), the organization responsible for measuring water levels, has three water level sensors, one of them close to a prone area under frequent flash floods (Figure 5). Water level sensor measurements are provided every 10 min by SAISP[2]. The precipitation data is provided also every 10 min by the National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN)[3].

The Aricanduva catchment was chosen for this study because it is one of the most important basins in Sao Paulo. The Aricanduva river basin has 22 tributaries on the outskirts of São Paulo. Frequent floods occur here during the rainy season, generating losses related to goods, human lives, infrastructure, health, and time. For instance, in January 2010, heavy rains caused landslides, which in turn originated human deaths, economic damages, and compromised the operation of public transport; in addition, water and electricity supply was interrupted, as several parts of the catchment flooded (Listo and Vieira, 2012). The watercourse slope ranges nearly from 0.025 m/m in the high section to less than 0.005 m/m in the medium and low sections (Silva, 2010). The Aricanduva catchment covers thirteen districts of the municipality of São Paulo with a population of over one million and five hundred thousand inhabitants and a population density of over 114 inhabitants/ha (Silva, 2010). The

[2] https://www.saisp.br/estaticos/sitenovo/home.xmlt

[3] http://www.cemaden.gov.br/

Aricanduva catchment is remarkable for its heavy urbanization including a strong roads system, residences and factories. Thus, the middle and lower area of the basin is already fully occupied. Although there are some green areas in the highest part of the basin, the expansion tendency of the urbanization towards the headwaters is evident, with the growth of deforestation activities. There is a highly preserved green area located on the right margin of the middle course of the catchment, corresponding to a portion of the so-called Carmo's Park, with its original forest. The Aricanduva River has several slums on the slopes of the basin, mainly in the upstream sub-basins. The bad practices of construction of houses and embankments in these areas have increased the erosion of slopes, which affects downstream strength and increases sedimentation and flooding (Listo and Vieira, 2012).

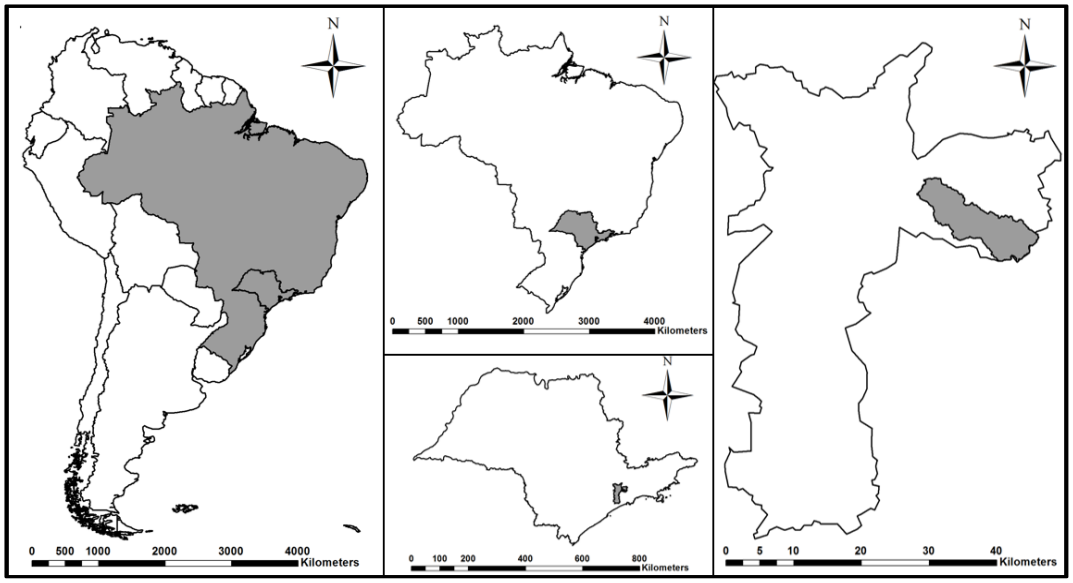


Figure 4. Aricanduva watershed, Sao Paulo Metropolitan Region, selected for this study.

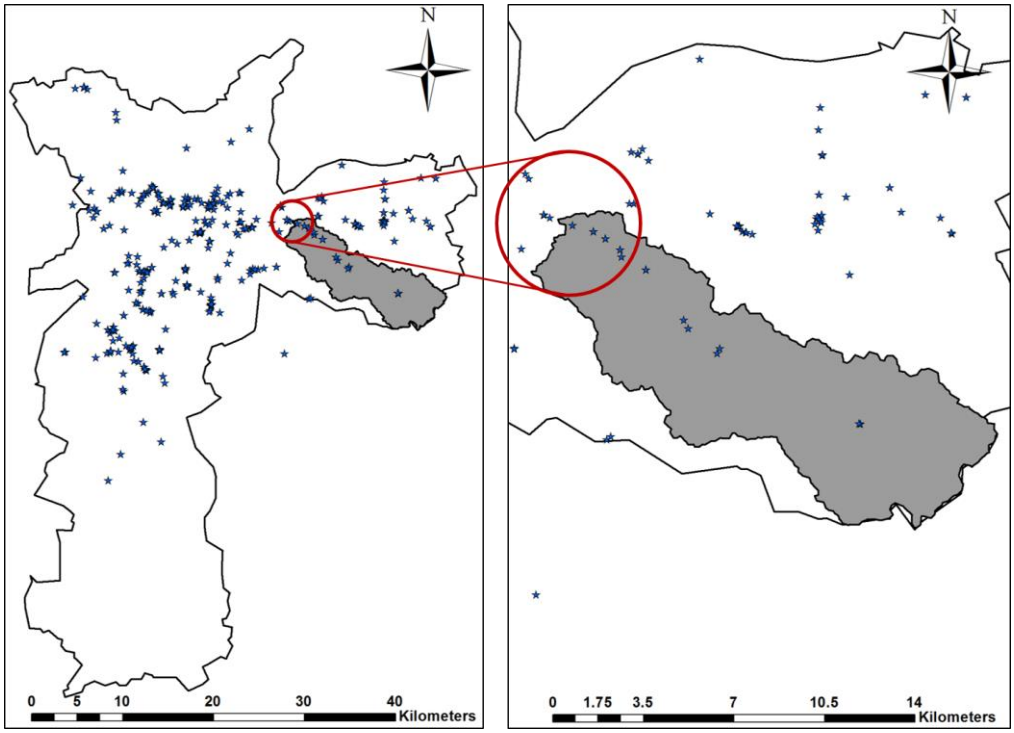


Figure 5. SAISP reported floods points November 2016-February 2017.

The time of concentration of the Aricanduva catchment is around 5.6 hours. This value was obtained by computing the average of the values in Table 1 (Velez-Upegui and Botero-Gutierrez, 2010; Almeida *et al.*, 2014; Salimi *et al.*, 2017; Michailidi *et al.*, 2018).

Table 1. Time of concentration

Model	Formula	Remark	Value (hours)
California Culvert Practice	$t_c = \left(0.87075\frac{L^3}{H}\right)^{0.385}$	Developed for small mountainous basins in California	4.3
Giandotti	$t_c = \frac{4*\sqrt{A}+1.5L}{25.3\sqrt{S*L}}$	Developed for small agricultural watersheds	5.8
Pilgrim and McDermott	$t_c = 0.76A^{0.38}$	Developed for watersheds <250 km^2	4.4
Témez	$t_c = 0.3\left(\frac{L}{S^{0.25}}\right)^{0.76}$	Developed for natural catchments in Spain	7.8

where t_c is the time of concentration in minutes, L is mainstream longitude in kilometres, A is the catchment area in square kilometres, S is the average slope of mainstream in (m/m), H elevation difference between divide and outlet in meters. For these computations the values of $L = 23.5\ [km], H = 260\ [m], A = 100\ [km^2], S = 0.011064\ \left[\frac{m}{m}\right]$.

1.2. Rainfall authoritative

Rainfall data were collected by CEMADEN using an API Application. The data is updated in intervals of 10 min if the cumulative volume in this interval is higher than 0.2 mm. However, if no rainfall is recorded, the data are available hourly. Then, since our modelling is aimed at providing a tool to predict floods, rainfall-runoff calibration is done for some previous events, which have total precipitation greater than 10 mm. For this, 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 other 15 were chosen for validation (from 2016-01-01 to 2016-01-30 and 2016-11-09 to 2017-02-27). Quality and consistency of the rain gauge available information were assessed by comparing them with gathered information by the University of São Paulo (USP), São Paulo, and its observatory, which provides the information of monthly amounts of rainfall[4]. This information allowed us to validate the accumulated magnitudes of rainfall stations. Three rain gauges that showed coherent values for both sources were used.

1.3. Probability Distributed Model (PDM)

The Probability Distributed Model (PDM) and similar models derived from it are conceptual rainfall-runoff models widely used in research and hydrological applications (Alvarez-Garreton *et al.*, 2014), such as parameter prediction updating, flood forecasting, regionalization of parameters using the Ensemble Kalman Filter (EnKF), among others (Lamb, 1999; Moradkhani *et al.*, 2005; Kay *et al.*, 2009). The PDM transforms rainfall and potential evapotranspiration series over a catchment into streamflow at the outlet of the catchment. A detailed description of the modelling process, parameters and model formulation is given by Moore (2007). In this work, the PDM has been chosen over distributed and physically-based hydrological models because it requires a reasonable number of hydrometeorological variables (i.e. rainfall,

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

potential evapotranspiration and streamflow), and it is lumped, parsimonious and friendly-user model. These last properties reduce the modelling time. In comparison, distributed and physically-based hydrological models have high computational requirements for simulating spatiotemporal processes in multiple control sections through nonlinear equations.

Some other advantages of the model are that it is a lumped model that allows a fast estimation of the total flow, and the base flow. It has few requirements for input variables, only precipitation and potential evapotranspiration. It can be implemented with relative ease in watersheds that have little data, being a simple model that allows integrating data assimilation methodologies. However, this model does not include the spatial distribution of the catchment, since it is a concentrated model and it considers the entire basin as a single reservoir, directly leaving out of the consideration important variables for the movement of water, such as solar radiation, type and distribution of soils and their uses, among others.

Figure 7 shows how the PDM model works and how its various components are connected. In this schematic diagram, three principal components can be identified. A reservoir of probabilistically distributed soil moisture, a surface reservoir and a sub-surface reservoir. The reservoirs within the model are conceived as if they were random variables. There are different proposals about distribution functions used as shown in detail in Moore, (2007). The version used in this work is that of the Pareto distribution (Moore, 2007), in which the distribution function of the store capacity (c) is given by

$$F(c) = 1 - \left(\frac{c_{max} - c}{c_{max} - c_{min}}\right)^{b} \tag{1.1}$$

where c_{min} and c_{max} are the maximum and minimum values of the storage capacity of the model, and b is a parameter belonging to the Pareto distribution that represents the degree of control of spatial variability over the catchment. In this sense, it should be noted that $b = 1$ will represent a triangular distribution and $b = 0$ will represent a constant storage capacity in the basin. Figure 6 shows these relations.

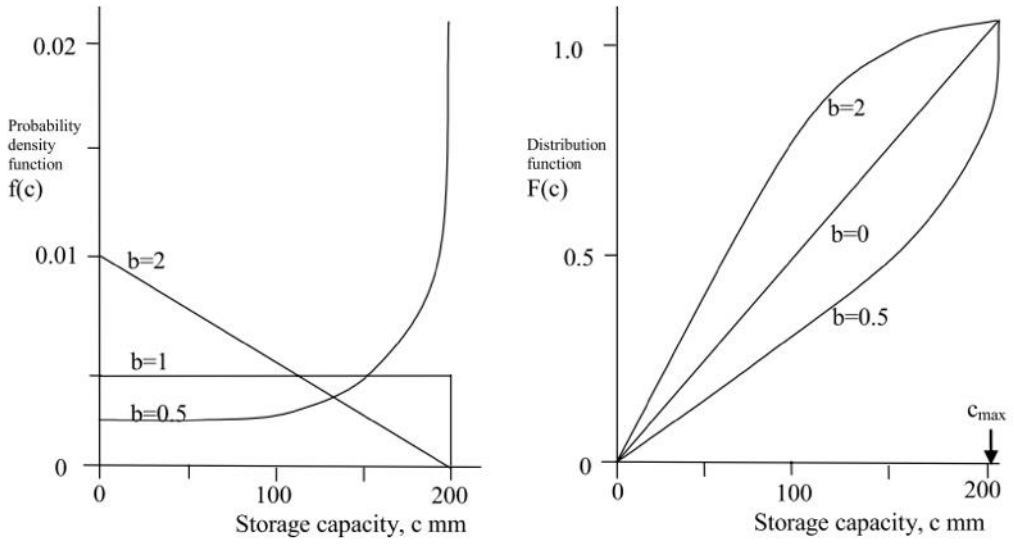


Figure 6. Pareto distribution of storage capacity. Source: Moore (2007)

Deriving this function, the density function of the storage capacity is obtained.

$$\frac{dF(c)}{dc} = f(c) = \frac{b}{c_{max} - c_{min}} \left(\frac{c_{max} - c}{c_{max} - c_{min}} \right)^{b-1} \tag{1.2}$$

The total available storage S_{max} is equal to the mean storage capacity $\bar{c}$

$$\bar{c} = S_{max} = \int_0^{\infty} \big(1 - F(c)\big)\, dc = \frac{b c_{min} + c_{max}}{b + 1} \tag{1.3}$$

the total water in storage over the catchment, $S(t)$, is

$$S(t) = c_{min} + (\bar{c} - c_{min}) \left[1 - \left(\frac{c_{max} - c^*(t)}{c_{max} - c_{min}} \right)^{b+1} \right] \tag{1.4}$$

where the critical storage capacity, $c^*(t)$, is

$$c^*(t) = c_{min} + (c_{max} - c_{min}) \left[1 - \left(\frac{S_{max} - S(t)}{\bar{c} - c_{min}} \right)^{\frac{1}{b+1}} \right] \tag{1.5}$$

The dependence of evapotranspiration loss on soil moisture content is introduced by assuming the following function between the ratio of actual to potential evapotranspiration

$$\frac{E_i'}{E_i} = 1 - \left\{ \frac{S_{max} - S(t)}{S_{max}} \right\}^{b_e} \tag{1.6}$$

where b_e is the exponent of the evapotranspiration function.

The rate of drainage over the interval, d_i, depends linearly on catchment moisture at the start of the interval

$$d_i = k_g^{-1}[S(t) - S_t] \tag{1.7}$$

The parameter k_g is a temporary constant of drainage of groundwater and S_t it the threshold storage below witch there is not drainage, water held under soil tension.

And in this manner

$$\pi_i = P_i - E_i' - d_i \tag{1.8}$$

The change of the volume in the time depends of the net rainfall π_i

$$V(t + \Delta t) = \pi_i \Delta t - [S(t + \Delta t) - S(t)] \tag{1.9}$$

The change in the volume of water held in the storage per unit area is

$$S(t + \Delta t) - S(t) = -\frac{1}{3k_b S^2(t)}\{\exp[-3k_b S^2(t)\Delta t] - 1\}[u - k_b S^3(t)] \tag{1.10}$$

where u is the rate of inflow to the store per unit area and k_b is a time constant of base flow.

The representation of the surface storage component is a cascade of two linear reservoirs, with time constants k_1 and k_2, expressed as the discretely coincident transfer function model (Moore, 2007):

$$q_t = -\delta_1 q_{t-1} - \delta_2 q_{t-2} + \omega_0 u_t + \omega_1 u_{t-1} \tag{1.11}$$

with

$$\delta_1 = -(\delta_1^* + \delta_2^*), \qquad \delta_2 = \delta_1^* \delta_2^*, \qquad \delta_1^* = \exp\left(-\frac{\Delta t}{k_1}\right), \qquad \delta_2^* = \exp\left(-\frac{\Delta t}{k_2}\right)$$

$$\omega_0 = \begin{cases} \dfrac{k_1(\delta_1^* - 1) - k_2(\delta_2^* - 1)}{k_2 - k_1} & k_1 \neq k_2 \\ 1 - \left(1 + \dfrac{\Delta t}{k_1}\right)\delta_1^* & k_1 = k_2 \end{cases} \tag{1.12}$$

$$\omega_1 = \begin{cases} \dfrac{k_2(\delta_2^* - 1)\delta_1^* - k_1(\delta_1^* - 1)\delta_2^*}{k_2 - k_1} & k_1 \neq k_2 \\ \left(\delta_1^* - 1 + \dfrac{\Delta t}{k_1}\right)\delta_1^* & k_1 = k_2 \end{cases} \tag{1.13}$$

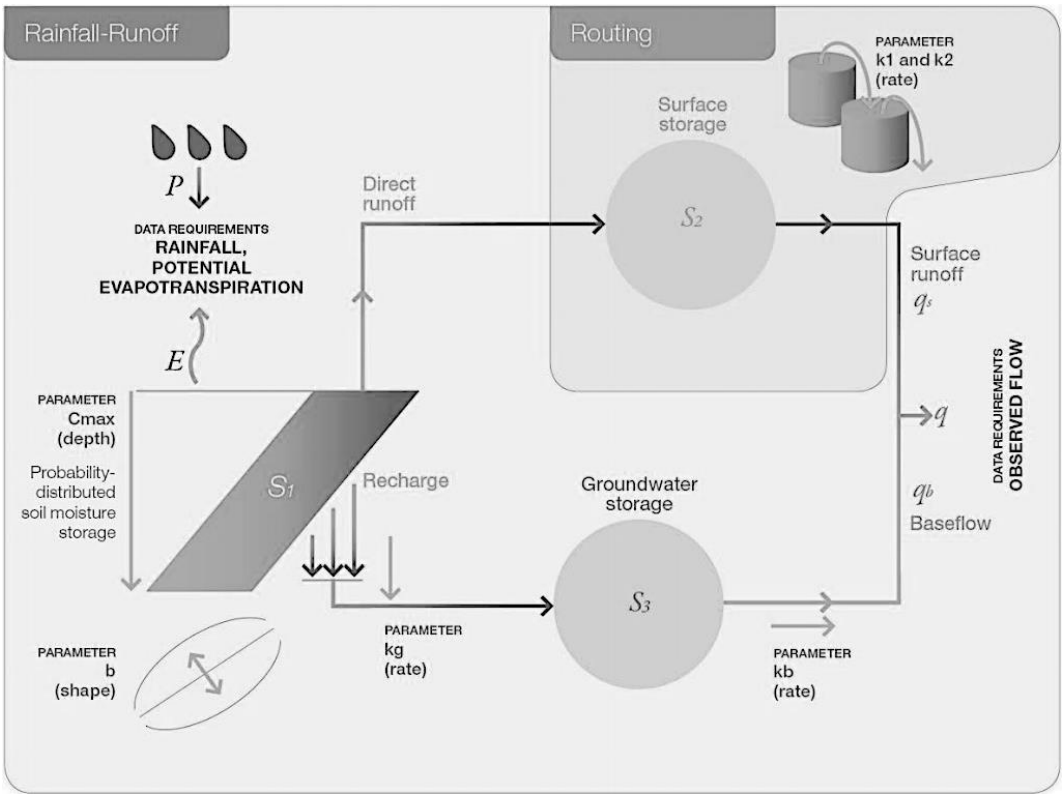


Figure 7. Scheme of PDM hydrological model. Source: Pagano, Hapuarachchi, and Wang (2009).

It is important to emphasize that in this thesis two versions of the PDM were used, one to calibrate and another to update states. This last version was based on the correction of errors using the values of the state variables at the output, using the approach proposed by Moore (2007). The PDM parameters are described in

Table 2.

Table 2. PDM parameters

Parameter	Description	Function	Units
C_{min}	Baseflow	Minimum storage capacity	(mm)
C_{max}	Baseflow	Maximum storage capacity	(mm)
b	Surface flow	Exponent of Pareto distribution controlling spatial variability of storage capacity	(-)
b_e	Baseflow	Exponent of actual evaporation function	(-)
k_1	Surface flow	Time constant for linear reservoir	(h)
k_2	Surface flow	Time constant for linear reservoir	(h)
k_b	Baseflow	Baseflow time constant	(h mm^{-1})

k_g	Baseflow	Groundwater recharge time constant	(h mm^{-1})

1.4. NSE and Nash log

It is always necessary to calibrate and validate the hydrological models. These calibrations consist of varying the parameters of the model while using input values of rainfall or evapotranspiration, among others, in the model. The outputs of the model are compared with the actual values using some statistics as indicators of how good the calibration is. The parameters are adjusted until reaching adequate values for the statistics. In this thesis, two of the most important statistics were used: The Nash-Sutcliffe Efficiency (NSE) and the Nash logarithm. The Nash-Sutcliffe Efficiency (NSE) and Nash-log are normalized statistics that determine the relative magnitude of the residual variance ("noise") compared to the measured data variance ("information") (Nash and Sutcliffe, 1970). In the case of the NSE, these calculations are done with both the observed streamflow and the simulated streamflow. In the Nash-log, the calculations are done with their logarithms. NSE and Nash-log are computed as shown in equations (1.14) and (1.15):

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}\left(Q_{sim,i} - Q_{obs,i}\right)^2}{\sum_{i=1}^{N}\left(Q_{obs,i} - \frac{1}{N}\sum_{i=1}^{N} Q_{obs,i}\right)^2} \tag{1.14}$$

$$\mathrm{Nash-log} = 1 - \frac{\sum_{i=1}^{N}\left(\log\left(Q_{sim,i}\right) - \log\left(Q_{obs,i}\right)\right)^2}{\sum_{i=1}^{N}\left(\log\left(Q_{obs,i}\right) - \frac{1}{N}\sum_{i=1}^{N}\log\left(Q_{obs,i}\right)\right)^2} \tag{1.15}$$

where $Q_{sim,i}$ is the model simulation discharge, $Q_{obs,i}$ observed discharge, N is the total number for simulation periods .

2

GEO-SOCIAL MEDIA AS A PROXY FOR HYDROMETEOROLOGICAL DATA FOR STREAMFLOW ESTIMATION AND TO IMPROVE FLOOD MONITORING

This chapter is based on the article "Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring" published in the journal Computers & Geosciences in 2018.

2.1. Geo-social media

Social media data used in this study were gathered from the Twitter platform using the public streaming Application Programming Interface (API) to fetch georeferenced tweets within two bounding box that encompasses the city of Sao Paulo[5]. Twitter provides data through a JSON file. Once extracted it is stored in a database. The JSON file brings in addition to the message text, the user, if the message is geolocated or not, data relative to the user who creates it as for example number of followers, number of users to whom it follows, language in which the user was written the tweet, among others. Data coming from social media are unstructured data that need to be processed before using them. To extract the signal coming from the social media used in this thesis, the message shared in the tweet was used.

The total number of tweets collected was 15,883,710(~28 Gigabytes). The georeferenced tweets (1,631,329, ~2.8 Gigabytes) (Figure 8) were then filtered by keywords (21,804). From the 1st to 30th January 2016 and from 8th November 2016, to 28th February 2016, were found 6,651 geotagged tweets related to floods within the city of Sao Paulo. Similar to our previous study (Andrade *et al.*, 2017), the messages were filtered in terms of words related to rain (*chuva* in Portuguese), intense rainfall and rainbow, excluding common unrelated expressions (Figure 9). Some examples for related tweets can be found in Table 3. Figure 10 shows the

[5] Crawler-twitter tool is available at https://github.com/sidgleyandrade/crawler-twitter

spatial distribution of the rainfall-related tweets in the city of Sao Paulo during this period.

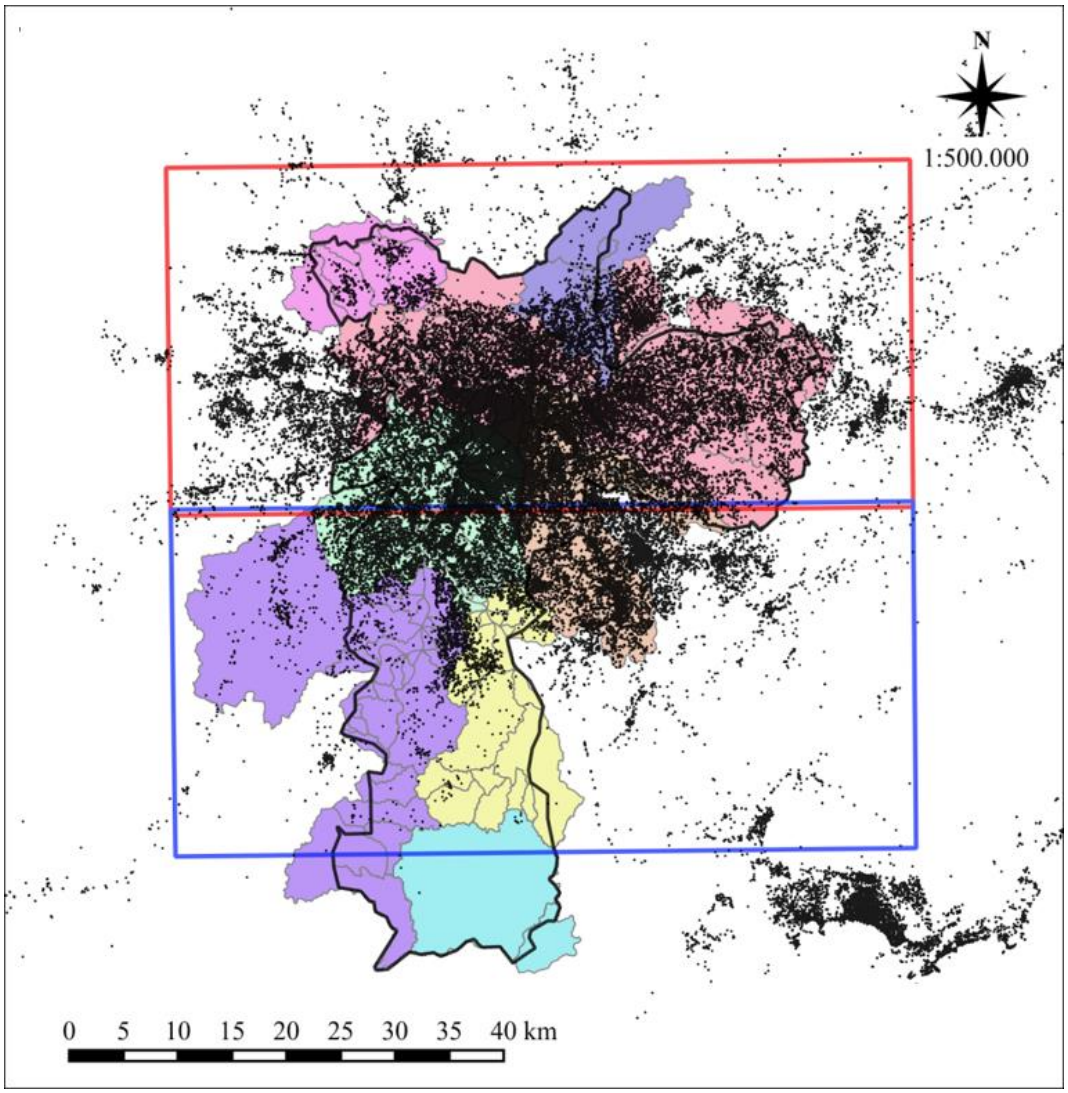


Figure 8. Georeferenced tweets for the analysed period.

In addition, the signal coming from social media could be generated through several methodologies, using frequency, entropy of information, through densities, among others. In this work, it was decided to use the absolute frequency of the keywords, removing those that generate noise. This with the aim of doing a job as simple as possible and be able to observe the characteristics of the social media data mentioned.

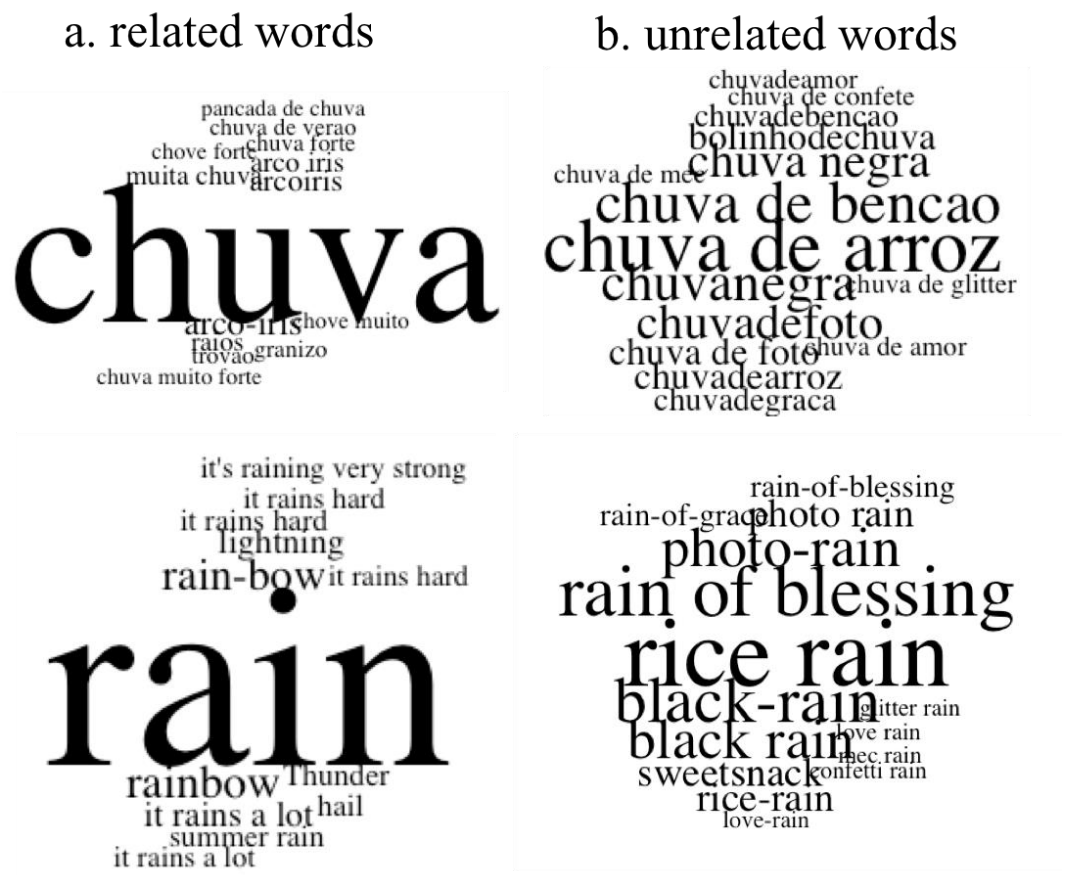


Figure 9. Frequent-related and unrelated words. All keywords are in unicode standard.

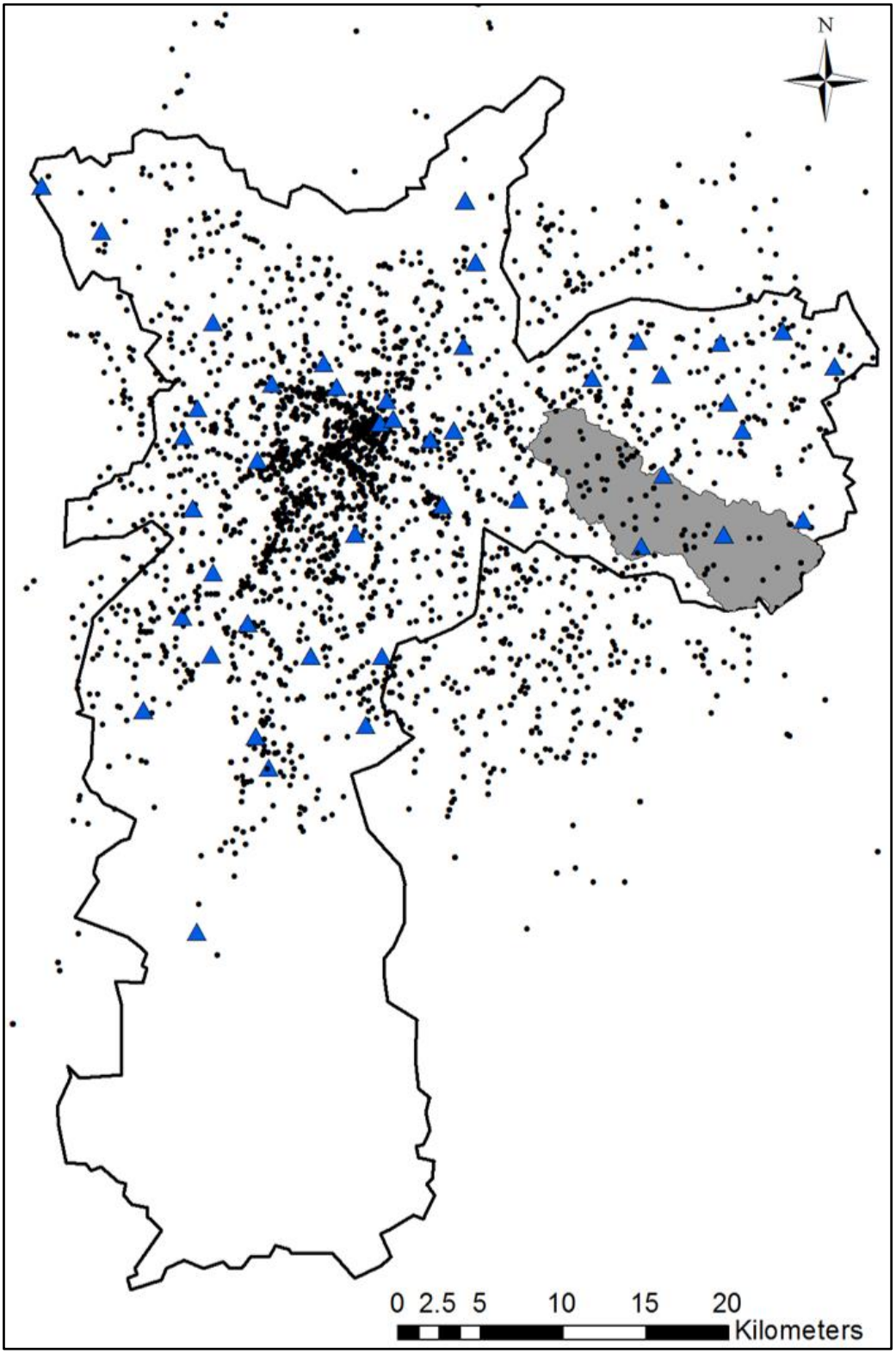


Figure 10. City of Sao Paulo with related tweets as black points, rainfall gauges as blue triangles and the Aricanduva catchment shaded grey, for the analysed period.

Geo-located tweets containing the selected 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 accumulated frequencies every Δt min. In section 3.3.2, it will discuss how this signal can be transformed into a proxy variable for rain.

Table 3. Some related tweets 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 taking a coke (@Hooters) ¦ https://t.co/KEFYXy8YM4"
2016-12-03 21:43:25	"Início da noite de 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"

2.2. Exploratory data analysis

An initial exploratory data analysis is displayed in Figure 11, in which the frequency series of two time-series are summarized. One is related to relevant words or, Twitter phrases related to, rainfall processes. The other one corresponds to rainfall depths measured by the authoritative sensors. Evidences by plotting the two-time series reveal time significant relationship of frequency of tweets and rainfall depths.

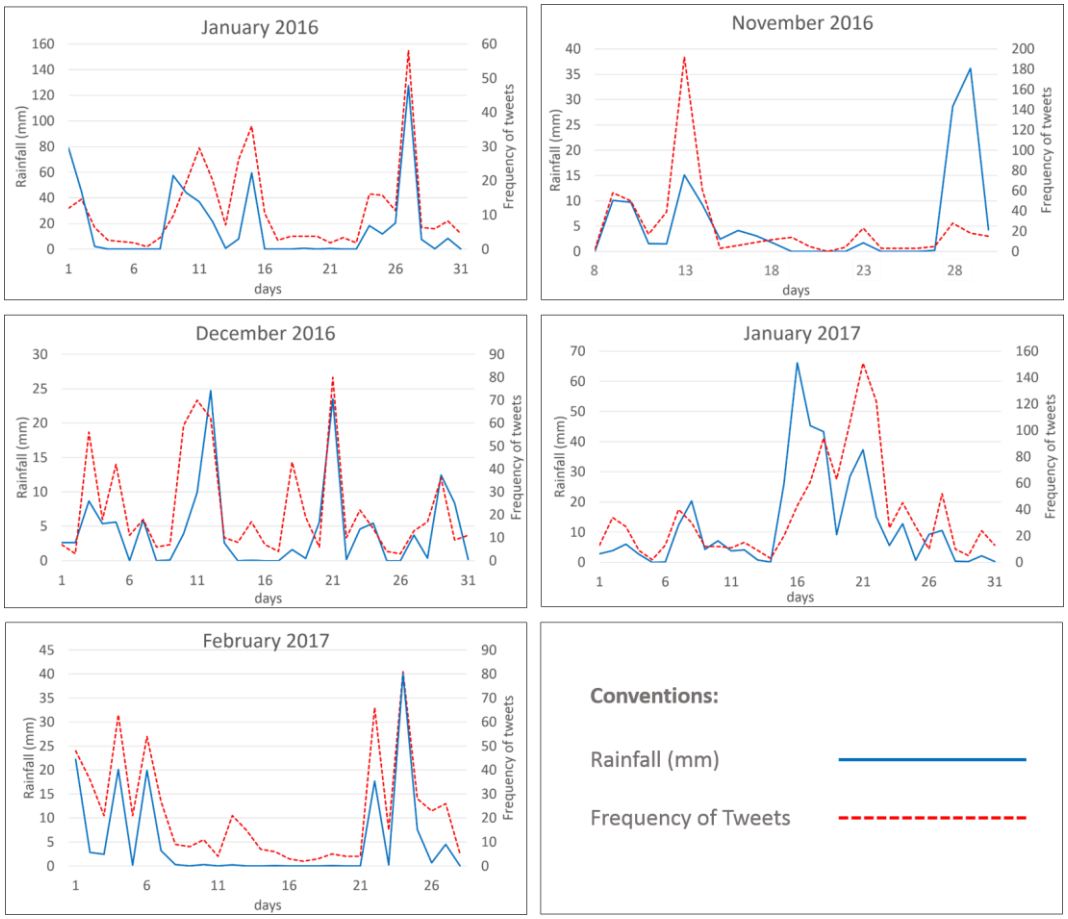


Figure 11. Time series of rainfall depths (left) with frequency of tweets (right) for the period of study January, 2016 (upper left chart) and since November 8, 2016 to February 28, 2017 (other charts, respectively).

In some events, the two series did not fit with the same behaviour or relative magnitude, as it is shown in Figure 11. For instance, on November 12th, 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 of this band. This reaction seems to be amplified by

the fact that it was interpreted while it was raining in the city. One example of how false positives may occur in detections could be illustrated by this tweet, as "*luizh.ap:* November Rain *com direito a chuva e balões vermelhos #GunsNRoses #gunsnrosesreunion #Axl #Slash #Duff #GNR*" translation: "November Rain entitled to rain and red balloons!!". !!". Those constraints call for methodology for refining geotagged data related to rainfall, explained as follows.

2.3. Methodology

Figure 12 displays the methodological structure adopted to transform data from social media into a hydrometeorological proxy variable and the way this can be used to provide further flood early warnings. The methodology is divided in four stages: (a) hydrological data (calibration and rainfall-streamflow modelling) (b) social media data (fitting the transformation function of proxy); (c) social media data (transformation of social media signal into hydrometeorological data) (d) comparison with real data. In each step a series of activities are carried out. Each of these processes is in turn explained in next sections.

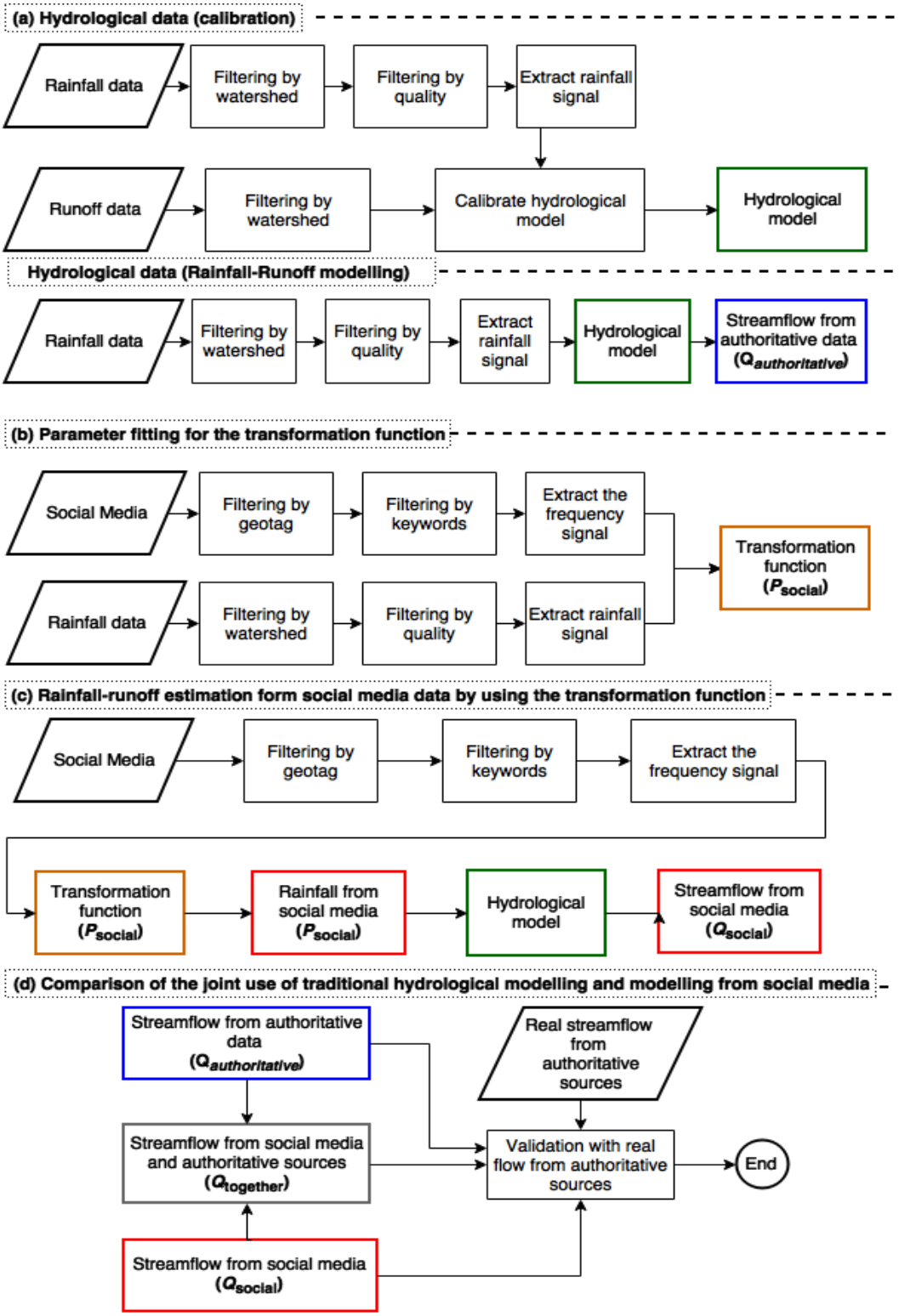


Figure 12. Methodological structure to transform authoritative and social media information to improve flood monitoring.

2.3.1. Hydrological data

The first methodological procedure carried out was the calibration of the hydrological model to be used in order to obtain a transformation of authoritative and social media rainfall values into a streamflow. This is a classic process in hydrology in which some hydrometeorological variables such as rainfall and streamflow are used to calibrate the model (Muleta, 2012). Considering that the methodology is intended to be used in ungauged and poorly gauged catchments or subject to sensor failures, simple modelling seems to be more appropriate (Sivapalan *et al.*, 2003).

In this work, the PDM has been calibrated and validated with time-steps of 10 min, considering the available 10-min rainfall data and the rapid response time, (ca. 30min) of the studied catchment. Using 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 (Nash and Sutcliffe, 1970; Muleta, 2012). Details of the model parameters were already described in Moore, 2007.

The streamflow was calculated by using not only three individual rainfall observed from gauges of CEMADEN official network, but also other two approximations: the maximum inter-station rainfall depth every 10 min, and the spatially-estimated mean precipitation depth using the Inverse Distance Weighting (IDW) method. Table 4 resumes NSE values for the five calibration and validation of the PDM model.

Transformation of authoritative rainfall data in streamflow depends on the calibration performed. In this case, rainfall from authoritative gauges is used to model the streamflow in the same period of social media harvesting. Simulated streamflow will be later compared with the one coming from 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.

Table 4. NSE performance to consider sensors.

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

2.3.2. Parameter fitting for the transformation function

It is clear that people are not obliged to post in social media about anything, that is, each user writes different things despite being in the same place subject to the same stimuli of the environment. So, what motivates an individual to post about a particular topic or phenomenon? and once motivated, what words do they use to express himself? To create the transformation function, three properties from people's behaviour in social media were assumed: Proportionality, randomness and semantic singularity. In Figure 13 can observes how these three aspects are related to the intensity of the phenomenon according to our work hypothesis.

First, it is supposed that people use more social media when discussing a phenomenon of great relevance. In that case, the number of people talking about it will depend on how they were affected and thus, the intensity of the phenomenon could be directly proportional to the number of related tweets. In this case, the intensity is referred to the phenomenon itself, i.e. with respect to a phenomenon to which people have usually exposed, one higher intensity on average will make people more motivated to write about it in social media.

Second, people would not "speak" in a synchronous way, namely, the users randomly post messages, before, during or after the phenomenon happens (Andrade *et al.*, 2017). It has then that the persistence of a topic in social media is related to the intensity of the phenomenon. This behaviour can be measured using bins of cumulative tweets over a certain period of time, depending on the phenomenon duration.

Third, people tend to use related words when the phenomenon becomes more intense/weak or singular/unusual, which could generate semantic singularities. What differentiates the phenomenon is the beauty or catastrophic that it may be. That is, people subjected to events that generate stress due to events such as floods, or a special light during thunderstorms, like lightning strokes among others will be more willing to share photos or texts about what is happening. Further, other hydrometeorological phenomena could be incorporated into the tweets because phenomenon's beauty makes people talk more about them, increasing posting, with phrases, photos or videos, like a rainbow, close to the end of the main rainfall storm.

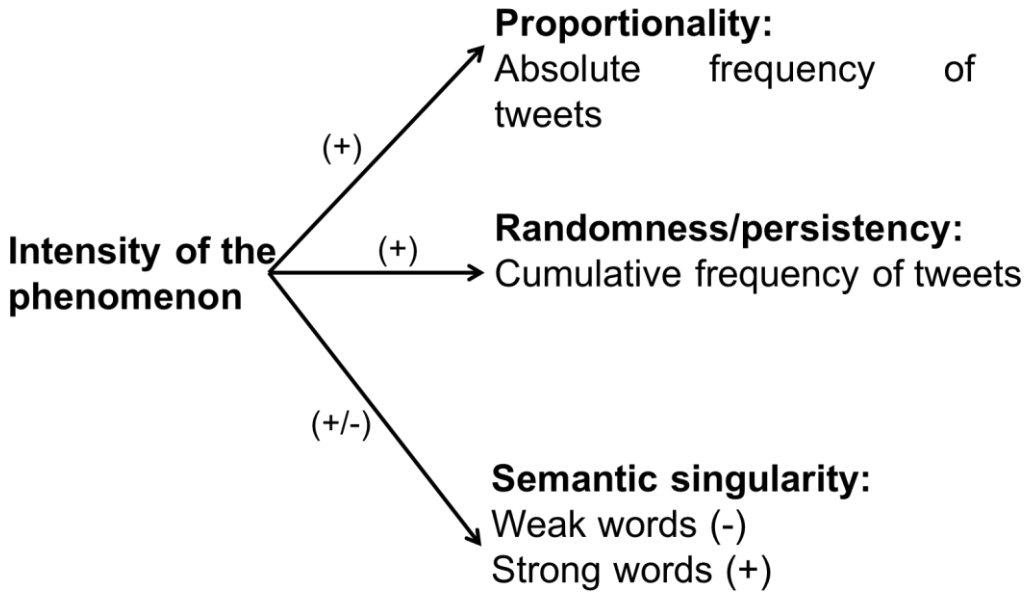


Figure 13. Hypothesis about people post in social media.

Inside this framework, a linear regression model is proposed 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 (Eq. 2.1), with the following functional structure:

$$p_{social} = \alpha\left(1 + \eta_{strong} + \eta_{soft}\right)\frac{f_{kw}}{A_{interest}} + \sum_{i=2}^{n} \beta_i \frac{F_{kw(i)}}{A_{interest}} \quad (2.1)$$

where p_{social} is the proxy of the variable precipitation that comes from the transformation from 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 a cumulative of i periods (with i = {20, 30, 40, …} minutes). $A_{interest}$

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, which have words that reinforce or diminish the intensity of the rainfall respectively. An example of a multiplicative strong effect is heavy rain, whereas for the weak effect it assumes the word rainbow.

For fitting the transformation function, the system collects social media data using an API. Then, the messages are filtered by geotags 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, referred to the whole city. The 5-fold calibration-validation procedure draws on the five-months sample, where one month is removed, while the remaining four permit fitting the transformation function. Then, the month removed is used for later validation of the transformation function of the very month. This will help to reduce any bias in the resulting function. These steps are repeated to obtain a transformation function for each month removed.

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

For the transformation of the social media data into a rainfall proxy, data was collected inside the catchment in order to obtain a rainfall proxy for this place. The same variables were collected with the same temporal resolution considered in the Section 1.2. Once the tweets were collected, the frequencies of the tweets were replaced into the function created in the past section. However, because hydrological processes like rainfall-runoff are only possible at systems like catchments, whose boundaries not necessarily agree with the administrative boundaries of the city, a regionalization of tweets into a catchment-area is performed by dividing the frequencies of related tweets every 10 min by the drainage area of the catchment. Then, this process differs from the parameter fitting process where the area of the whole city is employed. Finally, estimated rainfall values were used as an input of the PDM hydrological model to generate the streamflow data.

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

This step consists of comparing streamflow real values (SAISP) with estimated values by using either rainfall data from social media messages (Twitter), calculated from the transformation function developed in section 2.3.2, or authoritative rainfall data (CEMADEN) as input to the PDM model. This comparison is made by determining if streamflow real values are found within the confidence interval of the models, or instead they are overestimated/underestimated. This assessment allows to establish the accuracy of these cases when the modelling is performed by only using social media data, and employing the transformation function to estimate rainfall values for "ungauged" catchments, *i.e.* when it does not count on authoritative sensors. Additionally, the case when results from both models are employed was analysed, by selecting the maximum and minimum values of the confidence interval of each model and evaluating the accuracy to predict streamflow real values. This scenario is equivalent to the case of "poorly gauged" catchments, where data from both sources is available but authoritative data are inaccurate and/or imprecise.

2.4. Results

To create the transformation functions for each month, several linear regression models robust to heteroscedasticity were estimated (see Table 5). Following the 5-fold cross validation procedure, each column resumes 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 coefficients decrease in magnitude.

From these results, some simulations were carried out within the Aricanduva catchment using related tweets and authoritative rainfall data; replacing into the PDM rainfall-runoff model. For Figure 14 to 18 the convention is that in shaded blue with authoritative data and in red with social media proxy and observed streamflows in bold line. In the upper are the rainfall proxy from social media and authoritative rainfall.

Table 5. Regression coefficients for parameter fitting of 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^2	0.283	0.294	0.220	0.257	0.255

Figure 14 shows the period from January, 25th to January, 31st, 2016. It can be seen that for the rainfall events of January 26th and 28th, the proxy variable from Twitter performs better than the one with authoritative rainfall data. However, for the period after January 29th, the behaviour of the variable generated from social media strongly overestimated the streamflow values.

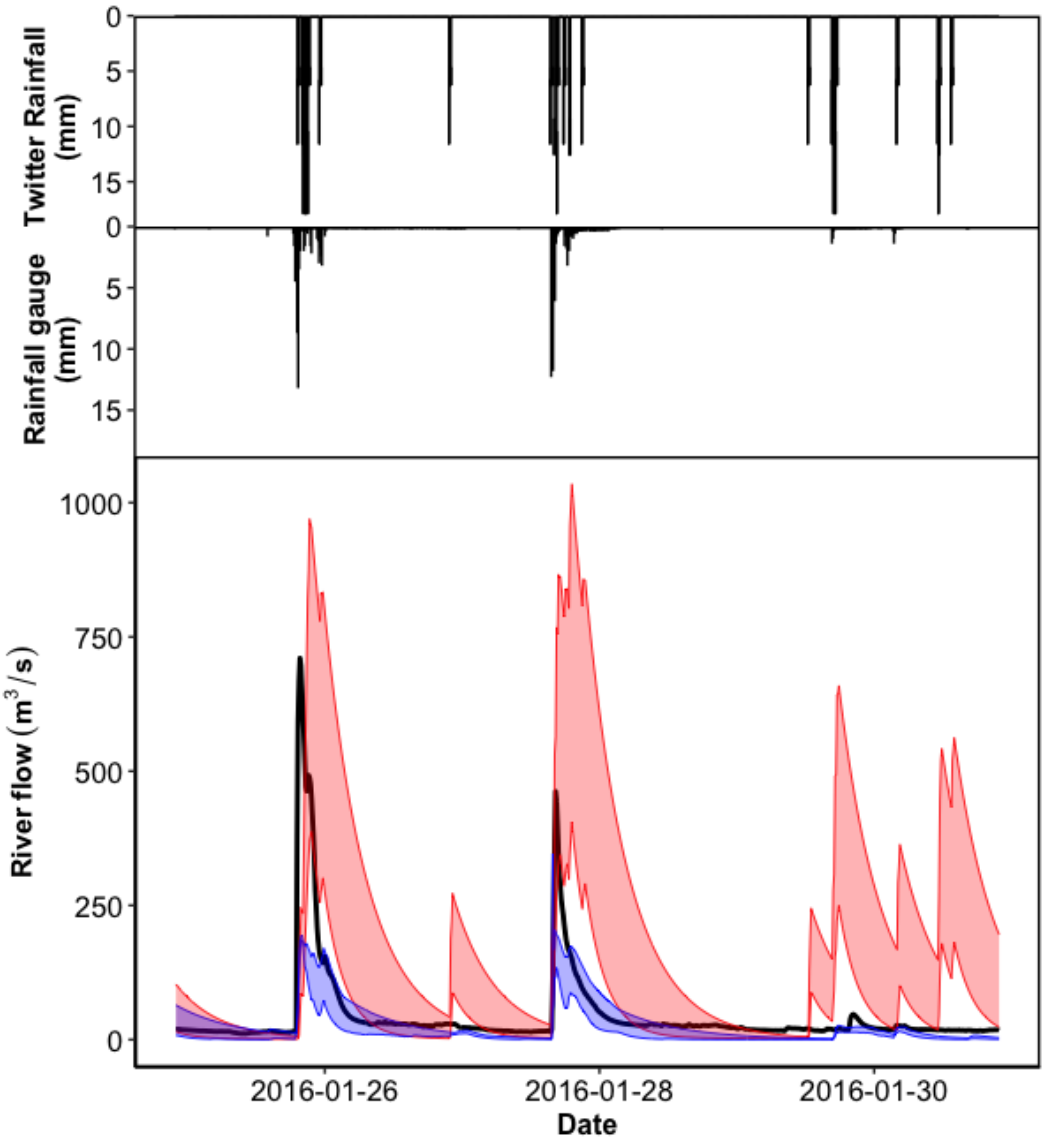


Figure 14. Streamflow simulation to the period January, 25th to January, 31st, 2016.

In turn, Figure 15, it is observe that on December 10th, there is a peak in the simulation carried out from the social media proxy, which did not exist in either the real value or the authoritative model. By the end of December 10th until December 12nd, it was observed that only the model with authoritative data followed the streamflow behaviour. However, for the highest peak streamflow, the one above 200 m^3/s, none of them offered an adequate estimative.

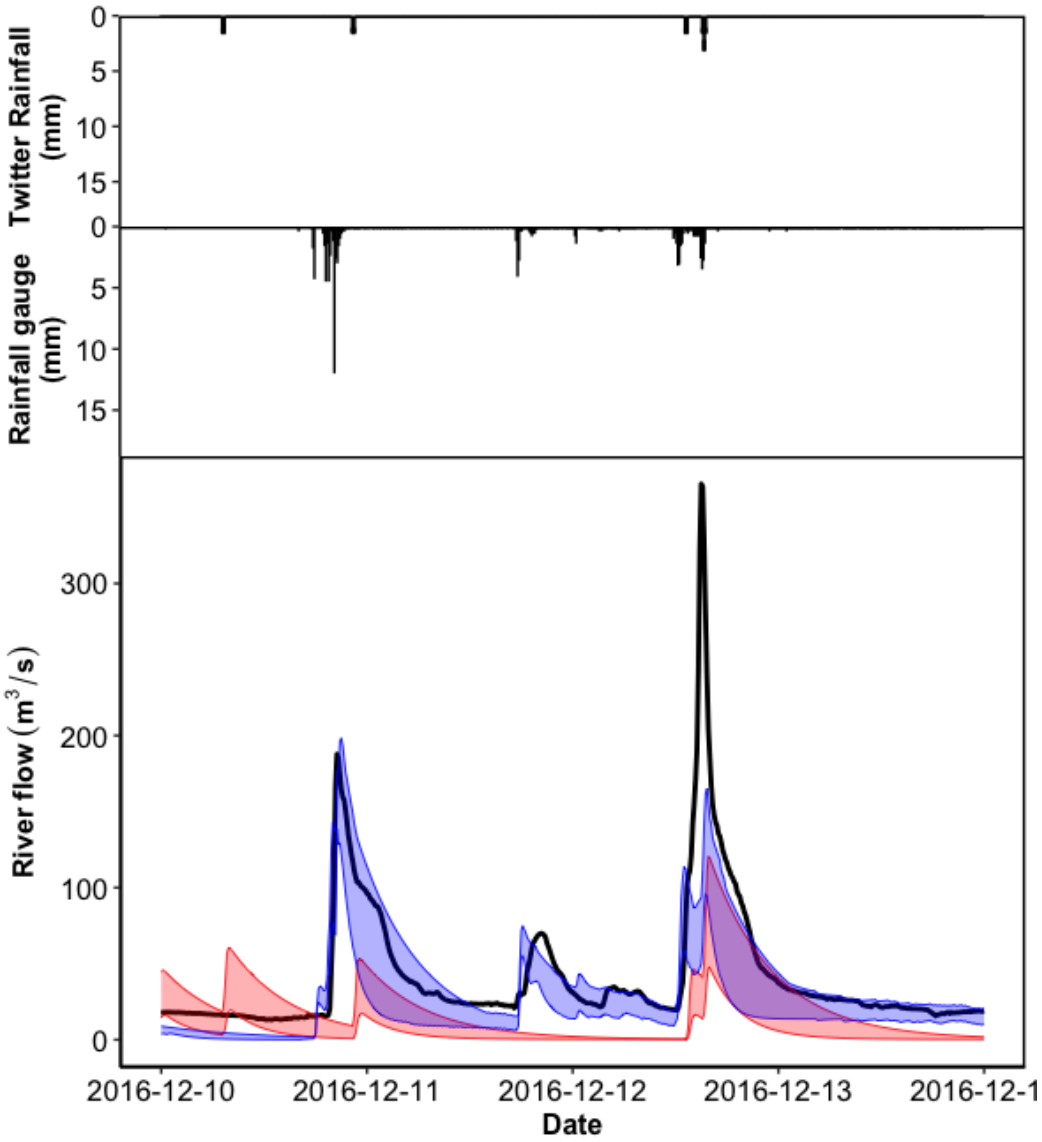


Figure 15. Streamflow simulation to the period December, 10th to 13nd, 2016.

Moreover, for the period from January 20th to 28th, 2017, Figure 16 shows how the Twitter proxy variable reacted to all observed peaks of the time series. Only in some cases, such as on January 25th, this reaction happened after the flood occurrence, except on January 26th, when the geo-social media reacted a bit earlier. Alternatively, the streamflow estimated with only the authoritative through modelling performed in a suitable way.

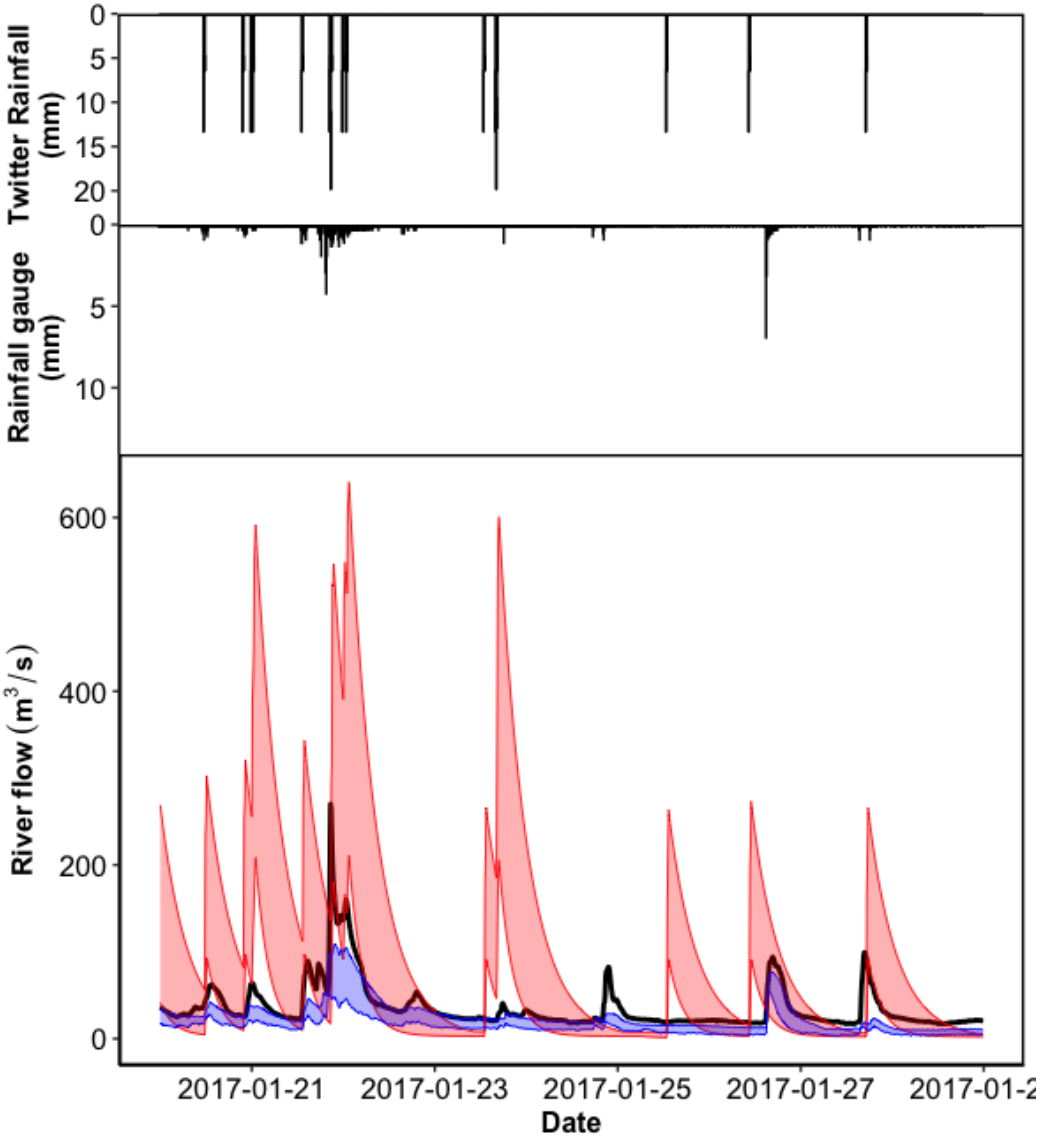


Figure 16. Streamflow simulation to the period January, 20th to 28th, 2017.

For the period from February 1st to February 9th, 2017 (Figure 17), it was observed that both simulations, carried out with the social media proxy and with authoritative data, follow the streamflow behaviour. However, for the first peak of streamflow, above 200 m^3/s, the authoritative model did not perform well; on the contrary, social media-based model reacted late, but with an adequate magnitude. Also, by the end of February 6th until February 7th, the model based on the social media reacted better.

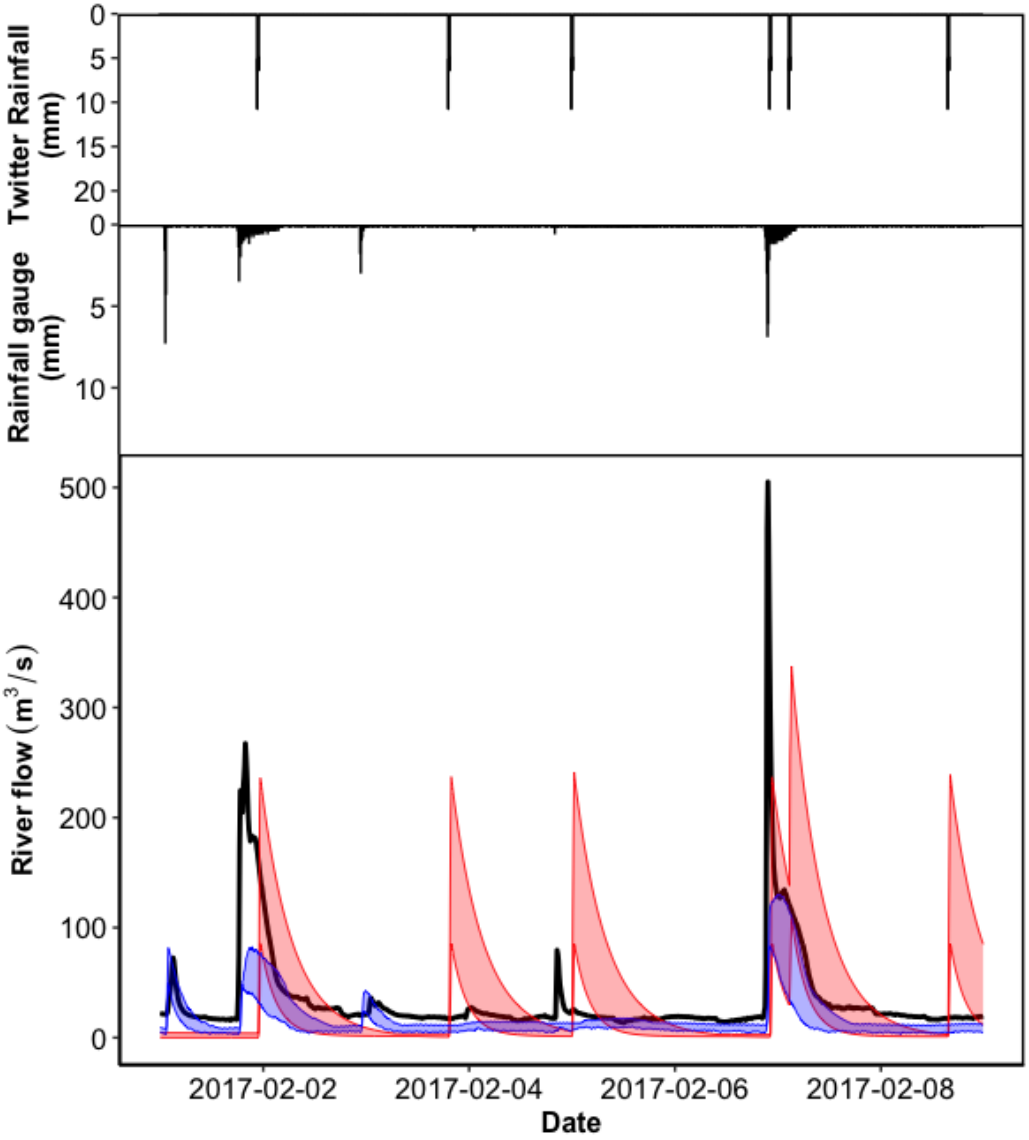


Figure 17. Streamflow simulation to the period February 1st to 9th, 2017.

In Figure 18, 5 peaks close to 100 m^3/s for the period from February 22nd to February 28th, 2017 can be observed that sometimes authorized data performs better while other times social media proxy data does it. Data on February 25th was better captured by the social media streamflow proxy, with a peak streamflow with a value greater than 700 m^3/s. This behaviour is probably due to convective rainfall, concentrated in some part of the catchment far away from the available rainfall gauges.

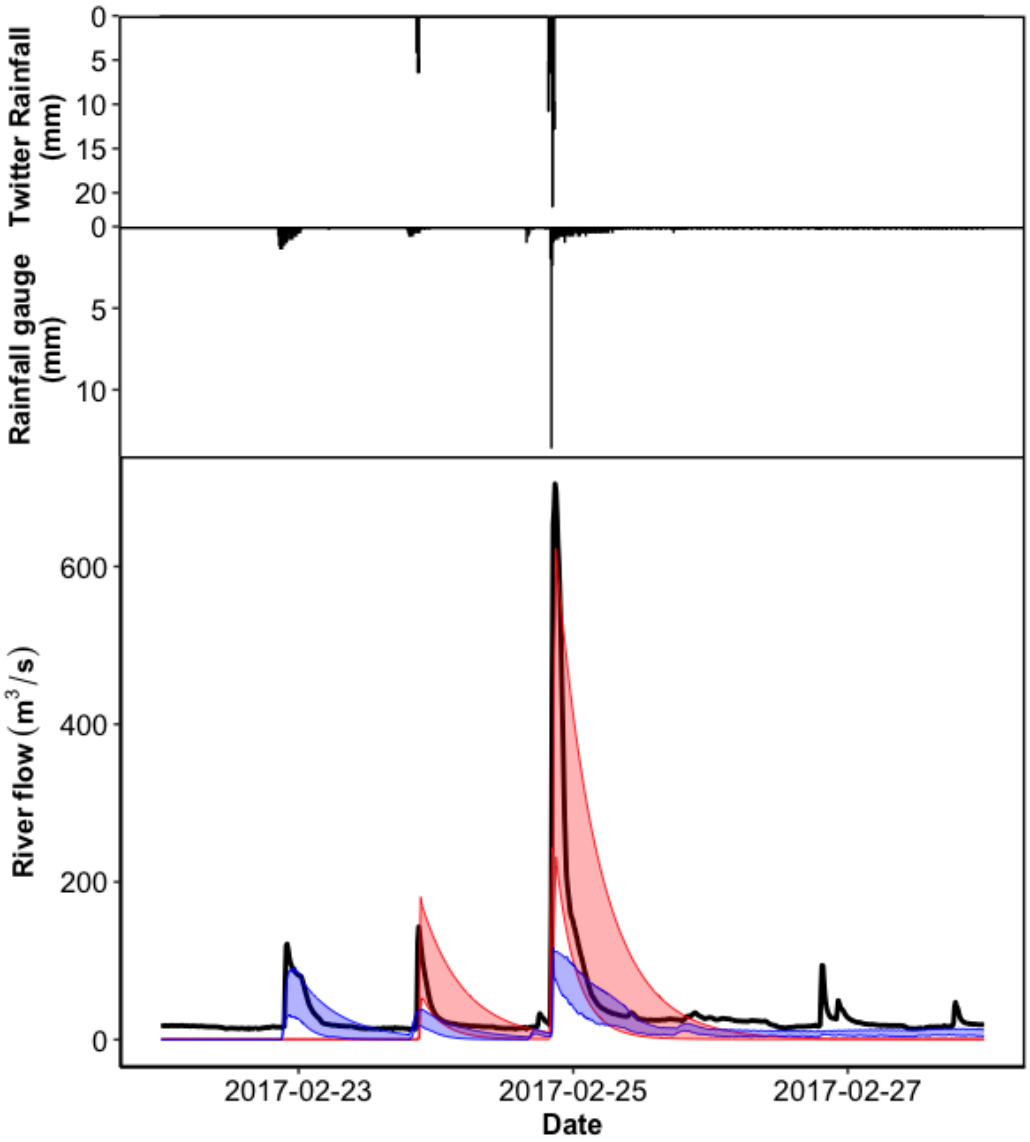


Figure 18. Streamflow simulation to the period February 22st to 28th, 2017.

Also it was simulated a combined rainfall variable composed by the social media proxy variable and the rainfall gauge. This combined simulation was taking the maximum and minimum values for the individual simulations of authoritative data and social media. In this case, the accuracy of the forecasting significantly increases, being able to predict correctly about 70.9% of the cases the value of the real streamflow. The underestimation is reduced to 28.6% and there is no overestimation for the period (see Table 6). This is an important result that clearly shows the potential in using data from social media to assist in monitoring environmental problems such as floods. An example of the combined simulation for the period from February 22nd to February 28th, 2017 is shown in Figure 19.

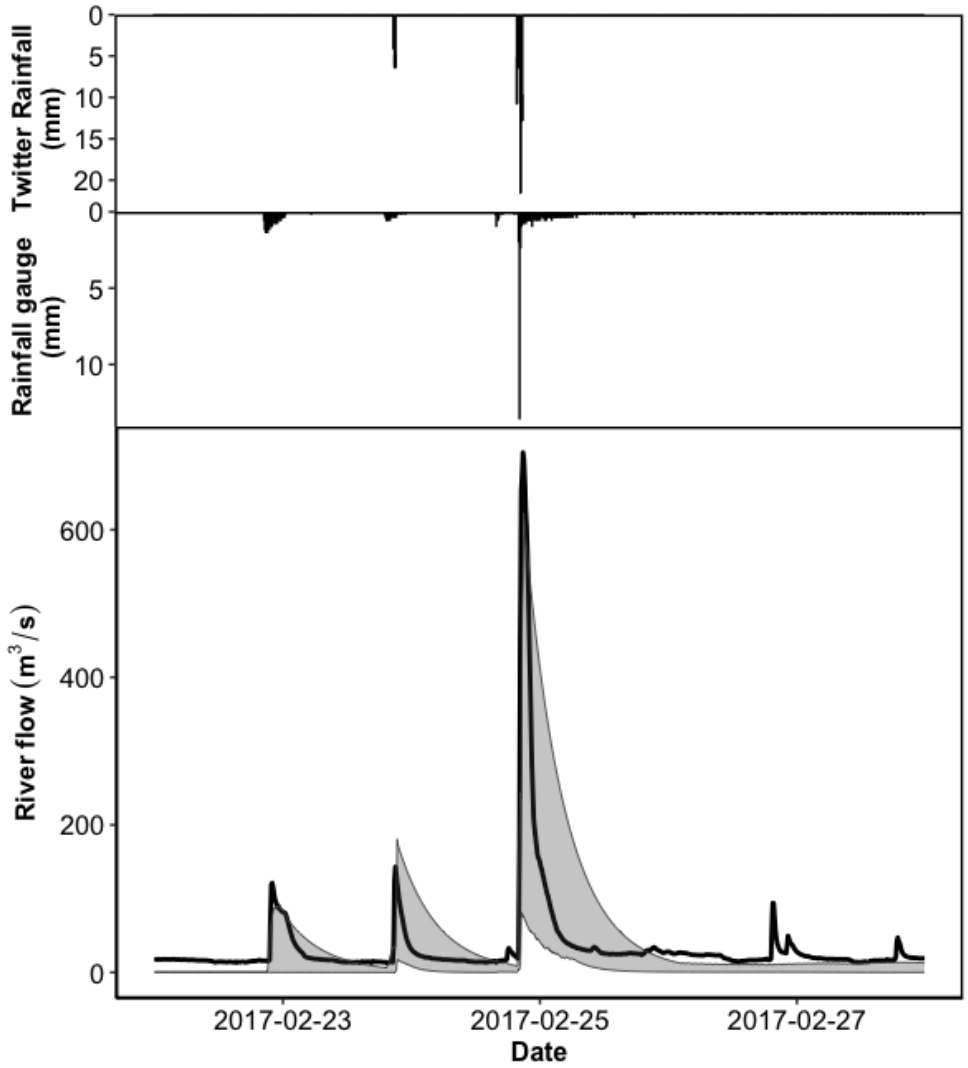


Figure 19. Combined streamflow simulation for the period from February, 22sh to 28th, 2016.

A summary of the streamflow simulation is shown in Table 6. Based on the values of the proxy variable obtained from Twitter, the simulation provides correct values in 31.3% of the cases, while overestimation represents 19.0% and underestimation 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 represent around respectively 58.4% and 3.0% of the cases.

Table 6. Percentage of correct estimate, overestimation and underestimation of the streamflows within the confidence interval using social media and authoritative data.

	Social media only	Authoritative sensor only	Composite of Social media and authoritative sensors
Observations estimated into model's confidence interval	31.3	38.6	70.9
Observations underestimated	49.5	58.4	28.6
Observations overestimated	19.0	3.0	0.5

2.5. Discussion

Results of this study supports the use of social media information to estimate precipitation or flow in poorly gauged catchments, which could help in issuing flood early warnings. In catchments in operation, but with incomplete records or with sensors under maintenance, the use alternative, social science proxy variables could become even more relevant. People's awareness by posting and sharing information through social media, capable of being transformed into viable proxy variables, as an alternative monitoring data source are of interest for community resilience, especially for streamflow forecasting purposes. Another possible application of social media-based information lies on

detecting authoritative sensors with on-line problems, thereby requiring maintenance.

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

In this thesis, modest values for the Adjusted Coefficient of Determination ($R^2_{adj} < 0.30$) were obtained in the equations that transforms social media data into precipitation, a result that complements our previous results discussed in Andrade et al. (2017). These values could be low due to, on the one hand, problems with the quality of rainfall gauge information, the modelling resolution and different synchronism time of sensors collected from different sources, i.e. national centers, state agencies with the social media posts. However, this temporal resolution is crucial for timing hydrological responses, like streamflows at an urban catchment. Nevertheless, these values could probably be improved with the use of other social media platforms (e.g. Instagram, Flickr) or considering other variables such as information quality protocols, spatiotemporal context like literacy and economic features of citizens posting social media, as well as the content of information, among others. In addition, other methods could be tested to transform the signal using other transformation algorithms to achieve better performance.

It is worth noting that the messages 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, which could be another limitation of the model. Additional research should also review the information according to the type of temporal context of the messages before, during or after the rainfall events or special thunderstorms.

3

DATA FUSION AND ASSIMILATION

In this chapter, the hypothesis is that it is possible to fusion authoritative rainfall and proxy from social media for flood monitoring. Chapter 2 described the authoritative rainfall, and in this chapter it will be explained how the rainfall proxy from social media is generated. In addition, it will discuss the fusion and assimilation of data using the Ensemble Kalman Filter.

3.1. Ensemble Kalman Filter (EnKF)

Data assimilation from different sources as a tool to improve the modelling of floods has been gaining more attention because nowadays there are more and more powerful processors that allow to analyse and manipulate a greater amount of data in increasingly shorter times (Moradkhani *et al.*, 2005). This has boosted the development of technologies and methods within what is now known as Big Data, understood as the management of large flows of structured and unstructured data, as well as statistical methods based on high dimensionality. The use of unstructured data sources, such as social media, in floods has shown its relevance when it comes to finding alternatives to monitoring and forecasting that quantitatively and qualitatively incorporate ways to data assimilation (Montanari *et al.*, 2013; Mazzoleni *et al.*, 2017). This is because having several sources of information and assimilating them is an efficient way to reduce uncertainty associated with measurements and models.

Kalman Filter (KF) (Kalman, 1960) is a technique used to assimilate data when the variable of interest can be measured only indirectly or when it has measurements from several sensors with noise. In it, the system is modelled with the use of state variables (e.g., soil moisture or flows), represented by x_t, of external forces called forcing (e.g., precipitation and evapotranspiration), represented with u_t, and the parameters of the model θ. The original filter is designed for linear systems, but hydrological systems are nonlinear dynamic systems with stochastic components (Pathiraja *et al.*, 2016). In this case, different variants of this filter have been developed. The most used filter is the Ensemble Kalman Filter (EnKF). Unlike other approaches in EnKF, it is not intended to linearize the system, but rather to deal with the non-

linear model itself. A general diagram of how the EnKF works is shown in Figure 20.

The EnKF is constituted by two steps alternately repeated. In the first step, called *a priori* or forecast step, a forecast is made with the model in a time step Δt generating the expected state vector. The change in the time of the system will then depend on the state variables at time t and the values of external forces u_{t+1}. In the second step, named as a posteriori or analysis step (Moradkhani *et al.*, 2005; Mazzoleni *et al.*, 2017), the state variables are updated with the help of a new measurement, y_{t+1}.

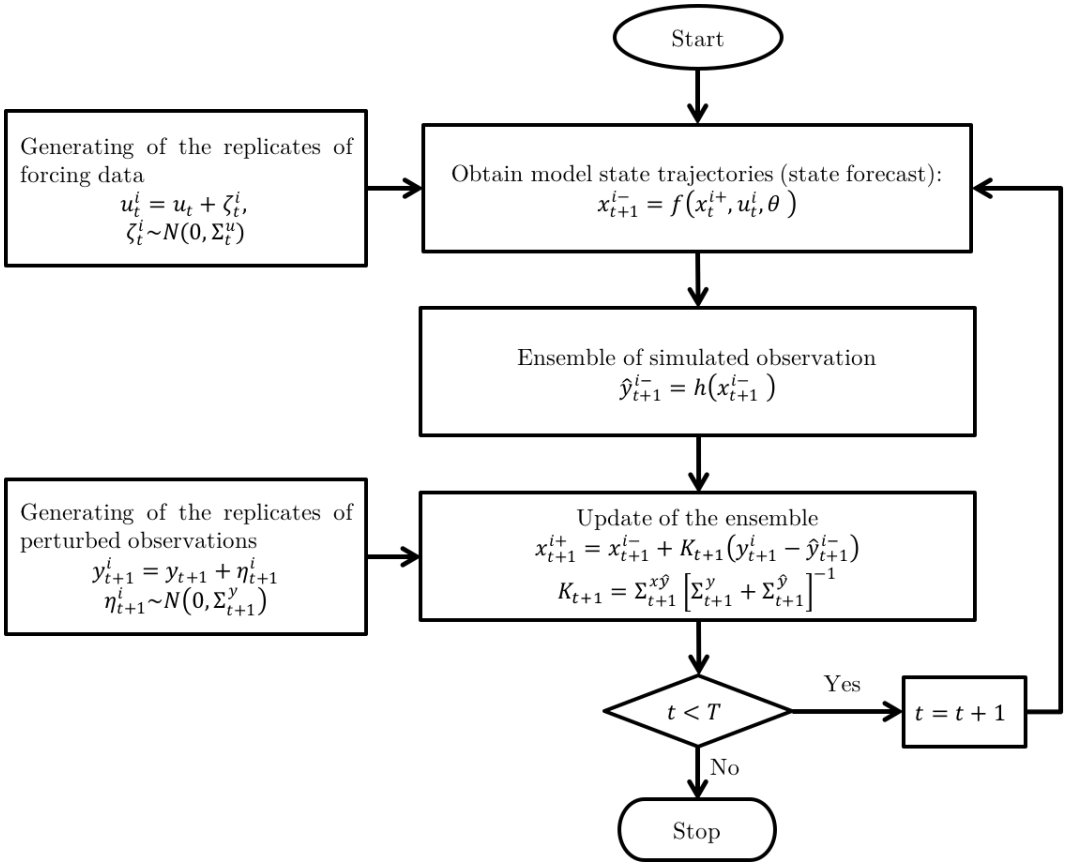


Figure 20. EnKF methodology. Adapted from Moradkhani et al. (2005).

This process is started by generating members of the forcing variable u_{t+1}^i assuming that it is constructed with the measured or predicted value u_t and that it has a Gaussian error ξ_{t+1}^i with zero mean and with covariance Σ_{t+1}^u. Each of the values generated from the forcing and the values of the state variables of previous steps are used inside the model f to generate the value of the state variables in the time step Δt. In what will be the *a priori* update of the hydrological system x_{t+1}^i. It can express in a general way the into the hydrological system as:

$$x_{t+1}^{i-} = f\left(x_t^{i+}, u_{t+1}, \Sigma_{t+1}^{x}, \theta\right) \tag{3.1}$$

where f is a function that maps the state variables from an instant t to an instant $t+1$. Σ_{t+1}^{x} is the forecast state error covariance. In addition to the parameters of the hydrological model, θ, that can be assumed as variants or invariant in time (Moradkhani *et al.*, 2005; Mazzoleni *et al.*, 2017).

In this case, the forecasted matrix of the ensemble of model states is:

$$x_{t+1}^{i-} = \left(x_{t+1}^{1-}, x_{t+1}^{2-}, \dots, x_{t+1}^{i-}, \dots, x_{t+1}^{N_{ens}-}\right) \tag{3.2}$$

and the ensemble error matrix is:

$$E_{t+1}^{-} = \left[x_{t+1}^{1-} - x_{t+1}^{true}, \dots, x_{t+1}^{i-} - x_{t+1}^{true}, \dots, x_{t+1}^{N_{ens}-} - x_{t+1}^{true}\right] \tag{3.3}$$

Since x_{t+1}^{true} is unknown, it is usually approximated by forecast ensemble mean $\bar{x}_{t+1}$, where

$$\bar{x}_{t+1} = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} x_{t+1}^{i-} \tag{3.4}$$

So, the forecast state error covariance is estimate as:

$$\Sigma_{t+1}^{x} = E_{t+1}^{-}(E_{t+1}^{-})^{T} \tag{3.5}$$

where

$$E_{t+1}^{-} = \left[x_{t+1}^{1-} - \bar{x}_{t+1}, \dots, x_{t+1}^{i-} - \bar{x}_{t+1}, \dots, x_{t+1}^{N_{ens}-} - \bar{x}_{t+1}\right] \tag{3.6}$$

In the second step, disturbances η_{t+1} in the measurement are generated to create the same number of measurements as the number of members. The different error perturbations η_{t+1} for the measurement are assumed Gaussian in this thesis with zero mean and covariance Σ_{t+1}^{y}.

Where

$$y_{t+1}^{i} = y_{t+1} + \eta_{t+1}^{i} \tag{3.7}$$

$$\eta_{t+1}^{i} \sim N\left(0, \Sigma_{t+1}^{y}\right) \tag{3.8}$$

Sometimes, it is necessary to transform the data to other units of measurement with the use of additional functions, in order to compare the values of the state variable x_{t+1} and measurements made by the sensors that will be considered to assimilate the data y_{t+1}. The function that maps x_{t+1} into $\hat{y}_{t+1}$ is denoted by h. If the model and the sensors result in the same variables, this transformation is not necessary (Moradkhani *et al.*, 2005; Mazzoleni *et al.*, 2017).

$$\hat{y}_{t+1}^{i-} = h\left(x_{t+1}^{i-}, \theta\right) \tag{3.9}$$

The Kalman gain K_{t+1}, is calculated using the covariance of the modelling and mapping errors $\Sigma_{t+1}^{x\hat{y}}$, the covariance of the mapping error $\Sigma_{t+1}^{\hat{y}}$ and the covariance of the measurement error Σ_{t+1}^{y}. It should be noted that is assumed that there is no covariance between the model error and the measurement error. The Kalman gain can be interpreted as a relative weight for measurement and forecast that is inversely proportional to the amount of noise of the variable.

$$K_{t+1} = \Sigma_{t+1}^{x\hat{y}}\left[\Sigma_{t+1}^{\hat{y}} + \Sigma_{t+1}^{y}\right]^{-1} \tag{3.10}$$

where

$$\Sigma_{t+1}^{\hat{y}} = \frac{1}{N_{ens}-1} E_{\hat{y}_{t+1}^{-}}\left(E_{\hat{y}_{t+1}^{-}}\right)^{T} \tag{3.11}$$

$$E_{\hat{y}_{t+1}^{-}} = \left[\left(\hat{y}_{t+1}^{1-} - \bar{\hat{y}}_{t+1}\right), \ldots, \left(\hat{y}_{t+1}^{i-} - \bar{\hat{y}}_{t+1}\right), \ldots, \left(\hat{y}_{t+1}^{N_{ens}-} - \bar{\hat{y}}_{t+1}\right)\right] \tag{3.12}$$

$$\bar{\hat{y}}_{t+1} = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} \hat{y}_{t+1}^{i-} \tag{3.13}$$

$$\Sigma_{t+1}^{x\hat{y}} = \frac{1}{N_{ens}-1} E_{t+1}^{-}(E_{t+1}^{-})^{T} \tag{3.14}$$

with the update given by

$$x_{t+1}^{i+} = x_{t+1}^{i-} + K_{t+1}\left[y_{t+1}^{i} - \hat{y}_{t+1}^{i-}\right] \tag{3.15}$$

Figure 21 shows how the members of the ensemble are propagating. In the a priori state it has some trajectories with their respective pdf. After this, observations appear with their probability distribution function. Finally, a posteriori state is estimated with the EnKF generating the new values and their associated pdf.

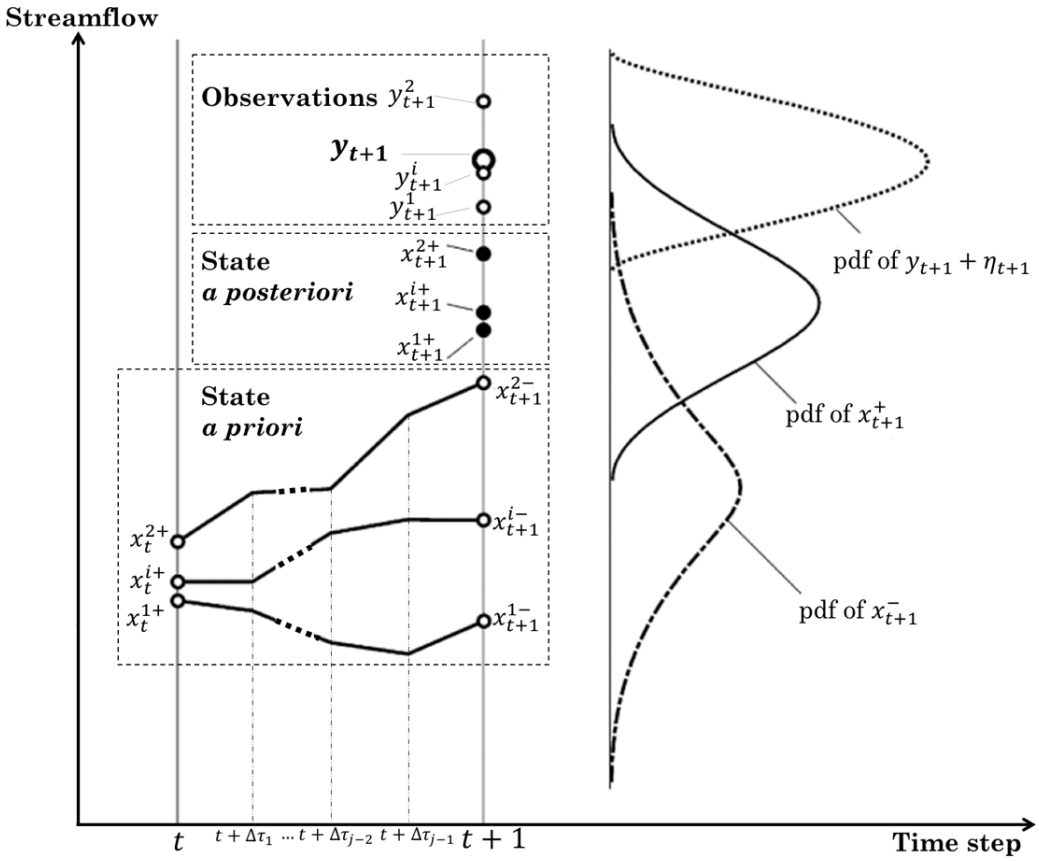


Figure 21. Schematic representation of the EnKF (adapted from Komma et al. (2008).

3.2. PDM

Here, due to a numerical instability presented in the PDM using the calibrations made in the first part of the work in conditions of data assimilation, it was decided to review and perform a new calibration of the PDM model. This new calibration was made considering that the hydrograph would fit better in the recession zone, instead of just thinking about the NSE. In this process, a single station corresponding to Boa Esperanza CEMADEN station was used, which had shown the best results in the previous calibration. Within this approach, the calibration NSE was 0.68, evidencing the improvement relative to the previous calibration, with a NSE 0.59 it should be noted, however, that this calibration was done with fewer events. The calibration was performed with data from January 24 until January 27, 2016, both dates included, every 10 minutes.

3.3. Methodology

To use the two available rain data sources, the authorized data and the proxy of social networks must be merged. This statistical process is carried out before applying the EnKF and generates as a result the forcing variable in the a priori step of the Kalman filter and, therefore, it becomes an additional step of the entire methodological process. The EnKF plus the fusion of rainfall data can be observed in Figure 22.

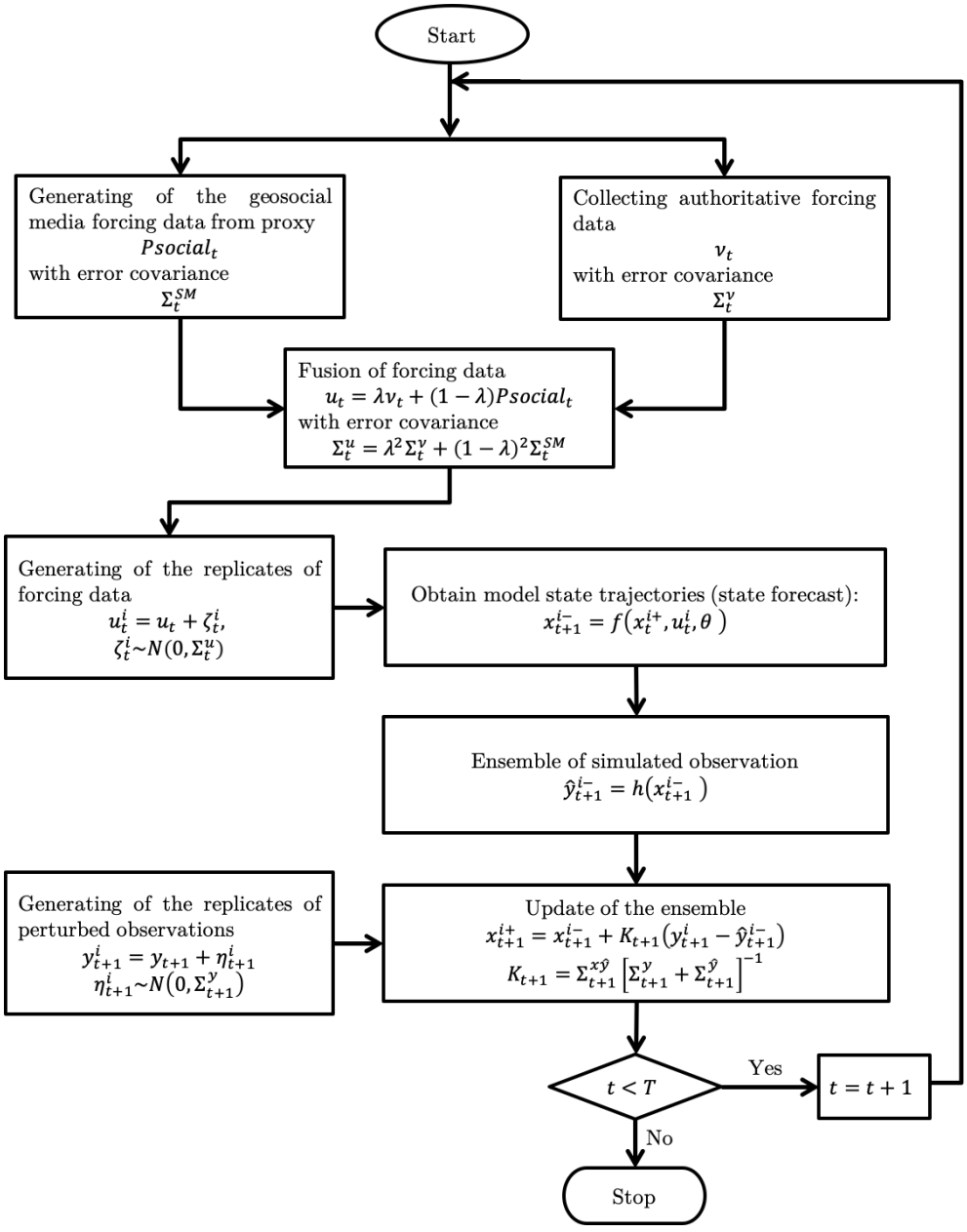


Figure 22. Methodological propose adapted from Moradkhani et al. (2005)

3.3.1. Fusion of authoritative and social media forcing data

Before using the data assimilation protocol into the EnKF, it is necessary to merge the two sources of forcing data. One of them from the authoritative rain gauges and the other one from rainfall in Twitter. v_{t+1} is defines as the rain that comes from rainfall gauges. The expected value of the rainfall from the authoritative rain gauges is the real value of the rain denoted as u_{t+1}^{true}. This value has an associated authoritative error variance Σ_{t+1}^{v}. It can write this as

$$E[v_{t+1}] = u_{t+1}^{true} \tag{3.16}$$

$$Var[v_{t+1}] = \Sigma_{t+1}^{v} \tag{3.17}$$

The rainfall coming from the transformation made in social media is s_{t+1}, its expected value is also the real value of the rainfall u_{t+1}^{true}, The variance of the social media proxy error is Σ_{t+1}^{s}.

$$E[s_{t+1}] = u_{t+1}^{true} \tag{3.18}$$

$$Var[s_{t+1}] = \Sigma_{t+1}^{s} \tag{3.19}$$

To merge these two values representing the same variable in a timely mode, but with different degrees of uncertainty, the weighted average was used, with λ as weight.

$$u_{t+1} = \lambda v_{t+1} + (1-\lambda)s_{t+1} \tag{3.20}$$

with

$$\Sigma_{t+1}^{u} = \lambda^2 \Sigma_{t+1}^{v} + (1-\lambda)^2 \Sigma_{t+1}^{s} \tag{3.21}$$

3.3.2. Errors in EnKF-PDM model

Models have errors because of simplifications and, in the input values, the estimation of parameters and, therefore, in their output values. When using data assimilation, all these errors are taken into account in the process in order to obtain the best possible estimate. Besides, the EnKF uses the structures of the errors of input variables mapped with the use of the model and the sensors error's measurement to generate the best possible estimate of the real state of the flow (Moradkhani *et al.*, 2005; Komma, Blöschl and Reszler, 2008; Alvarez-Garreton *et al.*, 2013). For this, reliably errors that are used in the modelling must be defined.

There are ways to model even the uncertainty in the model by incorporating data assimilation for the parameters, as shown by Moradkhani et al. (2005). In hydrological models, in general, many are the sources of uncertainty, as described in the introduction. This is one of the reasons why a simple model such as the PDM was chosen, in order

to avoid large parameterizations with uncertainties that are difficult to model in applications with few data and for basins with short response times. In PDM model, it has two input variables. Precipitation and potential evapotranspiration.

It is assumed that potential evapotranspiration is constant for simulation time and does not present error in the estimation, since the time step (every 10 minutes) and the values' range of input variables (between 2 and 3 mm per day) guarantee a small effect of this variable on the final flow, the variable of interest of the model.

The other input variable in the model is the rainfall. The estimation of its errors depends on factors such as the spatial distribution of the rain, the precipitated volume, the quantity and the calibration of the rain gauges, the number of rain gauges used to achieve the interpolated value, among others. In the literature, this error can be assumed as either homoscedastic (with a constant variance) (Komma, Blöschl and Reszler, 2008; Willems, 2009), or heteroscedastic (with a variable variance) (Carpenter and Georgakakos, 2004; Weerts and El Serafy, 2006; Rakovec *et al.*, 2012; Khorashadi, 2013; Rafieeinasab *et al.*, 2014). In this thesis will assume a heteroscedastic rainfall. In this case, the functional form of this variable can be expressed by (Khorashadi, 2013):

$$\Sigma_{t+1}^{u} = (a_h p + b_h)^2 \tag{3.21}$$

In this thesis, all sources of rainfall are assumed as heteroscedasticity. In the case of authoritative rainfall, it has:

$$v_{t+1}^{i} = v_{t+1} + \phi_{t+1}^{i} \tag{3.22}$$

where v_{t+1} represent the forcing variables of input, precipitation and potential evapotranspiration, the measurement error assumed Gaussian distribute $\phi_{t+1}^{i} \sim N(0, \Sigma_{t+1}^{v})$, and Σ_{t+1}^{v} is the variance of the measurement error of the rain gauges. The estimation errors of the rain gauges depend on factors such as the spatial distribution of the rain, the volume precipitated, the calibration of the rain gauges, the number of rain gauges used to achieve the interpolated value.

Here, it is assumed that for authoritative values, the parameter a_h, takes a value of 0.1, while b_h, will be assumed as 0.2 which is the minimum value measured by sensors in data from CEMADEN. The value of a_h was chosen taking values commonly used in the literature (Khorashadi, 2013).

In the case of rainfall from the proxy variable of Twitter, will assume that the value obtained through regression s_{t+1}, has an error ω_{t+1}^{i}.

$$P_{social}{}_{t+1}^{i} = P_{social}{}_{t+1} + \omega_{t+1}^{i} \tag{3.23}$$

where $\omega_{t+1}^{i} \sim N(0, \Sigma_{t+1}^{SM})$, and the variance Σ_{t+1}^{SM} thas the same functional form of the authoritative variance, only with values of the parameters a_h and b_h differentiated.

For b_h, assume the same value 0.2 as for authoritative data, because when the authoritative and proxy values are going to be merged the data value must be at least the one that the authoritative sensor has. Therefore, b_h in the case that both sensors record zero precipitation must have the same combined value. In contrast, after multiple simulations, it is observed that the maximum value that can be given to a_h for data from social media is 0.13. Higher values than 0.13 make the simulation divergent. This maximum value is taken as a safety factor, since the variable that is indirectly measured.

Measurement error

The streamflow measurement error η_{t+1}^{i} can be due to the calibration curve of the basin, the measurement error of the sensor that uses pressure or optical variables to determine the water level, the highly non-linear conditions of the flow, among others. This error will be modelled with a Gaussian distribution.

$$\eta_{t+1}^{i} \sim N(0, \Sigma_{t+1}^{v}) \tag{3.24}$$

For the variance of this error, it will have to consider that, as the flow increases, the measurement error tends to grow. From the literature, relations of the form in equation 3.25 are proposed (Komma, Blöschl and Reszler, 2008; Khorashadi, 2013):

$$\Sigma_{t+1}^{v} = (ez_k)^2 \tag{3.25}$$

Where the parameter e takes values between 0.001 and 0.1. For this thesis, it assumes a value of 0.001.

3.4. Experimental setup

In this section, the experiments performed to analyse the effects of a) assimilation only with authoritative rainfall data, b) assimilation only with geosocial media rainfall data, c) assimilation with authoritative and social media rainfall and d) assimilation with the artificial simulation of some common problems in authoritative rainfall data are described. In experiments, the updating frequency is considered different (30 minutes) to one model time step (10 minutes).

3.4.1. Experiment A. Assimilation only with the use of the authoritative rain gauge.

This experiment simulates the scenario in which only assimilating rainfall data from authoritative monitoring sources are used. This is the classic work scenario, in which authoritative data with adequate quality for monitoring are available for decision makers.

3.4.2. Experiment B. Assimilation only with the use of the geosocial media rain gauge.

This experiment simulates the scenario in which assimilation of rain data from geosocial sources are employed. This case is intended to show the alternative work scenario, based on the proxy built in the previous chapter (section 2.3.2.). In this experiment a scenario is supposed in which the authorized data are not available and the decision makers only have the data of the source proposed in this thesis.

3.4.3. Experiment C. Joint assimilation of authoritative and geosocial sources.

This experiment shows the scenario in which both data sources are used. It simulated a system with both sources working together. This redundancy would be used in the system in order to prevent possible known failures in the authoritative data.

3.4.4. Experiment D. Joint assimilation of authoritative and geosocial sources, simulating in an artificial way some typical failures of authoritative rain gauges.

As mentioned in the introduction, there are some common problems with rainfall data from authoritative sources. Some classic failures are those cases when the rain gauge does not register rainfall values, or the sensor stays marking a value in a static way, or even reaches very high extreme values that far exceed the actual occurring values.

Experiment D.1. the rain gauge does not register rainfall values

A classic failure that occurs in rain gauges, is when the sensor stops measuring rain. If this failure is not detected, the decision maker will think that in this watershed it is not raining and therefore will not see the need of monitoring what happens there, although it is really raining. To simulate this problem, it will assume that the authoritative sensor measures only zero anytime, while the geosocial media proxy works normally.

Experiment D.2. the rain gauge registers a constant rainfall value.

Another failure of the rain gauges is when the sensor measures the same value of rain for long time periods. This failure would lead to think that the accumulated rainfall is high and therefore overestimate what is really happening in the watershed. To simulate this failure, it will assume that the sensor constantly marks 0.2 mm which is the minimum measurement that it registers.

Experiment D.3. the rain gauge registers an atypical value.

Finally, another common failure that can occur is when in some moment the sensor registers very high outliers, but these values do not correspond with what it is really happening. This can lead the decision maker to issue a wrong warning. To simulate this scenario, the real value measured by the rain gauge is doubled.

3.5. Results and discussion

The colour convention in Figure 23 and Figure 24 is that in shaded blue with authoritative data, in red with social media proxy, in green both merged sources and observed streamflows in bold line. In the upper are the rainfall proxy from social media and authoritative rainfall. Figure 23 resumes those scenarios in which monitoring using both sources separately, and even in a joint way is simulated. Results show that when only one source is considered, underestimation and super esteem can occur, but at the time when both sources are used, forecasting improves. The average values of the NSE are 0.49 for modelling using authoritative data, -11.17 using only social media data and -0.79 using both sources at the same time. In the case of the Nash-log, the values were 0.66, 0.55 and 0.65 respectively.

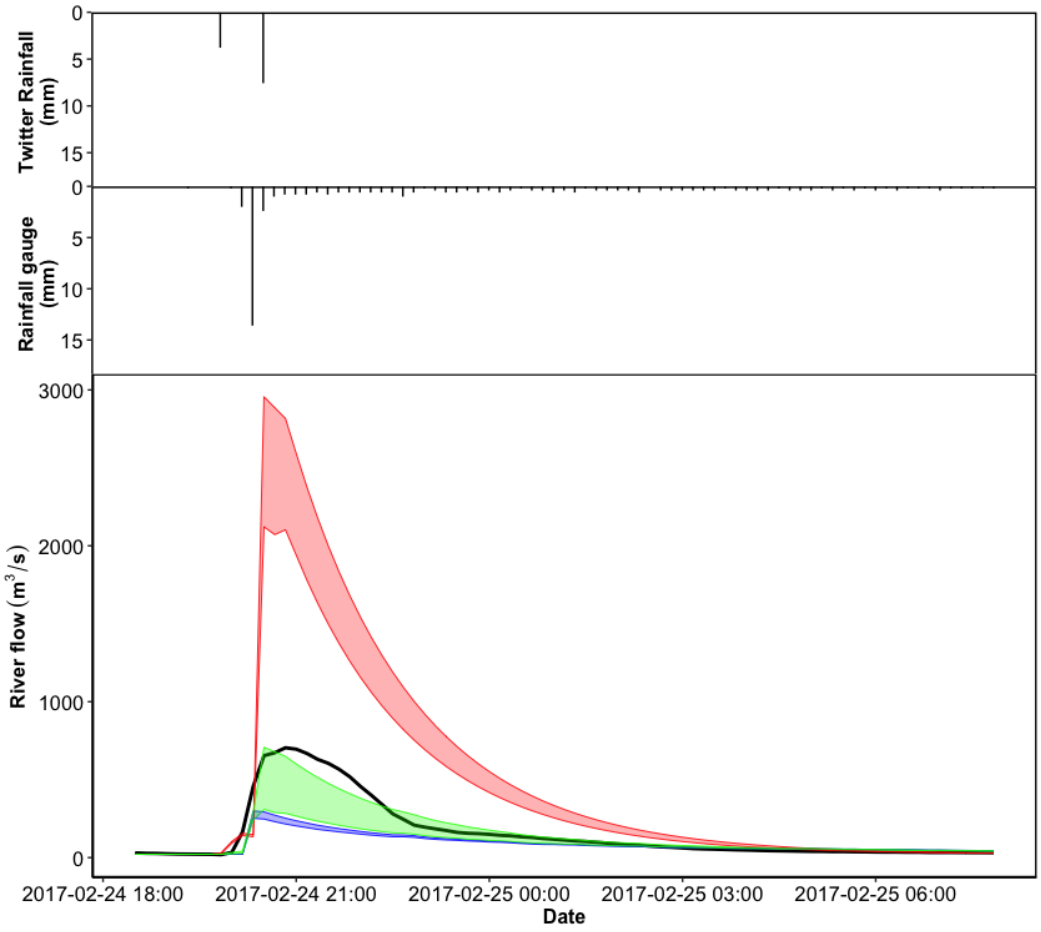


Figure 23. Simulation without assimilation.

Figure 24 resumes how the modelling behaves using data assimilation with the EnKF using only the authoritative source, the proxy coming from Twitter, and the case when both cases are merged. Merging of data, together with assimilation, represents better results. This is because it takes advantage of all the existing information.

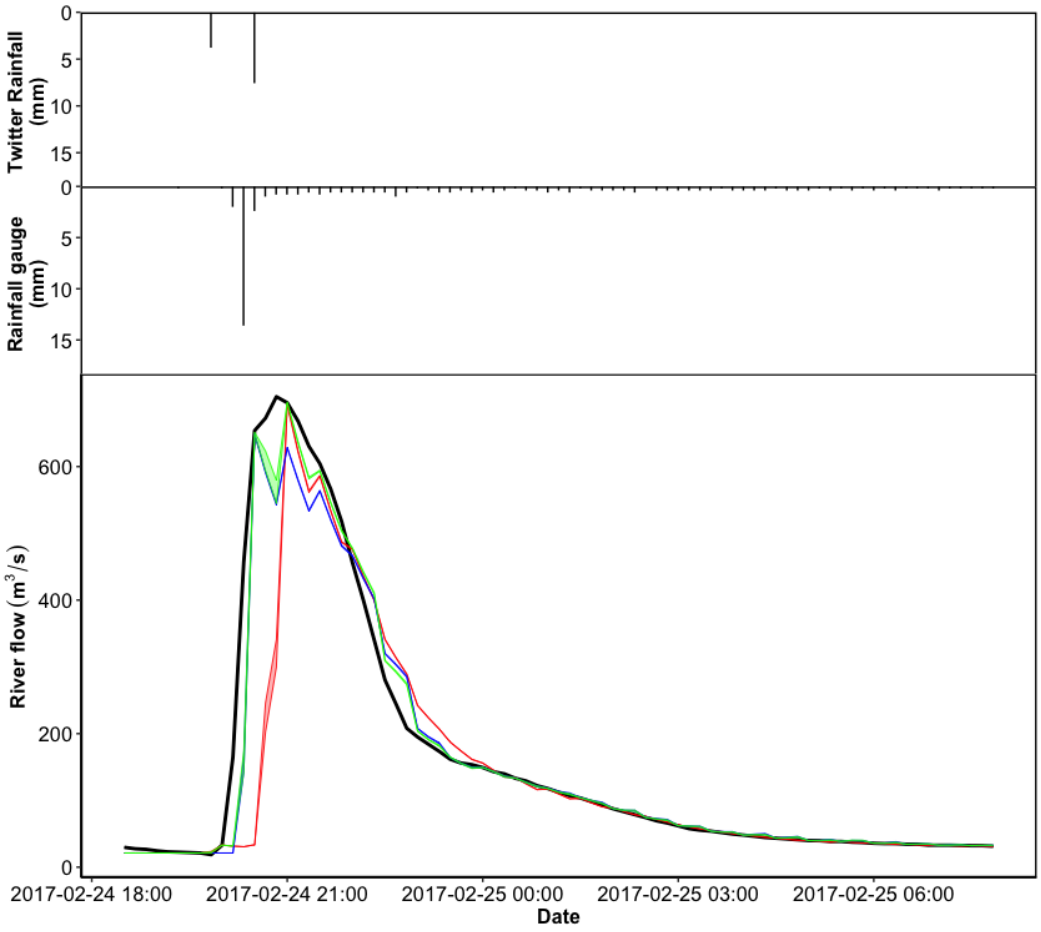


Figure 24. Simulation with assimilation

In the case of assimilation of data from authoritative sources, social media or data fusion, the NSE values are 0.90, 0.73 and 0.92, while for the Nash-log the values would be 0.53, 0.53 0.62, respectively (Figure 25). Showing an improvement in the indicator, being a sample that when assimilating takes advantage of the information coming from social media.

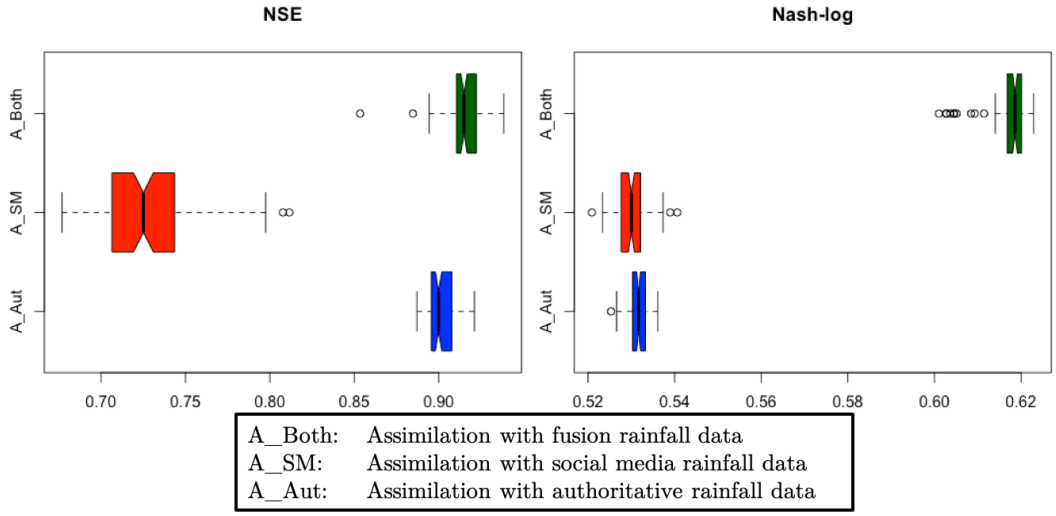


Figure 25. NSE and Nash-log for data assimilation

When authoritative data were analysed, a group of problems common to some of the sensors emerged. The problems included sensors that never changed from zero as a measured value. Values of 0.2 constants over time, i.e. every 10 minutes the sensor recorded 0.2 as the rainfall of those last 10 minutes and remained that way for days or even weeks. Another error, not so common, but that seems to occur, is the report of much higher values than seems to be occurring. To check how the proxy approach can collaborate in these 3 scenarios called D1, D2 and D3, additional runs were performed, including data merging and assimilation. The results can be seen in Figure 26, Figure 27 and Figure 28. The graphical convention in Figure 26, Figure 27, and Figure 28 is that in shaded purple both merged sources and observed streamflows in bold line.

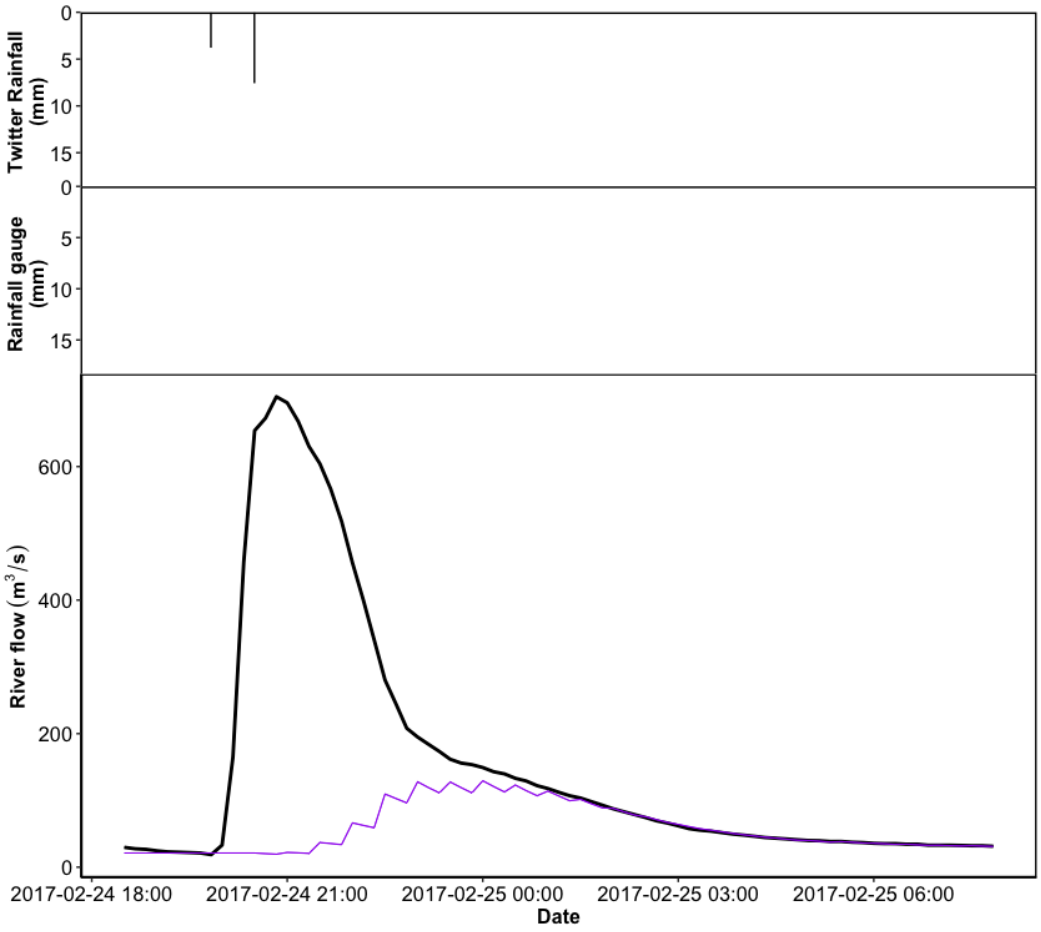


Figure 26. Simulation for scenario D1. Authoritative sensor fail.

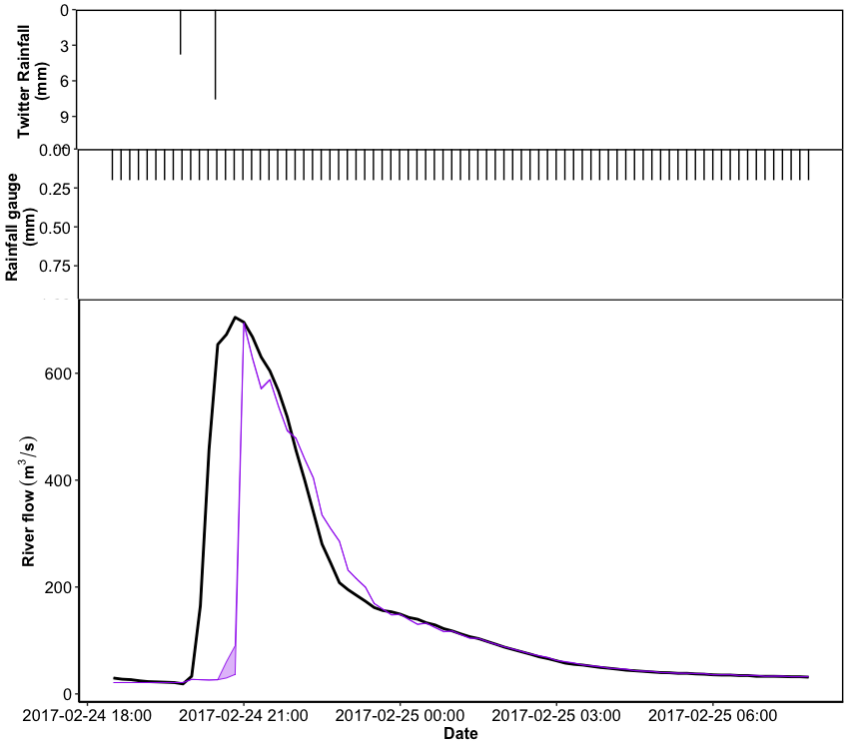


Figure 27. Simulation for scenario D2. Authoritative sensor fail

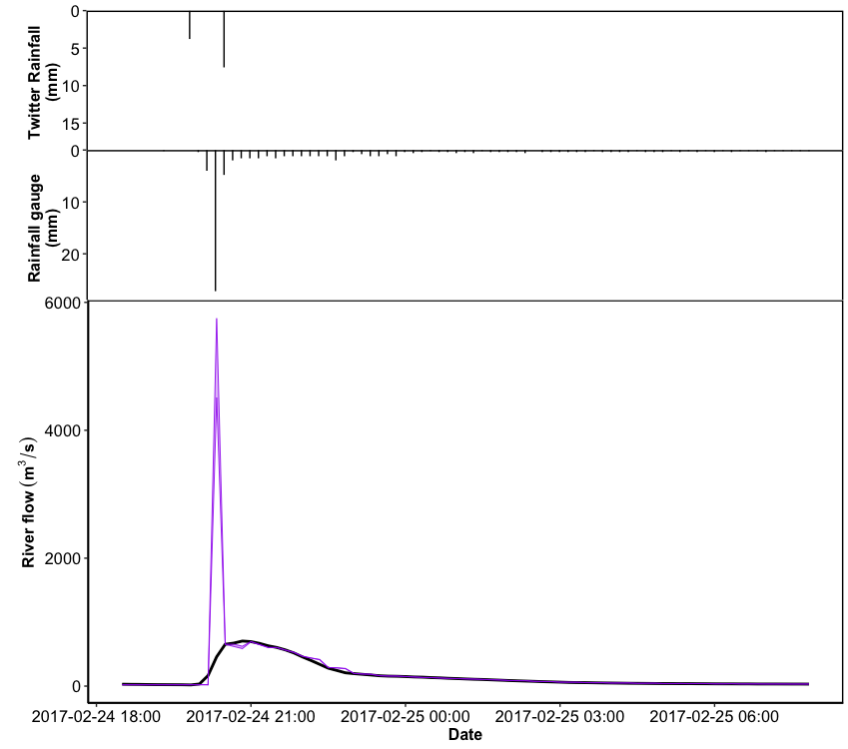


Figure 28. Simulation for scenario D3. Authoritative sensor fail

For scenario D1 it has that even with data assimilation the behaviour is not adequate. This may be due to the fact that there is not enough volume of water in the model to be able to make the hydrological balance adequately. For scenarios D2 and D3 present better results with respect to the cases in which these problems were encountered without the input of social media data. This is because the filter was raised relying more on the authoritative data (90%) with respect to the proxy of geo-social data. In addition, in the particular case of scenario D3, when overestimating the EnKF tends to privilege the values of the level sensor. This methodology serves to improve the hydrological forecast if are considered common problems in data. This would greatly help the persons in charge of making decisions to take early warnings. However, it must be explored other ways to capture signals and check other temporal and spatial scales in order to obtain better results.

The results show that, considering concentrated models in joint use with assimilation of streamflow data provides an improvement in the estimation of the output hydrograph. In particular, its concludes that improves even when the input data is more uncertain as in scenarios D2 and D3. The information contained in the virtual sensor from the social media proxy can be an alternative tool in watersheds with few data or basins where there are problems with sensor networks. It is important to note that this alternative source is intermittent and that it will not always have a real measure in time due to the limitations set forth in chapter 2. An application of this type of approach for other basins and/or cities should always be considered as a whole with assimilation of streamflow data.

According to Montanari et al. (2013) one of the concepts of Panta Rhei is that the interaction between hydrology and society is changing, generating new feedbacks which need to be understood, assessed, modelled and predicted by adopting an interdisciplinary approach. Within this concept, the widespread use of technological tools and communication aid in the consolidation of "citizen science". These data from these strategies (structured or unstructured) help to better understand the feedback between hydrological processes, society and ecosystems provide important information on the functioning of the catchment. In addition, data management techniques aid to understand and use these new sources of hydrological information. Tauro et al. (2018) present as current paradigms of hydrology the use of crowdsourcing data as aid to measure rainfall. They present as a goal of the Measurements and Observations in the XXI century (MOXXI) the construction of low cost and / or opportunistic systems that allow the measurement of hydrological variables. Therefore, this methodological

proposal that contains a look at the hydrological process of flood integrating people, social media and statistical and hydrological techniques framed within the Panta Rhei and MOXXI initiatives.

CONCLUSIONS AND FUTURE WORK

A. RESEARCH OUTCOMES

Below, the research outcomes in relation to the specific research objectives are reported.

The main objective of this thesis is to propose a methodology with the aid of data assimilation techniques that allows the incorporation of alternative and complementary data sources to improve short-term streamflow monitoring in cities where sensor networks fail or are inexistent.

This thesis shows the possibility of using a methodology that incorporates assimilation and fusion of data in the solution of common problems in hydrology. These problems can be scars or lack of data. These methodologies should incorporate the uncertainty associated with the data sources to make an adequate forecast or monitoring. Going one step further towards a quantitative integration of social media activity into flood forecasting models is important as a complementary resource for monitoring catchments, considering the fact that sometimes the rainfall gauges, usually used for this activity, are not available or fail for various reasons, such as a lack of maintenance.

1. To propose a rainfall proxy variable from data from social media to be used in flood monitoring.

   It is possible to use data from social media to build a rainfall proxy variable that can be used in monitoring and forecasting rainfall. Care should be taken when using it because it is an approximation and predictions, this would depend on human factors that can be overestimated and it is not always possible to motivate people to post messages related to rain or flooding in social media. If possible, its use should be complementary to other sources in order to validate the existence and magnitude of the phenomenon. In the case of not having other sources of contrast, it should only be used as an indicator that something is happening in the location.

2. To implement a data assimilation technique that allows to update the state variables in each step of time and the fusion of authoritative data and social media in flood monitoring.

   Data assimilation and fusion techniques serve to integrate various data sources which have uncertainties associated with the measurement. The decision-maker must understand the problems on this alternative socio-hydrological information data sources with the purpose to take the proper decision when issues

early warning. It is important that decision makers should be always informed about the magnitudes generated under this methodology to analyse it in a prudent manner since data comes from an unconventional source that has uncertainties in magnitudes and frequencies.

3. To develop guidelines to use alternative data sources in hydrology in cases of joint monitoring, for authoritative and alternative data, in “poorly gauged” or "ungauged" catchments.

   The methodology proposed in this thesis includes, in addition to the proxy from social media, the fusion with authoritative rainfall data. It also shows that the system is more efficient if a data assimilation method such as the EnKF is used considering sources uncertainties. The methodology includes the fusion of rain from authoritative sources and social media in a step prior to data assimilation. It should be made clear that the proxy should be reviewed periodically since it may vary over time for the same location. In addition, the spatial and temporal scales used in this thesis influence the indicators. These scales depended on response time of catchments. For basins with longer response times the results may change. The same as for smaller cities.

This thesis provides strong evidence that data from geo-social media can be used to derive proxy variables for rainfall and flow. In this thesis was used frequency of related messages from social media as a proxy for rainfall, which in turn can be used as an input for hydrological models to predict streamflows and flood conditions. To issue flood early warnings, data from social media were used to improve either rainfall monitoring from observational authoritative networks and even observed urban streamflow. Evidences showed that better results can be achieved by combining both authoritative data and social media together. By using only available social media data should be taken with caution, because of progressive biases and uncertainty in streamflow estimation. On the one hand, 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, which escapes the objectives of this thesis. Notwithstanding, methods presented in this thesis would improve the availability of multiple sources of data and information to make cities more resilient to extreme events such as floods.

The assimilation of data is, in general, a valuable tool when it comes to improve the hydrological modelling. In the case of having values from different sensors, it is an excellent tool to find the values of the state

variables considering uncertainty associated. That is, the Ensemble Kalman Filter procedure serves as a means to use models and sensors together, considering the uncertainty associated to each one and presenting a set of possible values for state variables.

It is important to know that the data coming from non-traditional sensors must be worked with great care. Because although this may capture the appearance of the phenomenon, the magnitude of this is uncertain. In the case of rain that is going to be used in a hydrological model, the cascade of uncertainty should be managed through data assimilation methods and, as much as possible, fusion with other sources. In the case of not having other sources the decision maker should think that proxy magnitudes are uncertain. However, it should also be noted that these data have information that is useful for a decision maker when issuing alerts.

B. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORK

This thesis has certain limitations which should be mentioned. It leads to formulation of some recommendations for future work.

One of the main limitations is the fact of not being able to control the behaviour of people. What is done is to try to capture this behaviour of users of social media to use it quantitatively in the prediction of a physical phenomenon. People are not asked to react in a certain way or with certain words. Neither can be controlled the number of people who react to the rain. This limitation means that on certain occasions it can overestimate or underestimate what really happens with the spatio-temporal variability of rainfall and runoff in a catchment. It is necessary to adequately assess the uncertainty associated with any proxy variable extracted from social media related to rain because it can be associated with floods.

Given that the behaviour in social media varies during the year, as well as the phenomenon, it would be interesting to see how these patterns evolve over intra- and inter-annually periods. Furthermore, it is worth recommending further works on how social media, say Twitter, signal can change in other spatial contexts, that is, other cities and even with minor aggregations. Other ways of generating members of assimilation should be explored, such as using uniform distribution or some extreme values distributions. Also some works directed to measure the uncertainty of the components.

Another limitation is related to the type of regression used to translate social media perception (a qualitative information) into pseudo rainfall units (a quantitative measure of the hydrologic balance). By adopting

this regression as a polynomial regression, as used/proposed in this thesis, would be further necessary to explore other models, such as neural networks or even self-adjusting regressions.

These results would have to be reviewed through a cost-benefit analysis since these systems are cheaper than an authoritative sensor network that requires large investments in installation and maintenance.

Another recommendation would be in terms of training. It is necessary that hydrologists and decision makers have clear aspects related to socio-hydrology and data assimilation. It is important to raise awareness of the importance of the redundant use of information in monitoring systems where some of the components may fail at critical times when its necessary to issue alerts. Likewise, the assimilation of data is fostered as a necessary methodology in the daily work of hydrology.

REFERENCES

Albuquerque, J. P. De, Herfort, B. and Brenning, 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*. Taylor & Francis, 29(4), pp. 667–689. doi: 10.1080/13658816.2014.996567.

Almeida, I. K. de, Almeida, A. K., Anache, J. A. A., Steffen, J. L. and Alves Sobrinho, T. (2014) 'Estimation on time of concentration of overland flow in watersheds: A review', *Geociencias*, 33(4), pp. 661–671.

Alvarez-Garreton, C., Ryu, D., Western, A. W., Crow, W. and Robertson, D. (2013) 'Impact of observation error structure on satellite soil moisture assimilation into a rainfall-runoff model', (December), pp. 1–6. Available at: www.mssanz.org.au/modsim2013.

Alvarez-Garreton, C., Ryu, D., Western, A. W., Crow, W. T. and Robertson, D. E. (2014) 'The impacts of assimilating satellite soil moisture into a rainfall–runoff model in a semi-arid catchment', *Journal of Hydrology*. Elsevier B.V., 519, pp. 2763–2774. doi: 10.1016/j.jhydrol.2014.07.041.

Andrade, S. C. De, Restrepo-estrada, C., Delbem, A. C. B., Mendiondo, E. M. and Albuquerque, J. P. De (2017) 'Societal Geo-innovation', pp. 19–37. doi: 10.1007/978-3-319-56759-4.

Boni, G., Ferraris, L., Pulvirenti, L., Squicciarino, G., Pierdicca, N., Candela, L., Pisani, A. R., Zoffoli, S., Onori, R., Proietti, C. and 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(6), pp. 2794–2805. doi: 10.1109/JSTARS.2016.2514402.

Brouwer, T., Eilander, D., Van Loenen, A., Booij, M. J., Wijnberg, K. M., Verkade, J. S. and Wagemaker, J. (2017) 'Probabilistic flood extent estimates from social media flood observations', *Natural Hazards and Earth System Sciences*, 17(5), pp. 735–747. doi: 10.5194/nhess-17-735-2017.

Carpenter, T. M. and Georgakakos, K. P. (2004) 'Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model', 298, pp. 202–221. doi: 10.1016/j.jhydrol.2004.03.036.

Chen, X., Zhang, L., Gippel, C. J., Shan, L., Chen, S. and Yang, W. (2016) 'Uncertainty of Flood Forecasting Based on Radar Rainfall Data Assimilation', *Hindawi Publishing Corporation*, 2016. doi: 10.1155/2016/2710457.

Crochemore, L., Ramos, M. H. and Pappenberger, F. (2016) 'Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts', *Hydrology and Earth System Sciences*, 20(9), pp. 3601–3618. doi: 10.5194/hess-20-3601-2016.

Crooks, A., Croitoru, A., Stefanidis, A. and Radzikowski, J. (2013) '#Earthquake: Twitter as a Distributed Sensor System', *Transactions in GIS*, 17(1), pp. 124–147. doi: 10.1111/j.1467-9671.2012.01359.x.

Degrossi, L. C., Albuquerque, J. P. De, Fava, M. C. and Mendiondo, E. M. (2014) 'Flood Citizen Observatory : a crowdsourcing- based approach for flood risk management in Brazil', in.

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

Fraternali, P., Castelletti, A., Soncini-Sessa, R., Vaca Ruiz, C. and Rizzoli, A. E. (2012) 'Putting humans in the loop: Social computing for Water Resources Management', *Environmental Modelling and Software*. Elsevier Ltd, 37, pp. 68–77. doi: 10.1016/j.envsoft.2012.03.002.

Goodchild, M. F. (2007) 'Citizens as sensors: The world of volunteered geography', *GeoJournal*, 69(4), pp. 211–221. doi: 10.1007/s10708-007-9111-y.

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

Hapuarachchi, H. A. P., Wang, Q. J. and Pagano, T. C. (2011) 'A review of advances in flash flood forecasting', *Hydrological Processes*, 25(18), pp. 2771–2784. doi: 10.1002/hyp.8040.

Horita, F. E. A., Albuquerque, J. P. de, Degrossi, L. C., Mendiondo, E.

M. and 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 and Geosciences*. Elsevier, 80, pp. 84–94. doi: 10.1016/j.cageo.2015.04.001.

Horita, F. E. A., de Albuquerque, J. P., Marchezini, V. and 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*. The Authors, 97, pp. 12–22. doi: 10.1016/j.dss.2017.03.001.

Huang, Q. and Xiao, Y. (2015) 'Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery', *ISPRS International Journal of Geo-Information*, 4(3), pp. 1549–1568. doi: 10.3390/ijgi4031549.

Huang, X., Wang, C. and Li, Z. (2018) 'Reconstructing Flood Inundation Probability by Enhancing Near Real-Time Imagery With Real-Time Gauges and Tweets', pp. 1–11.

Kalman, R. E. (1960) 'A New Approach to Linear Filtering and Prediction Problems', *ASME, Journal of Basic Engineering*, 82(Series D), pp. 35–45. doi: 10.1115/1.3662552.

Kay, A. L., Davies, H. N., Bell, V. A. and Jones, R. G. (2009) 'Comparison of uncertainty sources for climate change impacts: flood frequency in England', *Climatic Change*, 92(1–2), pp. 41–63. doi: 10.1007/s10584-008-9471-4.

Khorashadi, F. Z. (2013) 'Evaluating the Ensemble Kalman Filter in Flood Forecasting Updating', (June).

Komma, J., Blöschl, G. and Reszler, C. (2008) 'Soil moisture updating by Ensemble Kalman Filtering in real-time flood forecasting', *Journal of Hydrology*, 357(3–4), pp. 228–242. doi: 10.1016/j.jhydrol.2008.05.020.

Lamb, R. (1999) 'Calibration of a conceptual rainfall-runoff model for flood frequency estimation by continuous simulation', *Water Resources Research*, 35(10), pp. 3103–3114. doi: 10.1029/1999WR900119.

Leach, J. M., Kornelsen, K. C. and Coulibaly, P. (2018) 'Assimilation of near-real time data products into models of an urban basin', *Journal of Hydrology*. Elsevier, 563(April), pp. 51–64. doi: 10.1016/j.jhydrol.2018.05.064.

Li, Y., Grimaldi, S., Walker, J. P. and Pauwels, V. R. N. (2016) 'Application of remote sensing data to constrain operational rainfall-driven flood forecasting: A review', *Remote Sensing*, 8(6). doi: 10.3390/rs8060456.

Li, Z., Wang, C., Emrich, C. T. and Guo, D. (2018) '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*. Taylor & Francis, 45(2), pp. 97–110. doi: 10.1080/15230406.2016.1271356.

Listo, F. de L. R. and Vieira, B. C. (2012) 'Mapping of risk and susceptibility of shallow-landslide in the city of São Paulo, Brazil', *Geomorphology*. Elsevier B.V., 169–170, pp. 30–44. doi: 10.1016/j.geomorph.2012.01.010.

Mazzoleni, M., Verlaan, M., Alfonso, L., Monego, M., Norbiato, D., Ferri, M. and Solomatine, D. P. (2017) 'Can assimilation of crowdsourced data in hydrological modelling improve flood prediction?', *Hydrology and Earth System Sciences*, 21(2), pp. 839–861. 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), pp. 130–143.
Michailidi, E. M., Antoniadi, S., Koukouvinos, A., Bacchi, B. and Efstratiadis, A. (2018) 'Timing the time of concentration: shedding light on a paradox', *Hydrological Sciences Journal*. Taylor & Francis, 63(5), pp. 721–740. doi: 10.1080/02626667.2018.1450985.

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., Koutsoyiannis, D., Cudennec, C., Toth, E., Grimaldi, S., Blöschl, G., Sivapalan, M., Beven, K., Gupta, H., Hipsey, M., Schaefli, B., Arheimer, B., Boegh, E., Schymanski, S. J., Di Baldassarre, G., Yu, B., Hubert, P., Huang, Y., Schumann, A., Post, D. A., Srinivasan, V., Harman, C., Thompson, S., Rogger, M., Viglione, A., McMillan, H., Characklis, G., Pang, Z. and Belyaev, V. (2013) '"Panta Rhei—Everything Flows": Change in hydrology and society—The IAHS Scientific Decade 2013–2022', *Hydrological Sciences Journal*, 58(6), pp. 1256–1275. doi: 10.1080/02626667.2013.809088.

Moore, R. J. (2007) 'The PDM rainfall-runoff model', 11(1), pp. 483–499.

Moradkhani, H., Sorooshian, S., Gupta, H. V. and Houser, P. R. (2005) 'Dual state–parameter estimation of hydrological models using ensemble Kalman filter', *Advances in Water Resources*, 28(2), pp. 135–147. doi: 10.1016/j.advwatres.2004.09.002.

Muleta, M. K. (2012) 'Model Performance Sensitivity to Objective Function during Automated Calibrations', *Journal of Hydrologic Engineering*, 17(6), pp. 756–767. doi: 10.1061/(ASCE)HE.1943-5584.0000497.

Nash, E. and Sutcliffe, V. (1970) 'PART I- A DISCUSSION OF PRINCIPLES * The problem of determining river flows from rainfall , evaporation , and other factors , occupies a central place in the technology of applied hydrology . It is not only the essential problem of flood forecasting but a', 10, pp. 282–290.

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., Kroll, S., Murlà-Tuyls, D., Tisserand, B., Schertzer, D., Tchiguirinskaia, I., Onof, C., Willems, P. and Ten Veldhuis, M. C. (2015) 'Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation', *Journal of Hydrology*. Elsevier B.V., 531, pp. 389–407. doi: 10.1016/j.jhydrol.2015.05.035.

Pagano, T., Hapuarachchi, P. and Wang, Q. J. (2009) 'Continuous Soil Moisture Accounting and Routing Modelling to Support Short Lead-Time Streamflow Forecasting Water for a Healthy Country Flagship Report Water series', (July). Available at: http://www.csiro.au/partnerships/WIRADA.html.

Patankar, A. and 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*. Springer Netherlands, 80(1), pp. 285–310. doi: 10.1007/s11069-015-1968-3.

Patel, N. N., Stevens, F. R., Huang, Z., Gaughan, A. E., Elyazar, I. and Tatem, A. J. (2017) 'Improving Large Area Population Mapping Using Geotweet Densities', *Transactions in GIS*, 21(2), pp. 317–331. doi: 10.1111/tgis.12214.

Pathiraja, S., Marshall, L., Sharma, A. and Moradkhani, H. (2016) 'Detecting non-stationary hydrologic model parameters in a paired catchment system using data assimilation', *Advances in Water Resources*. Elsevier Ltd, 94, pp. 103–119. doi: 10.1016/j.advwatres.2016.04.021.

Porto de Albuquerque, J., Horita, F. E. A., Degrossi, L. C., Rocha, R. dos S., Camargo de Andrade, S., Restrepo-Estrada, C. and Leyh, W. (2017) 'Leveraging Volunteered Geographic Information to Improve Disaster Resilience', in *Volunteered Geographic Information and the Future of Geospatial Data*, pp. 158–184. doi: 10.4018/978-1-5225-2446-5.ch009.

Rafieeinasab, A., Seo, D., Lee, H. and Kim, S. (2014) 'Comparative evaluation of maximum likelihood ensemble filter and ensemble Kalman filter for real-time assimilation of streamflow data into operational hydrologic models', *Journal of Hydrology*. Elsevier B.V., 519, pp. 2663–2675. doi: 10.1016/j.jhydrol.2014.06.052.

Rakovec, O., Weerts, A. H., Hazenberg, P., Torfs, P. J. J. F. and Uijlenhoet, R. (2012) 'State updating of a distributed hydrological model with Ensemble Kalman Filtering : effects of updating frequency and observation network density on forecast accuracy', (1), pp. 3435–3449. doi: 10.5194/hess-16-3435-2012.

Rathore, M., Ahmad, A., Paul, A., Hong, W.-H. and Seo, H. (2017) 'Advanced computing model for geosocial media using big data analytics', *Multimedia Tools and Applications*. Multimedia Tools and Applications, 76(23), pp. 24767–24787. doi: 10.1007/s11042-017-4644-7.

Restrepo-Estrada, C., de Andrade, S. C., Abe, N., Fava, M. C., Mendiondo, E. M. and de Albuquerque, J. P. (2018) 'Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring', *Computers and Geosciences*, 111, pp. 148–158. doi: 10.1016/j.cageo.2017.10.010.

Rosser, J. F., Leibovici, D. G. and Jackson, M. J. (2017) 'Rapid flood inundation mapping using social media, remote sensing and topographic data', *Natural Hazards*. Springer Netherlands, 87(1), pp. 103–120. doi: 10.1007/s11069-017-2755-0.

Sakaki, T. and Matsuo, Y. (2012) 'Earthquake Observation by Social Sensors', in D'Amico, S. (ed.) *Earthquake Research and Analysis - Statistical Studies, Observations and Planning Downloaded*.

Salimi, E. T., Nohegar, A., Malekian, A., Hoseini, M. and Holisaz, A. (2017) 'Estimating time of concentration in large watersheds', *Paddy and Water Environment*. Springer Japan, 15(1), pp. 123–132. doi: 10.1007/s10333-016-0534-2.

Schnebele, E., Cervone, G., Kumar, S. and Waters, N. (2014) 'Real Time Estimation of the Calgary Floods Using Limited Remote Sensing Data', *Water*, 6(2), pp. 381–398. doi: 10.3390/w6020381.

Silva, G. C. M. da (2010) *Opções de adaptação às mudanças do clima para a bacia do rio Aricanduva*. Universidade de São Paulo. doi: 10.11606/D.8.2010.tde-25102010-162051.

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J. J., Mendiondo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S. and Zehe, E. (2003) 'IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences', *Hydrological Sciences Journal*, 48(6), pp. 857–880. doi: 10.1623/hysj.48.6.857.51421.

Skinner, J. C., Bellerby, J. T., Greatrex, H. and Grimes, D. F. D. (2015) 'Hydrological modelling using ensemble satellite rainfall estimates in a sparsely gauged river basin: The need for whole-ensemble calibration', *Journal of Hydrology*. Elsevier B.V., 522, pp. 110–122. doi: 10.1016/j.jhydrol.2014.12.052.

Smith, L., Liang, Q., James, P. and Lin, W. (2017) '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), pp. 370–380. doi: 10.1111/jfr3.12154.

Spinsanti, L. and Ostermann, F. (2013) 'Automated geographic context analysis for volunteered information', *Applied Geography*. Elsevier Ltd, 43, pp. 36–44. doi: 10.1016/j.apgeog.2013.05.005.

Tauro, F., Selker, J., van de Giesen, N., Abrate, T., Uijlenhoet, R., Porfiri, M., Manfreda, S., Caylor, K., Moramarco, T., Benveniste, J., Ciraolo, G., Estes, L., Domeneghetti, A., Perks, M. T., Corbari, C., Rabiei, E., Ravazzani, G., Bogena, H., Harfouche, A., Brocca, L., Maltese, A., Wickert, A., Tarpanelli, A., Good, S., Lopez Alcala, J. M., Petroselli, A., Cudennec, C., Blume, T., Hut, R. and Grimaldi, S. (2018)

'Measurements and Observations in the XXI century (MOXXI): innovation and multi-disciplinarity to sense the hydrological cycle', *Hydrological Sciences Journal*. Taylor & Francis, 63(2), pp. 169–196. doi: 10.1080/02626667.2017.1420191.

Tiesi, A., Miglietta, M. M., Conte, D., Drofa, O., Davolio, S., Malguzzi, P. and 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(6), pp. 2619–2627. doi: 10.1109/JSTARS.2016.2520018.

Tkachenko, N., Jarvis, S. and Procter, R. (2017) 'Predicting floods with flickr tags', *PLoS ONE*, 12(2), pp. 1–13. doi: 10.1371/journal.pone.0172870.

Velez-Upegui, J. J. and Botero-Gutierrez, A. (2010) 'De Rezago En La Cuenca Experimental Urbana De La Quebrada San Luis , Manizales (Estimation of the Time of Concentration and the Lag Time At San Luis Creek Basin , Manizales)', *Dyna*, 165(February 2011), pp. 58–71. Available at: http://www.scielo.org.co/pdf/dyna/v78n165/a06v78n165.pdf.

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

Weerts, A. H. and El Serafy, G. Y. H. (2006) 'Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models', *Water Resources Research*, 42(9), pp. 1–17. doi: 10.1029/2005WR004093.

Weng, J. and Lee, B. (2011) 'Event Detection in Twitter', pp. 401–408.
Willems, P. (2009) 'Environmental Modelling & Software A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models', *Environmental Modelling and Software*. Elsevier Ltd, 24(3), pp. 311–321. doi: 10.1016/j.envsoft.2008.09.005.

Appendix:

A.1. Fusion approach of social media and authoritative data.

Demonstration:

Be

$$E[u_{t+1}] = E[\lambda v_{t+1} + (1-\lambda)s_{t+1}]$$
$$E[u_{t+1}] = E[\lambda v_{t+1}] + E[(1-\lambda)s_{t+1}]$$
$$E[u_{t+1}] = \lambda E[v_{t+1}] + (1-\lambda)E[s_{t+1}]$$
$$E[u_{t+1}] = \lambda u_{t+1}^{true} + (1-\lambda)u_{t+1}^{true}$$
$$E[u_{t+1}] = u_{t+1}^{true}$$

besides

$$Var[u_{t+1}] = Var[\lambda v_{t+1} + (1-\lambda)s_{t+1}]$$
$$Var[u_{t+1}] = Var[\lambda v_{t+1}] + Var[(1-\lambda)s_{t+1}] + \lambda^2(1-\lambda)^2 Cov[v_{t+1}, s_{t+1}]$$

but as the variables v_{t+1} and s_{t+1} are independent, $Cov[v_{t+1}, s_{t+1}] = 0$

$$Var[u_{t+1}] = \lambda^2 Var[v_{t+1}] + (1-\lambda)^2 Var[s_{t+1}]$$
$$Var[u_{t+1}] = \Sigma_{t+1}^{u} = \lambda^2 \Sigma_{t+1}^{v} + (1-\lambda)^2 \Sigma_{t+1}^{s}$$

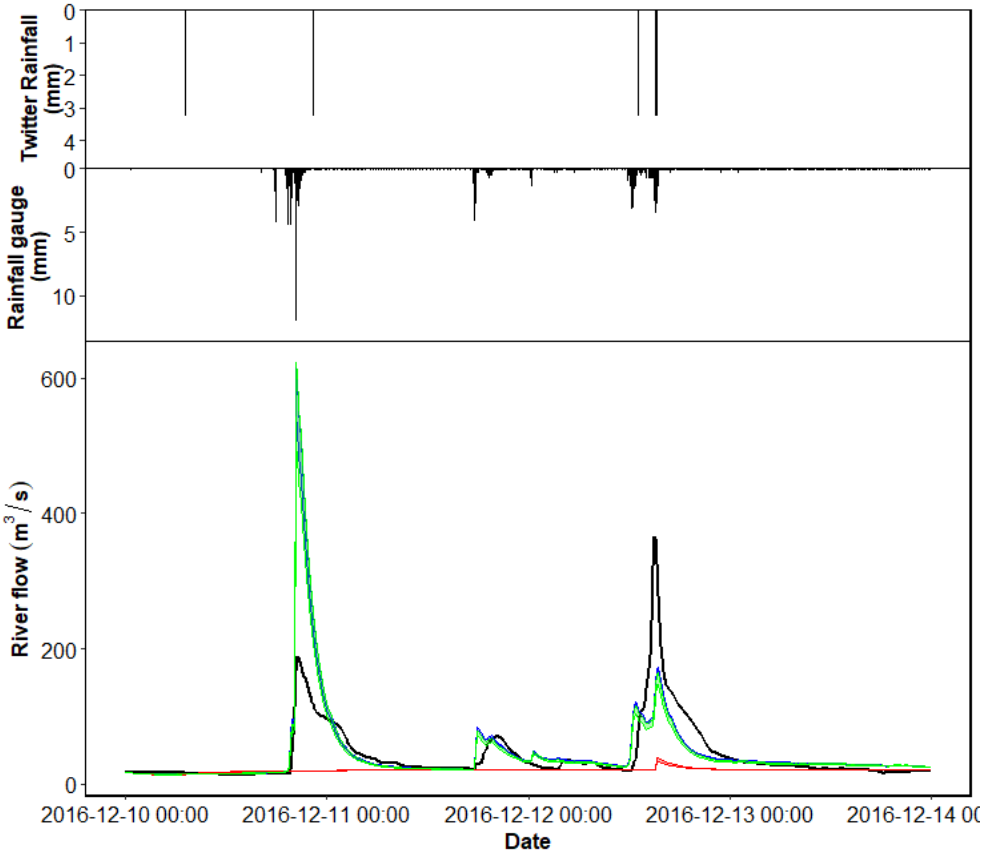
A.2. PDM Calibrations

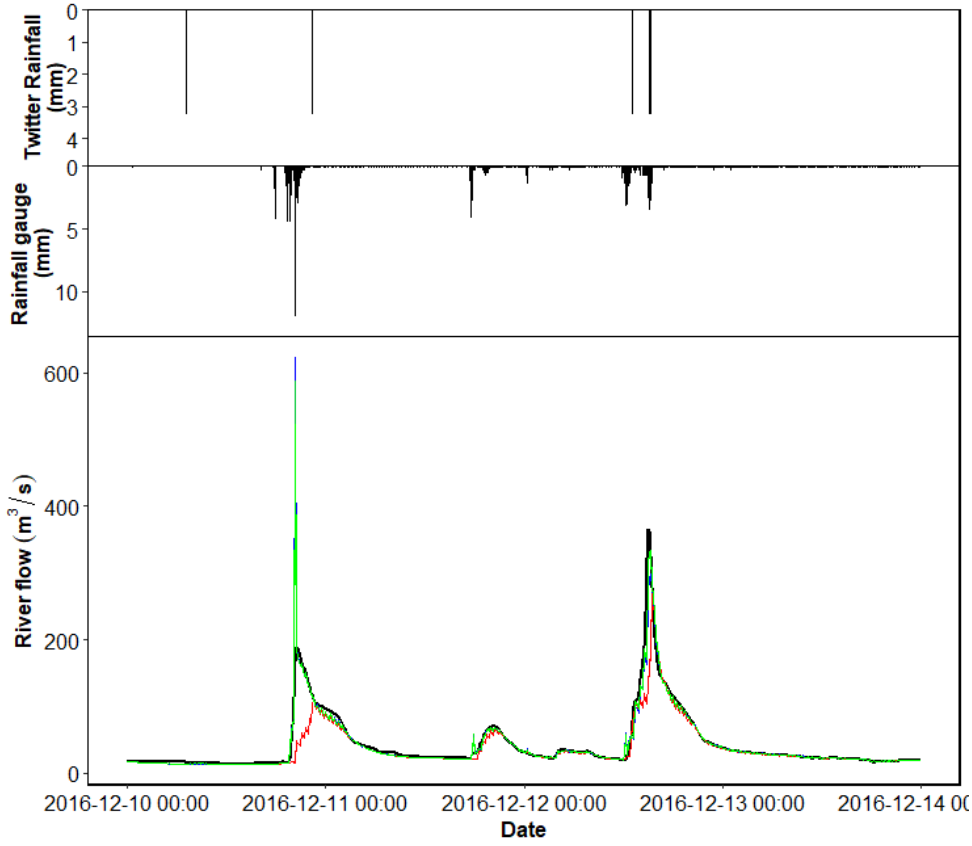
The following table presents a summary of the parameters that were obtained for the work of the thesis. It should be clarified at this point that calibrations made for the first part of the work were performed together several events, one after the other and were performed using an evolutionary algorithm, that is, it was a numerical calibration and not a hydrological one, maximizing the NSE. For the assimilation and fusion part, an optimization was performed for a single event, this calibration was manual and was thought more than to maximize the NSE in adequately representing the observed hydrograph.

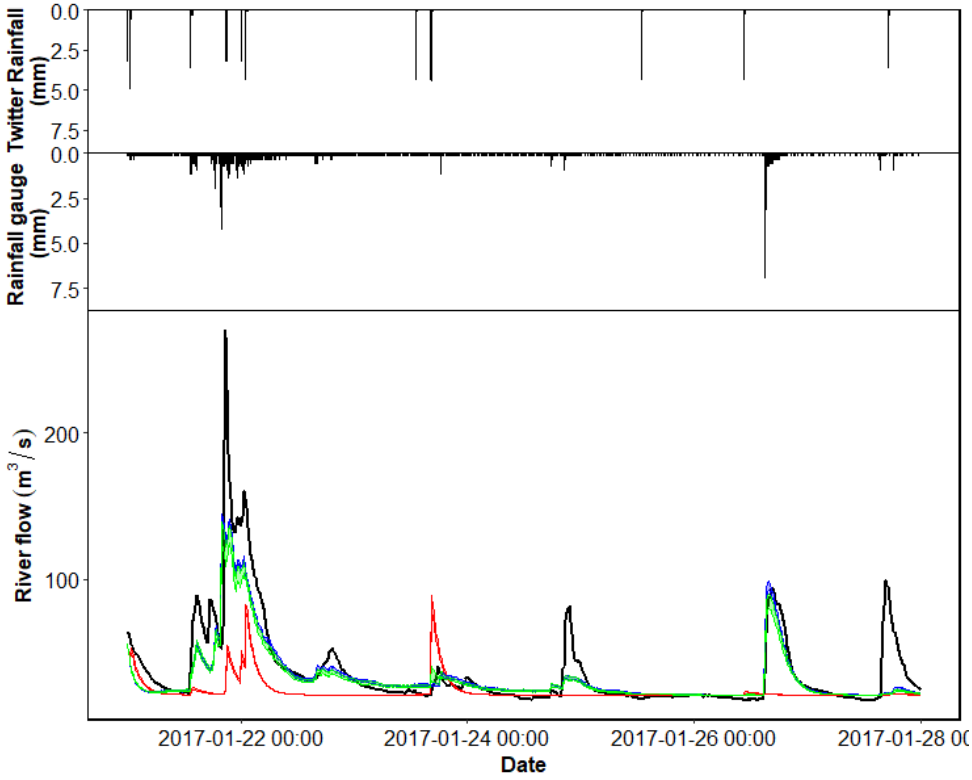
	C_{max}	C_{min}	b	b_e	k_g	k_b	k_1	k_2	NSE
Boa Esperança	1080	710	2.2	0	45	100	3900	9600	0.60
Cidade Tiradentes	75	55	1.0	0	28	110	50	200	0.38
Burgo Paulista	1130	740	3.0	0	65	1490	1000	5000	0.37
Int	1900	1000	1.0	0	15	1000	1840	1900	0.42
Max	260	0	1.0	0	22	250	30	250	0.63
Med	13000	6500	1.0	0	20	150	1200	1900	0.51
Ass_Fus	710	705	3.0	0	4450	0.3	0.01	1.9	0.68

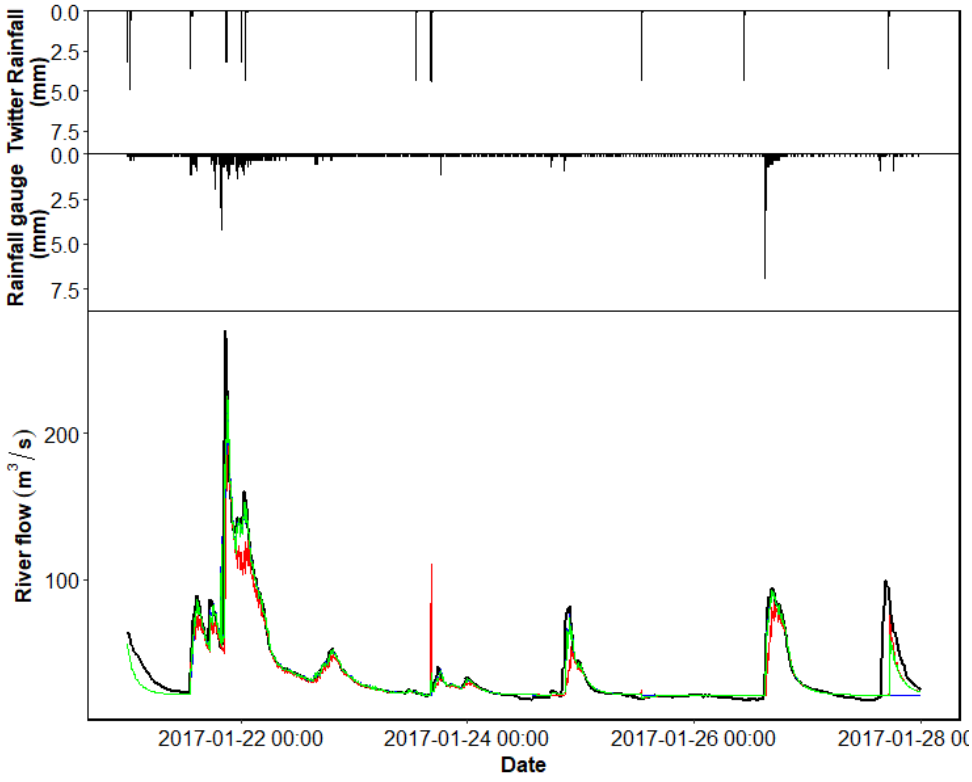
A.3. Other Assimilation events

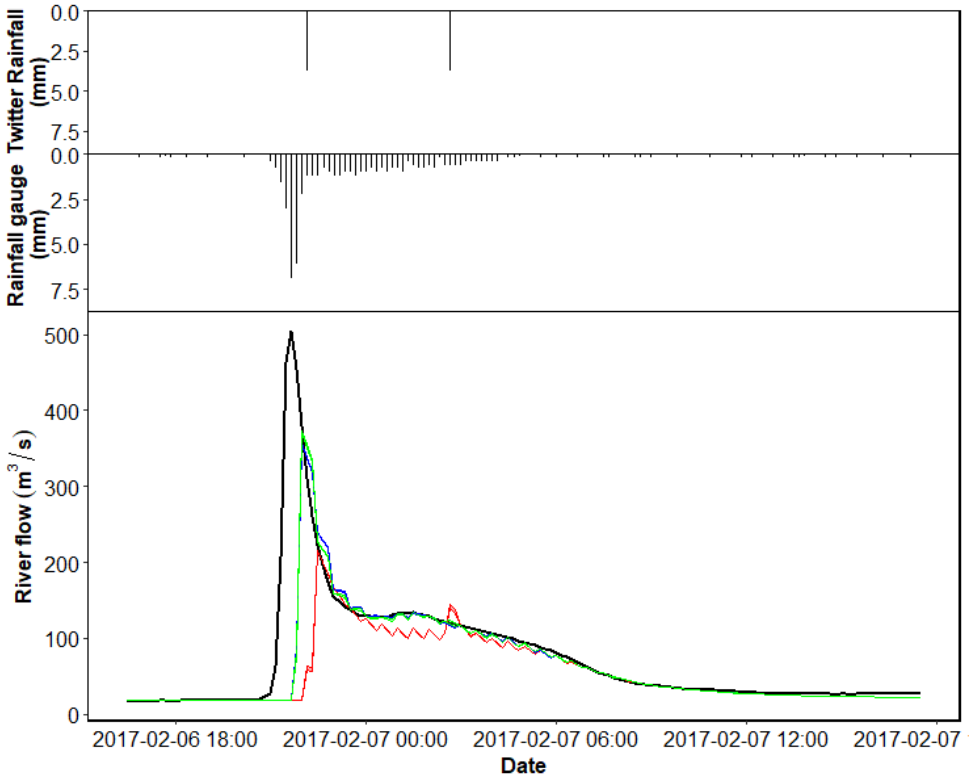
In Chapter 3, a simulated event was presented by assimilation and fusion. In that annex, the results show other simulated events under the methodology of that chapter.




Twitter Rainfall (mm)
0
1
2
3
4
Rainfall gauge (mm)
0
5
10
River flow (m^3/s)
600
400
200
0
2016-12-10 00:00
2016-12-11 00:00
2016-12-12 00:00
2016-12-13 00:00
2016-12-14 00
Date


Twitter Rainfall (mm)
0.0
2.5
5.0
7.5
Rainfall gauge (mm)
0.0
2.5
5.0
7.5
River flow (m^3/s)
200
100
2017-01-22 00:00
2017-01-24 00:00
2017-01-26 00:00
2017-01-28 00
Date


Twitter Rainfall (mm)
Rainfall gauge (mm)
River flow (m^3/s)
Date
2017-01-22 00:00
2017-01-24 00:00
2017-01-26 00:00
2017-01-28 0(


Twitter Rainfall (mm)
Rainfall gauge (mm)
River flow (m^3/s)
Date
2017-02-06 18:00
2017-02-07 00:00
2017-02-07 06:00
2017-02-07 12:00
2017-02-07