Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

Natural Hazards and Earth System Sciences
Discussions

Climate risks, digital media, and big data: following communication trails to investigate urban communities' resilience

Rosa Vicari[1], Ioulia Tchiguirinskaia[1], and Daniel Schertzer[1]

[1]Hydrology Meteorology and Complexity Laboratory, École des Ponts ParisTech, Marne-la-Vallée, Champs–sur–Marne, 77455, France

Correspondence: Rosa Vicari (rosa.vicari@enpc.fr)

Abstract. Nowadays, when extreme weather affects an urban area, huge amounts of digital data are spontaneously produced by the population on the Internet. These "digital trails" can provide an insight on the interactions existing between climate related risks and the social perception of these risks. According to this research "big data" exploration techniques can be exploited to monitor these interactions and their effect on urban resilience. The experiments presented in this paper show that digital research can bring out the most central issues in the digital media, identify the stakeholders that have the capacity to influence the debate and, therefore, the community's attitudes towards an issue. Three corpora of Web communication data have been extracted: press news covering the June 2016 Seine River flood; press news covering the October 2015 Alpes-Maritimes flood; tweets on the 2016 Seine River flood. The analysis of these datasets involves an iteration between manual and automated extraction of hundreds of key terms, network representations based on key terms co-occurrences, automated cluster visualisation based on adjacency matrix, and profiling of social media users. Visual observation of the network coupled to quantitative analysis of its nodes and edges allow obtaining an in-depth understanding of the most prominent topics and actors, as well as of the connections and clusters that these topics and actors tend to form in the journalistic sphere. Through a comparison of the three datasets, it is also possible to observe how these patterns change over time, in different urban areas and in different digital media contexts.

1 Introduction

This paper presents a study on how digital media represent urban resilience during extreme weather and in the following weeks. The approach proposed here aims at exploiting the huge amount of web data that are produced during and after a climate risk. Analysis of "big data" corpora drawn from digital media has been already employed in research fields related to communication on climate risks. A series of studies analyse and compare the existing controversies on climate change in the web sphere (Niederer, 2013 ; Rogers and Marres, 2000; Chavalarias, 2015; climaps.eu) . Other research works investigate the

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

Natural Hazards and Earth System Sciences
Discussions

use of social media during crisis due to climate hazards (Palen et al., 2010; Morss et al., 2017; Bruns et al. 2012; Lanfranchi et al., 2014 ; Gaitan et al., 2014 ; J.C. Chacon-Hurtado, 2017).

The originality of our approach lies in the fact that it is framed in the context of research on urban resilience assessment. For this reason, it aims at contributing to the comprehension of the interactions that exist among different urban resilience drivers. According to Mangalagiu, one of the challenges of the 21st century is giving the due attention to the "complex and causal linkages between human, technological, environmental and global biophysical systems" (Mangalagiu et al., 2012). In our view, quantifiable variables facilitate the investigation of the relations between different physic-environmental and socio-economic components of urban systems. Furthermore, quantitative indicators are helpful to cross-compare different locations and time points. A huge variety of quantitative metrics are proposed in the literature on different resilience assessment approaches (Cutter et al. 2008; 2010; UN/ISDR, 2008; Resilience Alliance, 2010; Keating et al., 2014). In this paper, the focus is put on those quantitative data that can be used to investigate the social representation of climate risks. Digital media are a source of quantitative information that can be automatically extracted and analysed through computer-aided exploration techniques such as advanced text mining and network representation. A quantitative analysis of digital communication patterns easily leads to an evaluation of how different space and time variables affect these patterns. It also facilitates the analysis of how communication trends and other resilience drivers (e.g. an environmental factor) mutually influence each other. When these correlations exist, they are a necessary basis to understand how social perception of climate risks affects urban resilience.

Besides examining these methodological challenges, we also intend to contribute to the comprehension of the social perception of urban resilience to climate risks through digital media. We present an analysis of three datasets in Sect. 3, 4 and 5: the press news covering the June 2016 Seine River flood; the press news covering the October 2015 Alpes-Maritimes flood; the tweets on the 2016 Seine River flood. We discuss an initial analysis of the most prominent topics and actors, as well as of the thematic subsets and connections that characterise each dataset. We also compare the three datasets and reflect on how the debate changes over time, in different urban areas and in different media contexts.

2 Benefits and constraints of digital media analysis based on advanced text mining techniques

2.1 Method

Stakeholders' perception of a controversial issue is a community characteristic (and a social impact when change occurs) that can be analysed through surveys, meetings and interviews. Surveys provide information on population attitudes at aggregated level, while interviews and meetings (e.g. focus groups) provide insights about how and why particular attitudes are developed at individual level or at small group level. However, big data exploration techniques allow getting beyond the dichotomy between the aggregated structure and the individual component when studying social connections (Latour, 2012). The following examples of analysis of digital texts illustrate how computer aided exploration techniques allow gaining insight both on the intensity and the quality of web communications. Indeed, thanks to an automated big data exploration tool such as Gargantext (Chavalarias and Delanoe, gargantext.org, 2017), it is possible to quickly navigating through huge masses of digital information

and following the connections among cultural contents (e.g. articles, blog posts, tweets), popular topics, and the names of public figures or organisations.

Gargantext allows extracting, automatically as well as manually, a list of key terms from a corpus of texts. This list of terms is then used by Gargantext to compute a network representation. A "network" or "graph" is a mathematical structure that represents a collection of interconnected objects. Network representation can be used to visualise which key terms co-occur in the same meaning unit (e.g. a press article or a social media post) and with what frequency. The key terms are represented by the "nodes" of the network, and the co-occurrency relation among these nodes is represented by the network "edges". Gargantext allows computing network representations that remain stable even when few documents are deleted from the corpus or few nodes are removed from the network. The nodes are assembled in cohesive subsets through a clustering algorithm, more specifically through Louvain modularity [1]. A weight is assigned to each node and to each edge: in the first case it corresponds to the node centrality, i.e. the number of edges associated to that node; in the second case it corresponds to the frequency of a connection between two nodes.

2.2 Data

Online press news and social media posts are second hand data. Indeed, as it is discussed by Venturini (Venturini et al. 2014), the researcher cannot directly control the production of these data and he should question himself about their production context and process. For instance, news publication follows a set of journalistic values, the so-called " newsworthiness" (Boyd, 1994), that determine if and how much a story is important for a media outlet and its audience. An example of news value is "the greater the drama, the greater its prominence in conversation": this kind of news, that is expected to get their audience talking, is considered more worthy than others. These values are translated in a hierarchy of information that will guide news programming. In the case of social media, the problem of the digital divide (i.e. internet access, skills and usage inequality) leads the researcher to consider the socio-demographic characteristics of social media users. For example, when analysing tweets, it should be taken into account that in France the population over 45 years is not well represented by a sample of Twitter users, while the population with a university degree is overrepresented. Indeed, in France in 2017, 16 % of Twitter users were over 45 years old and 40 % of the users had a university degree (excluding BTS) (blogdumoderateur.com). According to Venturini “digital traces are not natural items but artefacts created in a specific environment and with specific objectives”. However, this doesn’t reduce their value, since their publication process can be a source of information on the social representation of reality, for instance the media representation of climate related threats. Even though digital communications allow a more direct observation of social phenomena, these data need to be contextualised and interpreted.

The Seine River flood that occurred in June 2016 was picked as a case study because of its prominent media impact. According to a search of French press articles on Europresse (europresse.com), on the 3rd of June 2016 the press coverage of

[1]The Louvain method for community detection is used to maximise the network modularity. A network with high modularity has dense edges between the nodes within modules and sparse edges between nodes belonging to different modules. The modularity maximisation involves two stages: first the small clusters are detected, then the nodes that belong to the same cluster are aggregated and a new network is produced whose nodes are the clusters. These operations are repeated until a maximum of modularity is reached and a clusters hierarchy is built.

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

Natural Hazards and Earth System Sciences Discussions

this flood event reached a peak of 310 articles published in one day (corresponding to 29437 terms, as illustrated by Fig. 1a). This is a remarkable figure considering that the same French media published 591 articles in one day on Trump's victory on the 8th November 2016. Media visibility influences public opinion, hence stakeholders' attitudes towards risks and disasters, and related resilience policies or projects[2]. Therefore this flood event is worth exploring from the urban resilience perspective.

On the 3rd and 4th of October 2015 extreme rainfall caused rivers flood in the Alpes-Maritimes Department. Cannes, Antibes, Vallauris, Biot et Mandelieu-la-Napoule were the most affected municipalities. The press coverage of this flood event was more limited (286 articles over five months) in comparison to the Seine River flood (761 articles over five months), even though the first flood took a huge toll on human life with 20 deaths. This is probably due to the higher newsworthiness that events in the French capital have in comparison to those occurring in the rest of the country. Another reason is that the economic risks related to a flood event in Paris region are extremely high (OECD, 2014) since it is a densely populated area that represents a third of the national economy, and where companies' headquarters, national and international institutions are located; furthermore, it is an important transportation node and one of the first tourist destinations in the world.

As mentioned above, the press select and prioritise the news, hence it defines the prominent topics and their organisation in thematic clusters. In this way, editors and journalists obviously influence the public perception of an extreme weather event, even though a two-ways relationship exists between the press and the audience. This bond has been progressively fading since access to information has hugely increased in terms of variety and quantity, as a consequence of different factors, among others the development of public relations by non-journalistic organisations and the pervasive role of the Web sphere (Bucchi, 2013; Trench, 2008). In this context, a corpus of texts published on a social network deserves to be analysed and compared to the press articles corpus in order to have insight of the public perception of a flood event beyond the borders of the journalistic arena. The third dataset analysed in this paper is a corpus of tweets covering to the Seine River of June 2016. The choice to focus on Twitter is due to the fact that public authorities and citizens are increasingly using Twitter during natural disasters as a two-way early warnings and information channel. According to Bruns and Liang (2012), Twitter is particularly suitable for crisis communication, indeed its "flat and flexible communicative structure" allows any visitor to access public tweets: users that are not yet followers of the account that disseminates the information on a crisis event or even visitors that are not registered on Twitter. Furthermore, hashtags allow any visitor to search for tweets on a specific topic. Such communicative structure facilitates fast, large-scale collection of information.

3 Press coverage of Seine River flood in 2016

A corpus of 761 articles was first selected through Europresse archives on the base of the following criteria: French press articles published from 15/05/2016 to 15/10/2016, with a title including the terms "crue" or "inond*" and "Seine" or "Ile-de-France" or "Paris" or "Région Parisienne". The corpus was then analysed through the open source software Gargantext.

[2]Media contribute to our perception of reality (including risks and disasters) through selection and omission of information. For example the UK government reassurance campaign contributed to spread the mad cow disease, which resulted in millions of animals being destroyed and the deaths of 226 people. Furthermore, humans respond more forcefully to emotional appeals than to facts like in the case of the Indian Tsunami earthquake (2004): images and stories from tourists and the extreme language used by the media led to a higher donors' response compared to other disasters with more victims.

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

3.1 Aggregated analysis

A first histogram was created to illustrate the intensity of press coverage in terms of number of terms per day and how it evolved over one month (see Fig. 1a). Due to a very high inflow of publications in few days, the information in Fig. 1a) is presented in a semi-log plot to make the information more clear. The press coverage peak was reached on 3rd of June 2016: on the same day the Seine River discharge was highest.

The following step of the analysis consisted in selecting terms that correspond to a range of "flood resilience solutions" (see Table S1). "Flood resilience solutions" is understood here in a large sense, as any kind of solution that is implemented to reduce flood damages. 316 key terms were extracted from two corpora of articles: the corpus of articles covering the 2016 Seine River flood and the corpus of articles covering the 2015 Côte d'Azur flood. For each corpus, a first list was automatically established by Gargantext algorithms on the base of their occurrency, compared to other occurrences that characterise all Gargantext database corpora, as well as on the base of the co-occurrencies that characterise this specific corpus. These lists were then manually refined on the base of the relevance of each term (as a solution aimed to cope with a flood event) and finally merged. This list was used to analyse the occurrency and co-occurrency of the key terms in each corpus, except for the terms with less than five occurrences that were not considered in order to highlight the most frequent topics.

As shown in Fig. 1a, the key terms list allowed a comparison between the total number of terms per day with the number of key terms per day referring to the topic of flood resilience solutions. The comparison between the overall corpus with the sub-group of key terms (referring to flood resilience solutions) aimed to monitor how the quality of the contents evolved over time. The portion of flood resilience solutions key terms varies between 0% and 10%. Further insights can be obtained through a comparison with other histograms based on different corpora of texts. Figure 1b shows a similar histogram based on press articles covering the 2018 Seine River flood (21/01/2018 – 20/02/2018). In this case the portion of key terms referring to flood resilience solutions is reduced, since it varies between 0% and 5%. A comparison with two other corpora (a corpus of news covering a flood event in a different location and a corpus of tweets) will be presented in the next sections.

3.2 Network representation

After analysing the key terms frequency in an aggregated manner, the second step of the analysis was aimed at representing a complex communication network highlighting the connections that exist among different topics and public figures/organisations. The network was based on an adjusted version of the terms list referring to flood resilience solutions. Indeed new key terms corresponding to public figures and organisations involved in flood risk management, as well as affected infrastructures and properties were added to the initial list. Furthermore, synonyms ("habitants" and "riverains"), declensions of terms (e.g. "scientifique" and "scientifiques") and equivalent forms (e.g. "Établissement Public Seine Grands Lacs" et "EPTB Seine Grands Lacs") were merged. Once the terms list was refined, a network of 254 nodes and 445 edges (Figure 2) was generated on the base of the conditional distance between two key terms, i.e. the incidence of co-occurrences of two terms in the same articles. The size of a node represents its centrality that measures the number of node connections, in other words how frequently a term is associated to flood resilience related topics or risk management actors. Each subset or "cluster" cor-

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

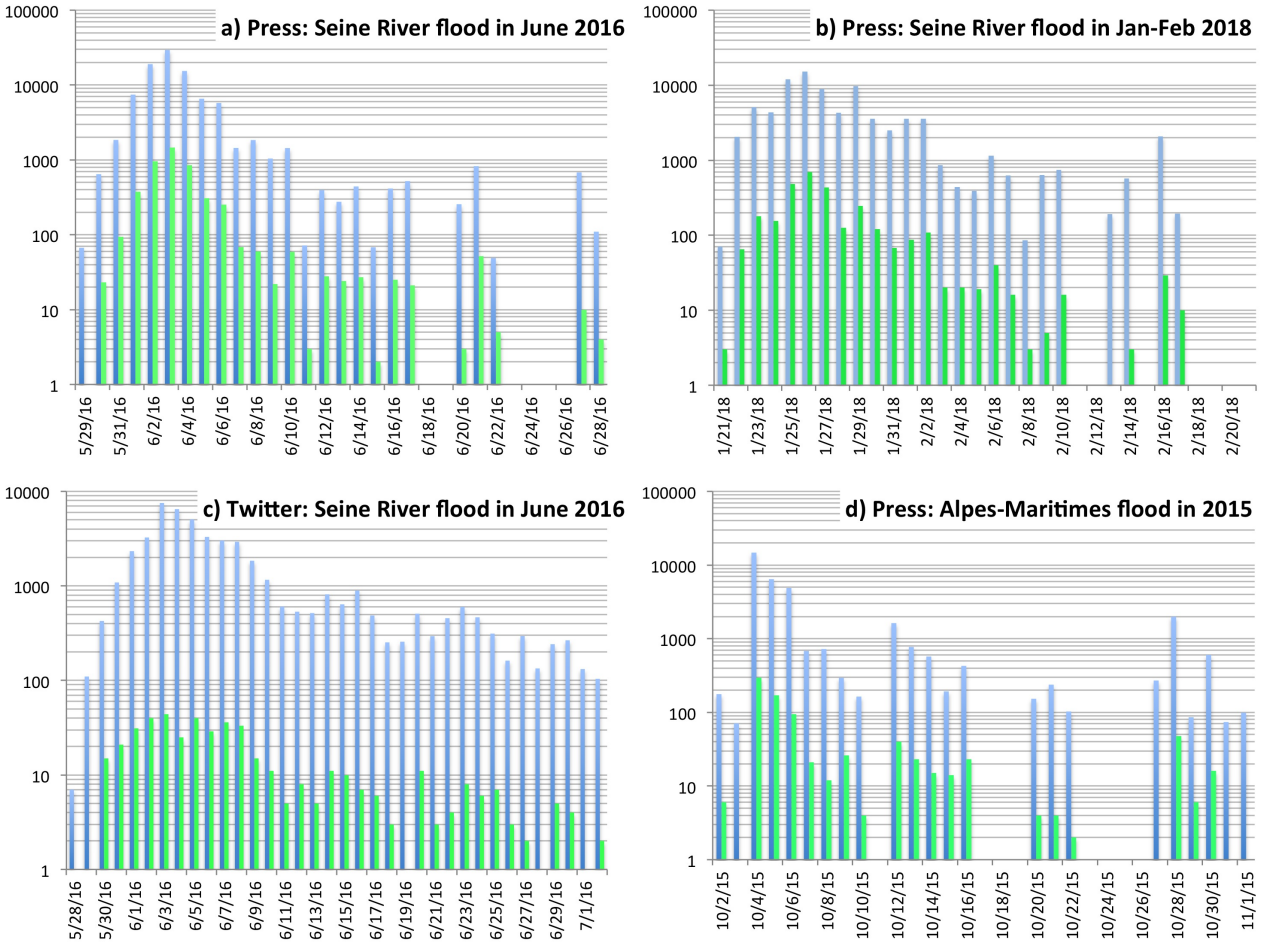

Figure 1. Comparison between the total number of terms per day (blue) and the number of terms per day referring to flood resilience solutions (green) in a semi-log plot (based on four different datasets).

responds to a group of key topics and key actors that frequently appear in the same article. Gargantext allows zooming in and out the network, moving around, selecting a node and highlighting its connections.

3.3 Visual observation of the network

Similarly to the network representation presented by Venturini (Venturini et al., 2014), a first analysis of the results can be based on visual observation of the network. By navigating the network, it is possible to observe very central, i.e. strongly connected, topics (“alerte” with 51 connections, “prise de conscience” with 25 connections, or “état de catastrophe naturelle” with 29 connections) and organisations (e.g. “SNCF” with 46 connections or “Louvre” with 52 connections) or public figures (e.g. “Manuel Valls” with 33 connections or “François Hollande” with 26 connections). Network navigation also allows examining

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

Natural Hazards and Earth System Sciences
Discussions

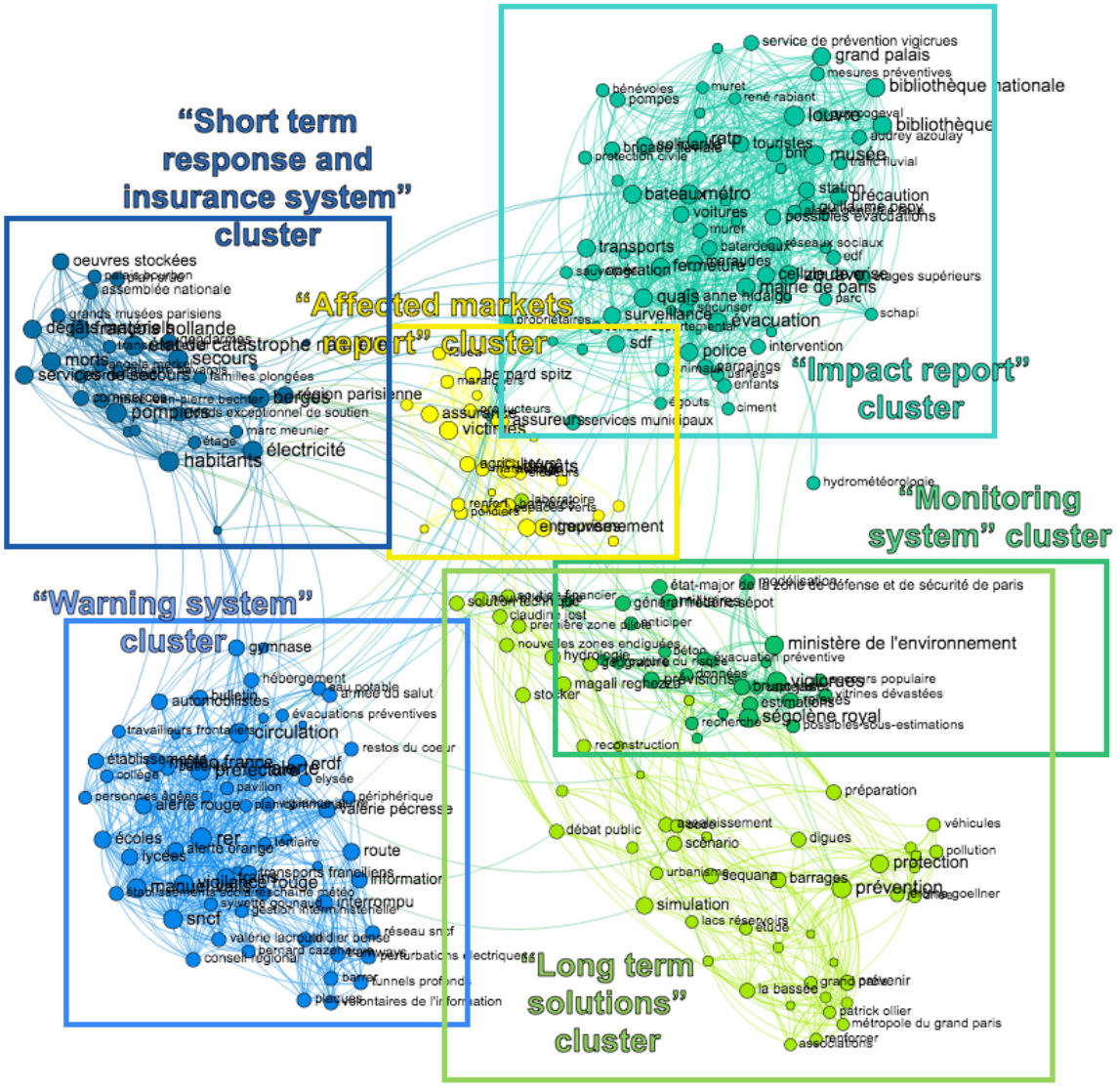

Figure 2. Complex network computed on the base of the co-occurrence of two key terms (the network nodes) in the same article: the nodes are assembled in clusters through modularity maximization. The network navigation (zooming in and out, moving around, selecting the nodes to highlight their connections) allows observing what are the most central topics and actors and if they are connected.

which topics the identified actors can be associated to (e.g. "Métropole du Grand Paris" and "désimpérméabilisation" co-occur), and which organisations/public figures assemble in clusters of stakeholders (for instance "Ségolène Royal", "Ministère de l'Environnement", "Bruno Janet" – Head of modelling centre at Vigicrues – and "IRSTEA"). By zooming out to visualise the entire network (Fig. 2) it is possible to identify six main subsets that correspond to six clusters of debate topics and actors.

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

Each cluster is assumed to be associated to a macro-theme, i.e. an expression that sums up to which resilience management area the key terms, included in the same cluster, refer to. The colour of the cluster is automatically defined by Gargantext algorithm, while the authors chose the name of each macro-theme:

1. The green macro-theme "Monitoring system" that brings together topics such as "sensors", "data", "modelling", "observation", "estimates", "possible under-estimations", as well as actors that include national authorities (the French Ministry of the Environment, the Minister of the Environment, Vigicrues) and researchers (IRSTEA, Vazken Andreassian);

2. The light blue macro-theme "Warning system" that covers key terms such as "red vigilance", "warning", "information", "municipal plan", "shelter", "traffic" and actors, including government representatives (the French Minister of the Interior, the Prime Minister), a law enforcement institution (Prefecture), a transport company (SNCF), the national meteorological service (Météo France), an electric utility company (ERDF), vulnerable population (elders);

3. The turquoise macro-theme "Impact report" that gathers topics such as "rescue", "closure" or "evacuation" of "museum", "station", "hospitals", "transports", "boats", "cars", "electric network", "zouave" and actors such as local authorities (the Mayor of Paris, Departmental Council), cultural institutions ("Louvre", "Grand Palais", National French Library), rescue services ("police", "volunteers", "civil protection", "René Rabiant"), affected population ("tourists", "tenants");

4. The dark blue macro-theme "Short term response and insurance system" that brings together "state of natural emergency", "fire brigade", "rescue services", "stored art works", "François Hollande", "the National Assembly", "inhabitants";

5. The yellow macro-theme "Affected markets report" with topics like "damages", "repair", "farms", "agricultural holdings", "companies", "market gardening", "insurance", "diagnosis" and actors that include the victims ("farmers", "producers", "victims", "shipowners"), organisations representing them ("farmers' union") and insurers ("Bernard Spitz");

6. The light green macro-theme "Long term solutions" includes topics like "awareness raising", "prevention", "soil sealing reduction", "La Bassée pilot project", "memory", "preparation", "first pilot area", "new structure", "public debate", "retention basin", "reinforce", "simulation"; and actors such as local authorities ("Regional Department of Environment and Energy", "Métropole du Grand Paris", "associations", "public territorial agencies", the mayors of Saint-Maur-des-Fossés and Rueil-Malmaison, the Hydrology Director at the Public Agency of the Seine Grands Lacs Basin).

3.4 Quantitative analysis of the nodes and the edges

Further information on the network can be obtained through a quantitative analysis: the values corresponding to the nodes degree and to the edges weight can be obtained through Gephi (an open source network visualisation software, gephi.com). As it is shown in Figure S2.1, the most central nodes (degree > 30) concern warning and emergency management, especially management of public infrastructures, indeed they are localised in the two biggest clusters (the light blue cluster and the turquoise cluster). For instance, the most central node is "RER" (the Paris region commuter rail service) with 55 connections:

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

this figure highlights that the press focuses on the immediate flood impacts, affecting a relevant portion of the population in their daily life. Concerning the actors, law enforcement and rescue services appear as the most central actors in the debate.

In Figure S2.2 the most frequent co-occurrences (i.e. the highest edges weight) are identified with the label of the corresponding couple of nodes. The figure highlights that frequently coupled terms are: terms that concern the same risk management field (e.g. rescue services); terms that concern the same area of affected activities (agriculture or transports); terms that concern infrastructure located in the same flood prone area; two terms that refer to an action and the object to which this action is directed (e.g. « clubs » and « closing ») or a subject and an action performed by it (e.g. « trains » and « normally circulating »).

4 Press coverage of the Alpes-Maritimes flood in 2015

286 articles were extracted through Europresse. We used the following criteria in the selection: French press articles from 15/09/2015 to 15/02/2015, with a title including the terms “crue” or “inond*” and a reference to at least one of the locations affected by the flood [3].

4.1 Aggregated analysis

Since the total amount of news for this case study is smaller, the peak of published terms per day (see Fig. 1d) is obviously reduced (108 articles on the 4th of October, corresponding to 14772 terms). As shown in Fig. 1d, after the maximum peak (on the 4th of October) the number of terms per day decreases less progressively than in the Paris region case studies (Fig.1a, 1b). This characteristic of the curve can be explained by the limited media visibility of the region, another reason might be a different evolution of the rivers flow (a very slow decrease of water levels in the case of the Seine River). A peculiarity in Fig. 1d is that there are two small peaks on the 12th of October (with the news of two rescued persons and two victims) and then on the 28th of October (with the news of the mayor of Cannes calling for help to the movie celebrities).

In order to assess the visibility that was given to the debate on flood resilience solutions, the term list, that we formerly created, was used to identify a subset in the second corpus (Fig. 1d). As in the previous experiment, terms with less than 5 occurrences were excluded. The portion of terms referring to flood resilience solutions varies from 0% to 9%, which is close to the percentages of the first case study.

4.2 Network representation and visual observation

In a second stage, the key terms list was enriched with terms corresponding to public figures and organisations, affected infrastructures and properties. After merging synonyms, declensions of terms and equivalent forms, the list of key terms was

[3]The titles of the articles selected for the second case study have a title referring to at least one of the following locations: “Alpes-Maritimes” or “Cannes” or “Antibes” or “Vallauris” or “Biot” or “Mandelieu-la-Napoule” or “Bouches-du-Rhône” or “Var” or “Vaucluse” or “Drôme” or “Siagne” or “Bragues” or “Fréjus” or “Reyran” or “Vallauris-Golfe-Juan” or “Cagnes-sur-Mer” or “Le Cannet Mougins” or “Nice” or “Roquefort-les-Pins” or “La Roquette-sur-Siagne” or “Théoule-sur-Mer” or “Valbonne” or “Villeneuve-Loubet” or “Les Arcs” or “Brignoles” or “Cabasse” or “Callas” or “Camps-la-Source” or “Flassans-sur-Issole” or “Côte d’Azur” or “sud-est” or “Flayosc” or “Forcalqueiret” or “Fréjus” or “Méounes-lès-Montrieux” or “La Motte” or “Néoules” or “Puget-sur-Argens” or “La Roquebrussanne” or “Saint-Antonin-du-Var” or “Saint-Raphaël” or “Le Thoronet” or “Trans-en-Provence”.

used to generate a complex network based on conditional distance (Fig. 3) with 104 nodes and 676 edges. Because of the smaller key terms occurrence, the network is less dense. However it is still possible to identify four macro-themes:

1. A turquoise subset of terms referring to the macro-theme "Emergency management";
2. A green subset of terms related to the macro-theme "Monitoring system and prevention";
3. A light blue subset of terms related to the macro-theme "Reconstruction";
4. A violet subset of terms related the macro-theme "Impact record".

Except for the macro-theme « Impact record », the other themes are not equivalent to those identified in the first network. This is indicative of a certain variability between two cases of floods in terms of the resilience levers that are covered by the press.

4.3 Quantitative analysis of the nodes and the edges

The nodes centrality (Fig. S3.1) highlights that « deaths », « missing » and « victims » are central nodes. Furthermore, the following actors are relevant: the government (including the Prime Minister and the French President), the inhabitants, the rescue services (the volunteers, the police, the fire brigade), Cannes mayor, celebrities and insurers. This is probably due to the high number of victims that led the attention to the affected populations and to those actors who were involved in rescue and compensation payment. National institutions were involved in allowing the necessary procedures to attract funding for the reconstruction; furthermore, government representatives spoke to the press in commemoration of the victims. The Mayor of Cannes, David Lisnard, was also in the spotlight, when he asked to the film stars to financially support his city.

The values corresponding to the edges weight (Fig. S3.2) show that the most frequent co-occurrences are conform to the trends described in the previous case study: some couple of terms concern the same area of flood resilience management (such as forecasts, impact report, awareness raising and prevention or compensations for the victims); other couple of terms can be identified as an action and a related object (e.g. an event cancellation). Two actors that are frequently coupled are CNRS and Météo France: indeed both organisations are mentioned as partners of HyMeX (a project on understanding, quantifying and modelling the hydrological cycle in the Mediterranean). Furthermore, experts from these two organisations have received considerable media attention in relation to debate on the flood causes (climate evolution and urbanisation) and on how flood resilience can be enhanced through more accurate forecasts. Other researchers from IRSTEA and University of Avignon have been questioned by the press about the importance of awareness raising (“sensibilisation”).

5 Twitter coverage of the Seine River flood in 2016

5.1 Extraction of the dataset

The corpus of tweets covering the Seine River flood of June 2016 was extracted through “Twitter Advanced Search” (twitter.com/search-advanced). The selection criteria were a time span (from 28/05/2016 to 2/7/2016) and relevant hashtags (“#crue” or “#crueparis”

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

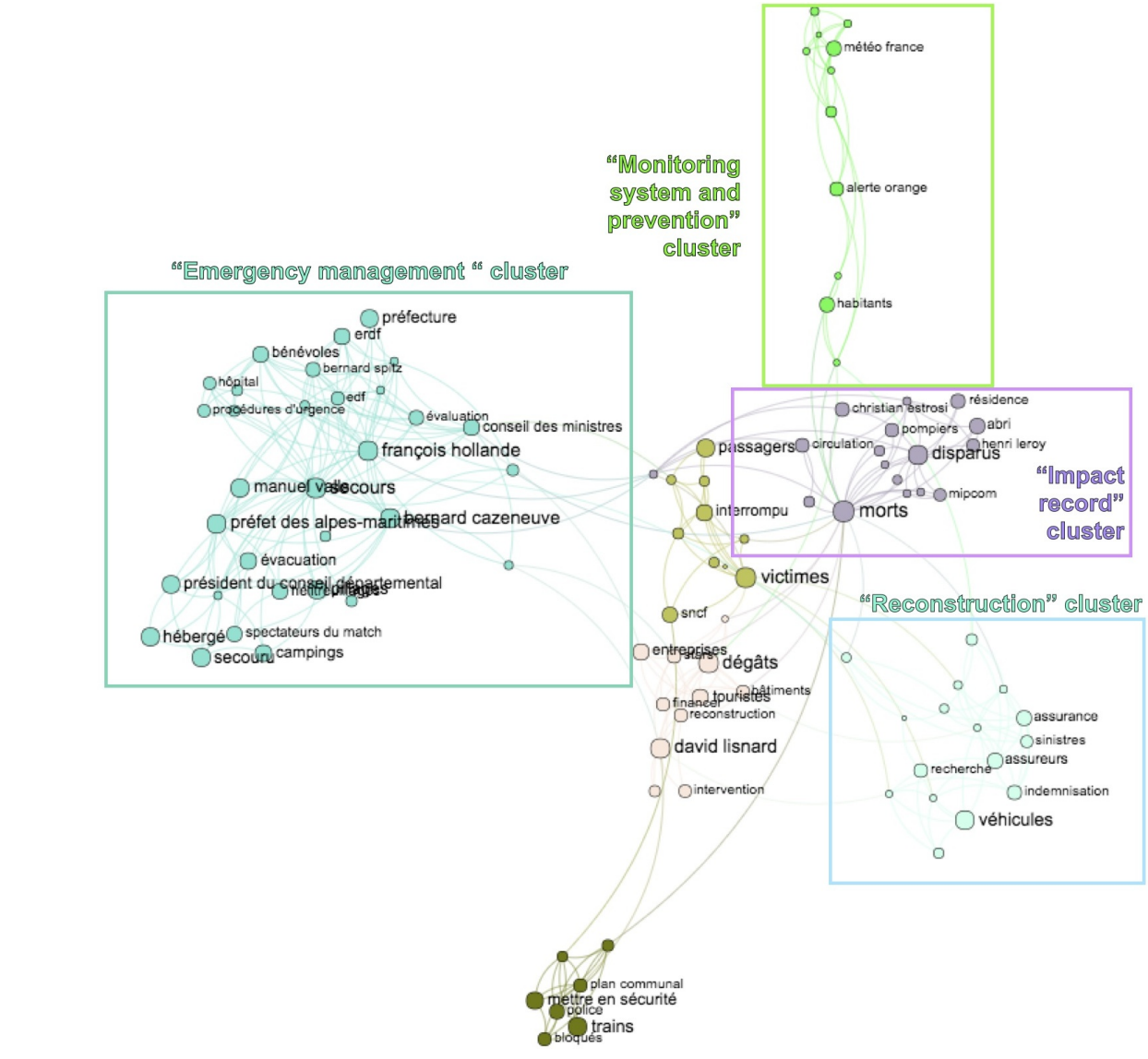

Figure 3. Network representation of the press articles on the 2015 flood in the Alpes-Maritimes Region.

and "#crueseine" or "#inondation" or "#inondations" or "#pluies" or "#Seine"). As it is suggested by Bruns et Liang (2012),hashtags allow focusing on those tweets where terms related to a flood are marked as important information. Geo-location was not used as a criterion because the sample of tweets would have been very small: few users provide such detail with their tweet (Bruns et Liang, 2012). The corpus was then refined by deleting duplicates. In order to facilitate a comparison with the previous corpora, the tweets that mentioned locations outside the Paris region were deleted as well. As a result, the final corpus included 7984 tweets.

As in the previous case studies, an aggregated analysis of the sample of tweets was made through Gargantext: Figure 1c highlights that a terms peak occurred on the 3rd of June 2016 when the Seine River discharge was highest. Like in the case of the press coverage of the Seine River flood, a correlation exists between the social representation of a meteorological event and its physical-environmental impact. However it should be noticed that the maximum peak is not significant as in the histograms based on the press corpora.

5.2 Aggregated analysis

A list of terms related to flood resilience solutions allowed highlighting that a minority of terms referred to these solutions. The list (first row in Table S4) was established on the base of the previous term list and new relevant terms that are specific to Twitter jargon and include also English terms. The green columns in Fig. 1c represent the portion of terms referring to flood resilience solutions that varies between 0% and 3,5% of the total amount of terms published per day. The occurrence of key terms was so limited that it was not possible to represent a significant network, even after extracting terms referring to stakeholders and affected infrastructures. Indeed the majority of tweets included in the corpus only describe the flood event and its impact. This is due to the limited numbers of characters that are allowed for each tweet (up to 140 in 2016) and that make the information essential, unless the tweet includes a link to an external webpage (the contents of which can't be automatically analysed through Gargantext). Twitter indeed is conducive to a fragmented communication.

Even if the thematic patterns of tweets cannot be represented through a network, we can push forward an aggregated analysis. We can identify thematic groups of key terms (Table S4) and their frequency as it was done by Vieweg et al. (2010). As shown in Fig. 4, a major portion of key terms (3143 occurrences) consists in purely factual information, with references to the time and location of the flood event. According to this result, Twitter might be primarily used as a mean to disseminate warnings. The category "flood resilience solutions" (1420 occurrences) includes a relevant portion of key terms, as well as the "stakeholders" category (1060 occurrences), however they are less frequent than terms describing the weather event (1506 occurrences) and its impact (1644 occurrences). On the contrary, the debate on the causes (72 occurrences) and risks (320 occurrences) is of little account.

5.3 Users' profile and behaviour

Besides this thematic analysis, the same sample of tweets can be used to investigate the behaviour of its users, their profiles and their interactions. This is the specific added value of social media data. We followed an approach which draws on the maps presented in the Climaps platform[4]. We identified the most active, liked and retweeted users in the sample. We consider as the "most active users" those accounts that published more than 10 tweets in one month: we counted 59 users with this characteristic. The "most liked users" are those accounts that received more than 50 likes per tweet in one month: 43 users have this characteristic. We name those accounts that received more than 50 retweets per tweet in one month as the "most retweeted users": we counted 58 users with this characteristic.

[4] "Reading the state of climate change from digital media" (climaps.eu, last access: 05/07/2018): http://disq.us/t/1gj2hci

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

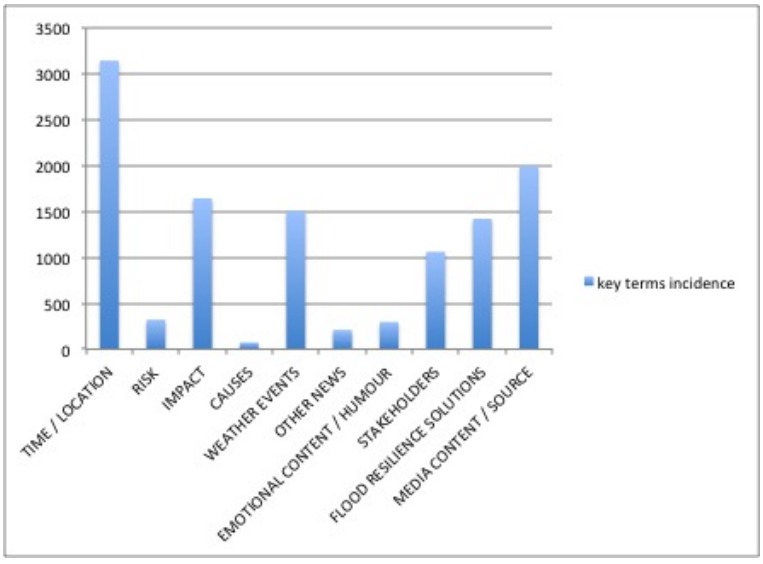

Figure 4. Twitter coverage of the 2016 Seine River flood: key terms incidence aggregated in ten thematic categories.

Figure 5a highlights that individual Twitter accounts (i.e. accounts that aren't owned by an organisation but by a person) represent a relevant portion of the most active users as well as of the most popular users. Among the most active users, 37% own an individual account. The most liked users are characterised by a majority of individual accounts (65%). The percentage is reduced in the case of the most retweeted users: only 46% of them own an individual account. Hence, with a percentage difference of 19%, tweets from individuals generated less retweets than likes. Twitter followers seem to prefer supporting individual accounts by liking their tweets, while they tend to retweet less frequently – probably only when the tweet content is valuable and worth being relayed.

Figure 5b presents the area of activity of the 59 Twitter users that published at least ten tweets in June 2016 on the topic of the Seine River flood. The activity area of these Twitter users was established on the base of the description included in their account. The majority of users deal with weather forecasts (20%) or public administration / policy making (19%), the next biggest area of activity gathers those users that are active in the field of journalism (14%). A first inference can be made on the base of these percentages: information tweeted by public authorities and policy makers seems more frequent than information tweeted by rescue services (5%). This marked difference could be explained by the fact that rescue services usually centralise information management.

Figure S5 zooms in on the first five most active users. It is possible to observe the impact that frequent tweeting has in terms of popularity, i.e. in terms of likes and retweets. By highlighting the mean values and the upper bound[5] values, it is possible to see that retweets and likes follow similar patterns. If the number of likes is high, the number of retweets will probably be high as well.

[5]After finding the interquartile range (IQR) and the upper quartile (Q3), the upper bound is calculated with the following formula :
Upper bound = Q3+1.5*IQR.

Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-200
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Discussion started: 26 July 2018

Natural Hazards and Earth System Sciences
Discussions

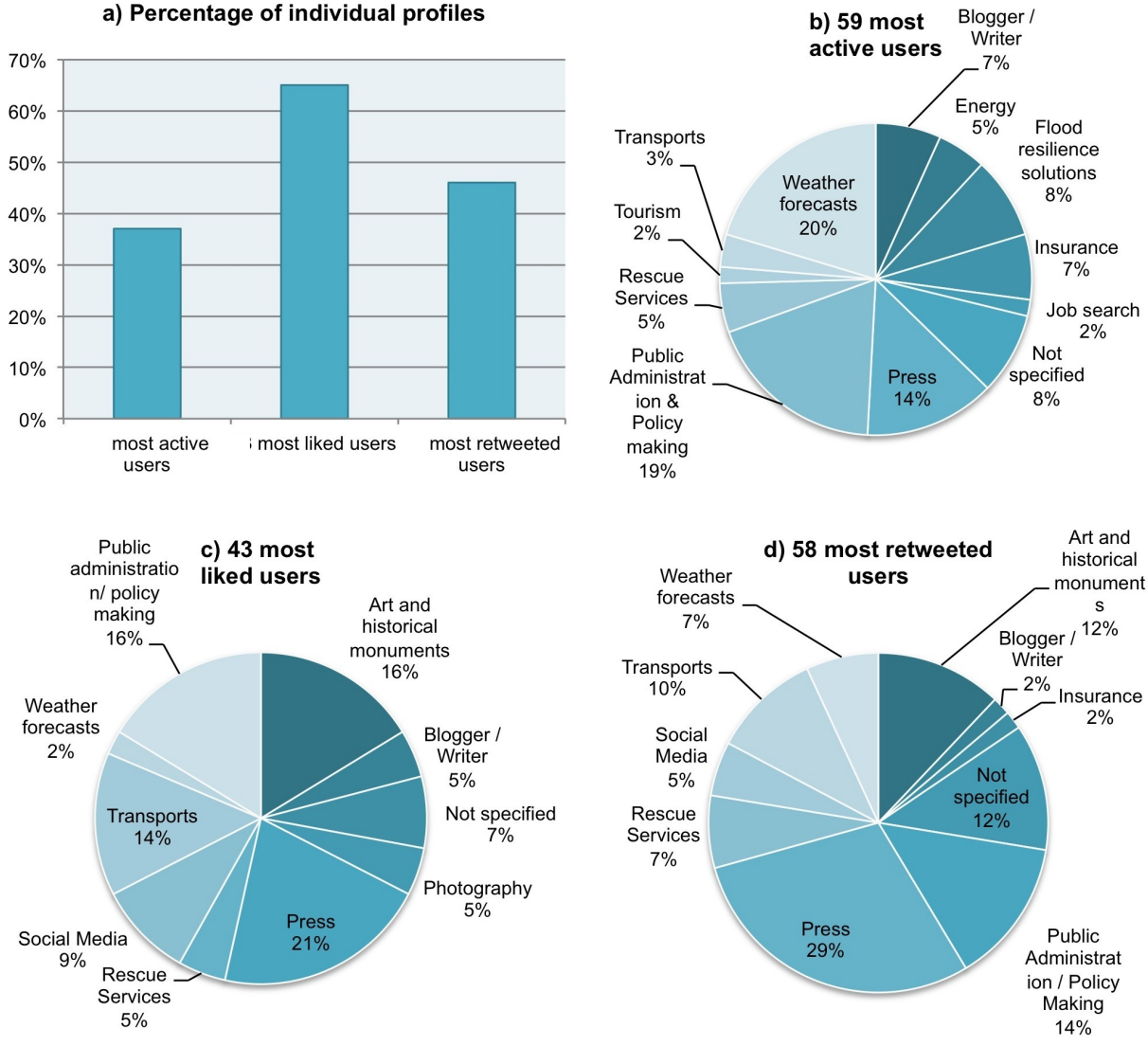

Figure 5. The users' behaviour (Twitter coverage of the 2016 Seine River flood): (a) percentage of individual profiles; (b) area of activity of the most active users (59 users that published more than 10 tweets in one month); (c) area of activity of the most liked users (43 users that received more than 50 likes per tweet in one month); (d) area of activity of the most retweeted users (58 users that received more than 50 retweets per tweet in one month).

As shown in Fig. 5c the most liked users are journalists (21%), followed by users that operate in the field of public administration / policy making (16%), art / historical monuments (16%) and transports (14%). The press seems to raise broad interest: in the changing landscape of digital media, the press continues to be considered as a source of reliable information. Public

authorities and policy makers are frequently in the social media spotlight, as their views can have direct consequences for the society. The popularity of the accounts related to art /historical monuments and transports is likely due to the fact that Paris region inhabitants, as well as tourists and people travelling across the region, felt strongly affected by the flood impacts on museums and transport infrastructures.

By looking at the areas of activity of the most retweeted accounts in Fig. 5d, we can notice similar trends to those presented in Fig. 5c: the most relevant segment is the press (29%), followed by public authorities / policy making (14%) and art / historical monuments (12%). The portion of tweets published by users that deal with transports is smaller (10%) than in Fig. 5c with a percentage difference of 4%. A small difference that could be explained by the fact that these tweets describe how transport workers cope with the flood, but they don't convey helpful information for the passengers.

6 Results and discussion

6.1 Comparison of the four histograms

The case studies discussed in this article aim at illustrating how big data exploration techniques allow complex analysis of the digital media debate on urban resilience to extreme weather. The first step of the analysis consisted in observing the media coverage distribution over one month. The four histograms presented in Fig. 1 are based on four different datasets: press news on the 2016 Seine River flood; press news on the 2018 Seine River flood; press news on the 2015 flood in the Alpes-Maritimes Department; tweets on the 2016 Seine River flood. A comparison of the four histograms highlights that, in the press as well as in social media, the peak of publications per day is determined by the date of the highest river discharge. This clearly proves a correlation between a meteorological event and its social representation. A possible explanation could be that the press tend to rather focus on the immediate consequences of natural disasters (Houston et al., 2012). Twitter seems to follow the same trend as the press. Figure 1 also shows a noteworthy difference between the Alpes-Maritimes flood and the two Seine River floods. The 2015 and 2018 Seine River floods are marked by a slower decrease of press coverage, after the maximum coverage peak is reached. This kind of evolution is probably due to the high media visibility of Paris, and it could have been reinforced by the very slow decrease of water levels of the Seine River.

A list of terms related to flood resilience solutions was then defined through iteration between manual analysis of the datasets and automated text mining. An aggregated analysis of these key terms allowed identifying thematic subsets in each dataset and observing the distribution of the number of published key terms per day (Fig. 1). The comparison between the four datasets calls attention to the minor discussion on flood resilience solutions on Twitter, if we compare it to the debate in the press. Indeed Twitter is a social media that is typically used as an early warning system: to disseminate factual information on the time and location of a flood.

6.2 Comparison of the network representations

The next stage of the analysis involved network representation applied to the 2016 Seine River flood case study and the 2015 Alpes-Maritimes case study. Thanks to an observation of the thematic clusters it was possible to gain a qualitative insight on how the debate on flood resilience is structured. A quantitative evaluation of the most central nodes and of the most frequent connections in the network enabled us to push forward the analysis: we identified the most prominent topics and actors and which of these are often coupled together. A comparison between the first and the second case study reveals some interesting differences between the two. In the first corpus of articles, the debate on various levers for flood resilience is well developed and detailed. Furthermore, specific stakeholders can be associated to specific resilience levers. In the second corpus, the important number of victims drew the media attention to the tragic impact of the flood event and to emergency management. Hence, the most relevant stakeholders are the victims and those organisations that were involved in rescue activities and economic compensation. The macro-theme "impact record" appears in both graphs and it is probably a recurring topic in the press coverage of natural disasters.

6.3 Tweets analysis

While network representation supported by Gargantext isn't suitable for a corpus of tweets, an aggregated analysis of thematic categories of Twitter terms is more appropriate for the purpose of this research. Moreover, Twitter data are valuable since they contain information on the profile of its users and metrics that describe how the users interact with each other. Indeed, it is possible to identify the most active users and the users that publish the most popular tweets. By focusing on the tweets that obtained more than fifty likes or more than fifty retweets, it is possible to observe that the most popular tweets are published by the press and public authorities, i.e. those actors that are also visible in the press. Percentages presented in Fig. 5c and Fig. 5d suggest that Twitter users operating in the media sector and in the public administration / policy making sector are leading opinion makers in the debate on Paris flood risk. Twitter is a media open to any contributor, however it seems that the most popular users are those actors that are also visible in the press. Fig. 5c and 5d also call attention to a widespread interest among Twitter users in the flood impact on transports and cultural heritage: this is probably specific to a population that lives or travels in a metropolis with a dense transport network and a very high concentration of historical monuments and museums.

7 Conclusions and perspectives

In this article we employed big data exploration techniques to investigate how urban resilience to extreme weather is perceived in the digital media debate. Through this study, we firstly intended to test how these techniques can be used to define metrics of social representation of urban resilience. In our view, these metrics can be integrated to a wider assessment of urban resilience to weather extremes. Secondly, we aimed at gaining an insight of the social perception of flood resilience in two French urban areas. This research is still in progress, however the experiments presented in this paper allowed us to obtain quantitative data on:

- The correlation existing between the intensity of the digital media debate (a social factor) and the level of the river discharge (an environmental factor);
- The evolution of the intensity of the debate over time, in two different locations and in two different media contexts (French press and Twitter);
- The differences that exist in terms of quality of the contents (i.e. reference to flood resilience solutions) between the press coverage of a flood event and the Twitter coverage of the same event;
- The most central topics and actors in the press, and how these patterns change in two different urban areas;
- The most frequent connections that exist among these topics and actors, and how these patterns change in two different urban areas;
- The most prominent topics in the Twitter debate;
- The profile of the most active and the most popular Twitter users.

These initial results are promising: they allowed a complex understanding of the intensity and quality of the digital media debate. In the future we intend to push forward our research by considering a longer time scale. We also intend to study the correlations that might exist between the intensity and quality of the debate and other resilience variables, such as: the number of citizens affected by extreme weather, the surface of regreened areas, the amount of insurance compensations for natural disasters, etc. Tweets analysis could be more fruitful if supported with automated exploration of Twitter accounts and network representation of likes and retweets. It would be then possible to analyse larger samples of data and easily move from an aggregated level to a detailed level of analysis.

Data availability. Supplements to this article are available online.

Acknowledgements. We are thankful for the technical support provided by Institut des Systèmes Complexes Paris Île-de-France and Frédérique Bordignon from École des Ponts ParisTech during the implementation of Europresse, Gargantext and Gephi. This research in led in the framework of the Chair “Hydrology for Resilient Cities” supported by Veolia.

References

Blog du moderateur: blogdumoderateur.com, last access: 18 May 2018.

Boyd, A.: Broadcast Journalism, Techniques of Radio and TV News. Oxford: Focal, 1994.

Bruns, A. and Liang, Y.E.: Tools and methods for capturing Twitter data during natural disasters, First Monday, 17, 4, doi:10.5210/fm.v17i4.3937, 2012.

Bruns, A., Burgess, J., Crawford, K. and Shaw, F.: #qldfloods and @QPSMedia: Crisis communication on Twitter in the 2011 south east Queensland floods, ARC Centre of Excellence for Creative Industries and Innovation, Brisbane, 58 pp., 2012.

Bucchi, M.: Style in science communication. Public Understanding of Science, 22, 8, 904-915, 2013.

Chacon-Hurtado, J. C., Alfonso, L. and Solomatine, D.: Dimensioning of precipitation citizen observatories in an uncertainty-aware context, EGU General Assembly, Vienna, Austria, 23-28 April 2017, EGU2017-18523-1, 2017.

Chavalarias, D.: Le Tweetoscope Climatique. Une représentation collective des enjeux autour du climat. La lettre de l'INSHS, 38, 39-42, 2015.

Climaps by Emaps: climaps.eu, last access: 18 May 2018.

Cutter, S.L., Barnes, L., Berry, M., Burton, C., Evans, E., Tate, E. and Webb, J.: A place–based model for understanding community resilience to natural disasters, Global Environ. Chang., 18, 598–606, 2008.

Cutter, S.L., Burton, C.G. and Emrich, C.T.: Disaster Resilience Indicators for Benchmarking Baseline Conditions, J. Homel. Secur. Emerg., 7, 1, 51, 2010.

Europresse: europresse.com, last access: 18 May 2018.

Gaitan, S., Calderoni, L., Palmieri, P., Ten Veldhuis, M.C., Maio, D., and van Riemsdijk, M.B.: From Sensing to Action: Quick and Reliable Access to Information in Cities Vulnerable to Heavy Rain. IEEE Senspors Journal, 14, 4175-4184, 2014.

Gargantext: gargantext.org (D. Chavalarias and A. Delanöe, 2017), last access: 18 May 2018.

Gephi: gephi.org, last access: 18 May 2018.

Houston, J.B., Pfefferhaum, B. and Rosenholtz, C.E.: Framing and Frame Changing in Coverage of Major U.S. Natural Disasters, 2000–2010, Journalism and Mass Communication Quarterly, 89, 4, 606-623, 2012.

Keating, A., Campbell, K., Mechler, R., Michel Kerjan, E., Mochizuki, J., Kunreuther, H., Bayer, J., Hanger, S., McCallum, I., See, L., Williges, K., Atreya, A., Botzen, W., Collier, B., Czajkowski, J., Hochrainer, S. and Egan, C.: Operationalizing Resilience Against Natural Disaster Risk: Opportunities, Barriers and A Way Forward, Zurich Flood Resilience Alliance, 43 pp, 2014.

Lanfranchi, V., Ireson N., When U., Wrigley S.N. and Fabio C.: Citizens' Observatories for Situation Awareness in Flooding, in: Hiltz, S.R., Pfaff, M.S., Plotnick, L. and Shih P.C. (ed.) Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2014): 18-21 May 2014, University Park, Pennsylvania, USA, 145–154, 2014.

Latour, B.: The whole is always smaller than its parts. A digital test of Gabriel Tarde's Monads, British Journal of Sociology, 63, 4, 591-615, 2012.

Mangalagiu, D., Wilkinson, A. and Kupers, R.: When futures lock-in the present: Towards a new generation of climate scenarios, in: K. Hasselmann, C. and Jaeger, C. (ed.) Reframing the Problem of Climate Change: From Zero Sum Game to Win-Win Solutions, Routledge, 2012.

Morss, R., Demuth, J., Lazrus, H., Palen, L., Barton, C. Davis, C. Snyder, C., Wilhelmi, O., Anderson, K., Ahijevych, D., Anderson, J., Bica, M., Fossell, K., Henderson, J., Kogan, M., Stowe, K. and Watts, J.: Hazardous Weather Prediction and Communication in the Modern Information Environment, Bull. Amer. Meteor. Soc., doi:10.1175/BAMS-D-16-0058.1, 2017.

Niederer, S.: Global warming is not a crisis!: Studying climate change skepticism on the Web. NECSUS European Journal of Media Studies, 2, 83–112, DOI:10.5117/NECSUS2013.1.NIED, 2013.

OECD: Seine Basin, Île-de-France, 2014: Resilience to Major Floods, OECD Reviews of Risk Management Policies, Éditions OCDE, Paris, 204 pp, 2014.

Palen, L., Starbird, K., Vieweg, S. and Hughes, A.: Twitter–based information distribution during the 2009 Red River Valley flood Threat, Bul. Am. Soc. Info. Sci. Tech., 36, 13-17, doi:10.1002/bult.2010.1720360505, 2010.

Resilience Alliance: Assessing resilience in social–ecological systems: Workbook for practitioners. Version 2.0., Resilience Alliance, 54 pp, 2010.

Rogers, R. and Marres N.: Landscaping climate change: A mapping technique for understanding science and technology debates on the World Wide Web. Public Understanding of Science, 9, 141–163, DOI:10.1088/0963- 6625/9/2/304, 2000.

Trench, B.: Internet: turning science communication inside-out?, in: Bucchi, M. and Trench, B. (ed.) Handbook of Public Communication of Science and Technology. Routldedge, 2008.

Twitter: twitter.com, last access: 18 May 2018.

Twitter Advanced Search: twitter.com/search-advanced; last access: 8 July 2018.

UN/ISDR: Indicators of Progress: Guidance on Measuring the Reduction of Disaster Risks and the Implementation of the Hyogo Framework for Action, United Nations secretariat of the International Strategy for Disaster Reduction (UN/ISDR), Geneva, Switzerland, 59 pp, 2008.

Venturini T., Baya Laffite, N., Cointet, J.P., Gray, I., Zabban, V., De Pryck, K.: Three maps and three misunderstandings: A digital mapping of climate diplomacy, Big Data and Society, 1, 2, 1–19, 2014.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L.: Microblogging during two natural hazard events: what Twitter may contribute to situational awareness, in: CHI '10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA, 10-15 April 2010, 1079-1088, 2010.