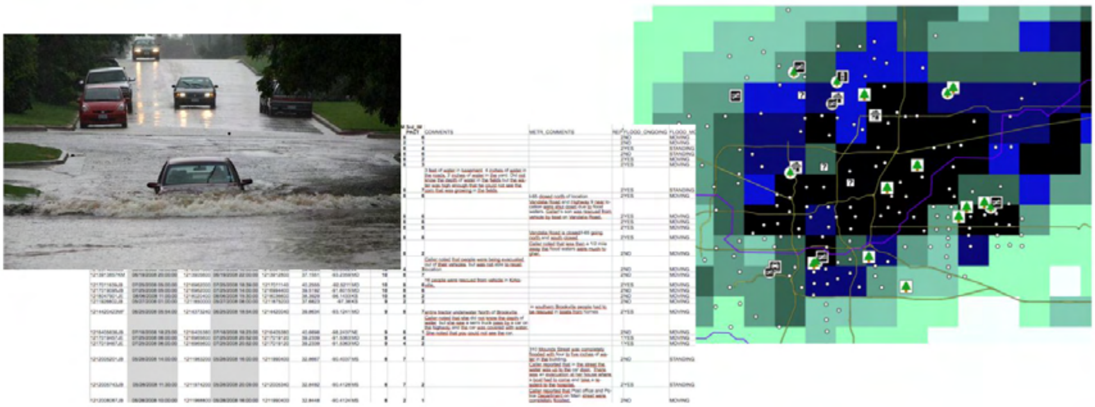
M2R HydroHazards internship report :

Impact classification of flash flood reports and ability of US flash flood forecasting tools to predict these impacts.

Martin Calianno

martin.calianno@ujf-grenoble.fr

January 2012



Degree :
Master2 Research HydroHazards
University J. Fourier, 38000 Grenoble (France) and University of Thessaly, 38221 Volos (Greece)
Supervisors :
- Isabelle Ruin
Laboratoire d'étude des Transferts en Hydrologie et Environnement (LTHE), Grenoble (France)
- Jonathan-J. Gourley
National Severe Storm Laboratory (NSSL), Norman, OK (USA)
Internship period :
- July 15th - September 30th 2011 (LTHE, Grenoble)
- October 5th - December 21st (NSSL, Norman)

Abstract

Surveys have been undertaken in the US on operational basis to report flash flood events, as well as their magnitude and spatio-temporal extent. The first dataset is the National Weather Service (NWS) *Storm Data*, which consists of spotters reports. The second is the Severe Hazards Analysis and Verification Experiment (SHAVE), conducted by the National Severe Storms Laboratory (NSSL) in Norman [Gourley et al., 2010]. SHAVE dataset is based on a near real-time public survey.

This study provides an impact classification of these datasets, to evaluate the ability of three US flash flood forecasting tools (FFG, GFFG and DHM-TF) to predict impacts. SHAVE impacts are used in a first spatio-contextual analysis, based on interviewees answers already included in the dataset, as well as GIS-sampled spatial attributes. This analysis showed consistent results, indicating that impact classification was made correctly and that the SHAVE dataset (even if based on public polls) is a reliable tool for flash flood characterisation. Moreover, interesting results emerged. Evacuations are not only observed in urban zones, but also in rural areas, and Rescues, Fatality or Injuries are mostly observed in low population density areas. Moreover, these severe impacts are not always perceived by the interviewees as extreme, rare events, which may indicate how people have difficulties to estimate an event frequency.

For two extreme cases of flash flood in Oklahoma (the Erin and Oklahoma City events) FFG, GFFG, and DHM-TF are evaluated on a YES/NO-forecast basis, but also as function of impacts. For the Erin event (using NWS reports), DHM-TF shows the best Probability Of Detection (1), followed by FFG (0.94) and GFFG (0.78). But on maps, DHM-TF (and GFFG, to a lesser extent) show larger forecast areas, which may indicate more false alarms, even if False Alarm Ratio can not be computed with the NWS dataset. Also, FFG and GFFG seem to show a relation between average tool values and impacts, when ranked by severity, whereas DHM-TF does not. For the Oklahoma City event (using SHAVE reports), only GFFG and DHM-TF are available but with SHAVE, FAR and Critical Success Index can be computed. GFFG shows the best value of CSI (0.14), despite a lower POD value than DHM-TF, but thanks to a better FAR. These CSI values are very low, even if they stay in the same order of magnitude as the highest value found by Gourley et al. [2011a]. For this particular smaller scaled, urban case, no clear link is found between impact type and tools values.

Keywords : Flash flood ; Database ; Impacts ; Forecasting ; Model evaluation

Περίληψη

Στις Η.Π.Α έρευνες έχουν διεξαχθεί σε επιχειρησιακή βάση ώστε να μελετηθούν φαινόμενα ξαφνικών πλημμυρών καθώς επίσης το μέγεθος και η χωροχρονική τους έκταση. ;ο πρώτο σύνολο δεδομένων είναι τα δεδομένα καταιγίδων του National Weather Service (NWS) που αποτελείται από εκθέσεις παρατηρήσεων. ;ο δεύτερο είναι το πείραμα SHAVE (Severe Hazards Analysis and Verification Experiment) που διεξήχθη από το NSSL (National Severe Storms Laboratory) , στο Norman [Gourley et al., 2010]. Το σύνολο δεδομένων του SHAVE βασίζεται σε σχεδόν πραγματικού χρόνου δημοσκοπήσεις. Η μελέτη αυτή παρέχει μια ταξινόμηση επιπτώσεων των δεδομένων αυτών για την εκτίμηση της ικανότητας τριών προγνωστικών εργαλείων για ξαφνικές πλημύρες στις Η.Π.Α. (FFG, GFFG kai DHM-TF), με σκοπό να γίνει η πρόγνωση αυτών των επιπτώσεων. Οι επιπτώσεις του πειράματος SHAVE χρησιμοποιούνται σε μια πρώτη χωρική ανάλυση συμφραζομένων, βασισμένη σε απαντήσεις των δημοσκοπούμενων, οι οποίες συμπεριλαμβάνονται στο σύνολο των δεδομένων, καθώς επίσης και στα δειγματοληπτικά χωρικά χαρακτηριστικά των Γεωγραφικών Συστημάτων Πληροφοριών (ΓΣΠ). Η ανάλυση αυτή έδειξε συνεπή αποτελέσματα, αποδεικνύοντας ότι η ταξινόμηση των επιπτώσεων ήταν ορθή, και ότι το σύνολο δεδομένων του SHAVE (αν και βασίζεται σε κοινωνικές δημοσκοπήσεις) είναι ένα αξιόπιστο εργαλείο για τον χαρακτηρισμό των ξαφνικών πλημμυρών. Επιπλέον προέκυ;αν ενδιαφέροντα αποτελέσματα. Εκκενώσεις παρατηρήθηκαν όχι μόνο σε αστικές ζώνες αλλά και σε αγροτικές περιοχές, και επιπλέον διασώσεις, θνησιμότητα ή τραυματισμοί παρατηρούνται κυρίως σε αραιοκατοικημένες περιοχές. Επίσης αυτές οι σοβαρές επιπτώσεις δεν εκτιμώνται πάντοτε από τους ερωτηθέντες ως ακραία, σπάνια περιστατικά, το οποίο μπορεί να αποδεικνύει πως οι άνθρωποι αντιμετωπίζουν δυσκολία στο να αντιμετωπίσουν την συχνότητα του φαινομένου. Για δύο ακραίες περιπτώσεις ξαφνικών πλημμύρων στην Οκλαχόμα (περιστατικά Erin και Oklahoma City) τα εργαλεία FFG, GFFG και DHM-TF εκτιμήθηκαν σε μια προγνωστική βάση τύπου ΝΑΙ/ΟΞΙ, καθώς επίσης και ως συνάρτηση των επιπτώσεων. Για την περίπτωση της Erin (χρησιμοποιώντας τις αναφορές του NWS), το DHM-TF δείχνει την καλύτερη πιθανότητα ανίχνευσης (1), ακολουθεί το FFG με πιθανότητα 0.94 και το GFFG με πιθανότητα 0.78. Ωστόσο στους χάρτες το DHM-TF (και το GFFG σε μικρότερη έκταση) δείχνει μεγαλύτερες περιοχές πρόγνωσης, το οποίο μπορεί να υποδεικνύει περισσότερες εσφαλμένες αναφορές εκτάκτου ανάγκης, αν και οι αναλογία τους μπορεί να υπολογιστεί από το σύνολο δεδομένων του NWS. Επιπλέον το FFG και το GFFG φαίνεται να δείχνουν μία σχέση μεταξύ των μέσων τιμών των εργαλείων και των επιπτώσεων, όταν ταξινομούνται βάση της δριμύτητάς τους, ενώ αυτό δεν συμβαίνει με το DHM-TF. Για το περιστατικό της Oklahoma City (χρησιμοποιώντας τις αναφορές του SHAVE) μόνο το GFFG και το DHM-TF είναι διαθέσιμα αλλά σ΄ αυτήν την περίπτωση μπορούν να υπολογιστούν το FAR και ο δείκτης CSI (Critical Success Index). Το GFFG δείχνει την καλύτερη τιμή για τον CSI (0.14), παρά την χαμηλότερη τιμή POD απ΄ ότι το DHM-TF, αλλά εξαιτίας καλύτερης τιμής του FAR. ;υτές οι τιμές του CSI είναι πολύ χαμηλές, αν και παραμένουν στην ίδια τάξη μεγέθους με την μέγιστη τιμή που βρήκαν οι Gourley et al. [2011a]. Για αυτήν την συγκεκριμένη μικρής κλίμακας αστική περίπτωση, δεν βρέθηκε κάποια καθορισμένη διασύνδεση μεταξύ του τύπου επίπτωσης και των τιμών των εργαλείων.

Λέξεις-Κλειδιά : Ξαφνική Πλημμύρα· Βάση Δεδομένων, Επιπτώσεις· Πρόγνωση· Αξιολόγηση Μοντέλου.

Résumé

Des études ont été menées aux USA, de manière opérationnelle, pour reporter les crues éclaires, ainsi que leur magnitude et leur étendue spatio-temporelle. La première base de données est le *National Weather Service (NWS) Storm Data*, qui consiste en des rapports d'observateurs professionnels. La seconde est le *Severe Hazards Analysis and Verification Experiment (SHAVE)*, mise en place par le National Severe Storms Laboratory (NSSL) à Norman [Gourley et al., 2010]. SHAVE est basée sur un système d'enquête par téléphone en quasi-temps réel.

La présente étude fournit une classification d'impacts pour ces deux bases de données et évalue la capacité de trois outils américains de prédiction des crues éclaires (FFG, GFFG et DHM-TF) à prédire ces impacts. Les impacts SHAVE sont utilisés dans une première analyse, basée sur les réponses déjà incluses dans la base de données et sur des attributs spatiaux échantillonnés via GIS. Cette analyse montre des résultats cohérents, indiquant que la classification des impacts a été faite correctement, et que SHAVE (même si basée sur des sondages publics) est un outil fiable pour la caractérisation des crues éclaires. De plus, des résultats intéressants émergent. Les évacuations ne sont pas seulement observées dans les zones urbaines, mais aussi dans les zones rurales, et les sauvetages, décès ou blessés sont majoritairement observés dans des zones de faible densité de population. De plus, ces impacts sévères ne sont pas toujours perçus par les sondés comme des événements rares, extrêmes, ce qui pourrait indiquer la difficulté que les gens ont pour estimer la fréquence de tels événements.

Pour deux cas extrêmes de crues éclaires en Oklahoma (les événements d'Erin et de Oklahoma City) FFG, GFFG et DHM-TF sont évalués de manière binaire (événement prédit ou pas), mais aussi en fonction des impacts. Pour l'événement d'Erin (en utilisant les rapports NWS), DHM-TF montre la meilleure Probabilité De Détection (1), suivi de FFG (0.94) et GFFG (0.78). Mais sur les cartes, DHM-TF (et GFFG, dans une moindre mesure) montrent de plus larges étendues de crues détectées, ce qui pourrait indiquer plus de fausses alertes, même si malheureusement le Ratio de Fausses Alertes ne peut être calculé avec les rapports NWS. Aussi, FFG et GFFG semblent montrer une relation entre les valeur moyennes des outils de prévision et les impacts, étant classés par sévérité croissante, alors que DHM-TF ne montre pas telle relation. Pour l'événement d'Oklahoma City (en utilisant les rapports SHAVE), seuls GFFG et DHM-TF sont disponibles, mais dans ce cas-ci, le RFA et l'Index Critique de Succès peuvent être calculés. GFFG montre la meilleure valeur de ICS (0.14), malgré une PDD plus basse que DHM-TF mais grâce à un meilleur RFA. Ces valeurs de ICS sont très faibles, même si elles restent dans le même ordre de grandeur que la plus grande valeur trouvée par Gourley et al. [2011a]. Enfin, pour ce cas particulier urbain et à plus petite échelle, aucun lien clair n'est trouvé entre le type d'impact et les valeurs d'outils de prédiction.

Mots-clefs : Crue éclair ; Base de données ; Impacts ; Prévision ; Evaluation de modèle

Acknowledgements

Many thanks to Isabelle Ruin and JJ Gourley for the supervision from both sides of the Atlantic.

Special thanks to Zac Flaming and Race Clark from the NSSL for data preparation.

Contents

1 Introduction

Flash flooding differs from river flooding in terms of space and time. A flash flood is defined as "a rapid flooding of water over land caused by heavy rain or a sudden release of impounded water (e.g., dam or levee break) in a short period of time, generally minutes op to several hours" [Hong et al., 2010]. In the United States, flash flooding is considered as one of the first cause of death among weather-related hazards [Ashley and Ashley, 2008]. However, comparatively to its human and environmental impacts, this phenomenon remains poorly documented [Gaume and Borga, 2008]. Efforts have been made over the last decades to collect data of physics (rainfall-runoff processes) and spatio-temporal scope of flash floods : stream gauges measurements, remote sensing of water surface extents, post event field investigations or rainfall-runoff modelling. Beside these *process understanding* datasets, surveys have been undertaken in the US on operational basis to report flash flood events, as well as their magnitude and spatio-temporal extent. The first dataset is the National Weather Service (NWS) *Storm Data*, which consists of spotters reports collected by US Weather Forecast Offices. The second is the Severe Hazards Analysis and Verification Experiment (SHAVE), conducted by the National Severe Storms Laboratory (NSSL) in Norman [Gourley et al., 2010]. SHAVE dataset is based on a near real-time public survey. These products are primarily designed to evaluate US flash flood forecasting tools, on a yes/no event basis. Nevertheless, information contained in these databases (e.g., flood magnitude, damages, fatalities, contextual comments) can be used to further portray flash flood events in terms of impacts, and eventually see how forecasting tools are able to predicts such impacts.

Indeed, as societal impacts of flash floods are resulting from the combination of flooding hazard and human/environmental vulnerability, research now also focuses on these vulnerability aspects, to ultimately integrate hydrometeorology and social science. This *integrated* concept was first introduced by the Weather and Society Integrated Study (WAS*IS) [Demuth et al., 2007] and recent integrated studies about flash floods include topics such as human behaviour and mobility [Ruin et al., 2008] or road susceptibility [Versini et al., 2010]. Efforts are also made at the NSSL to build an enhanced database that would combine NWS and SHAVE reports, but also US Geological Survey stream flows. Impact characterization would be a valuable addition to this project.

This report introduces an impact classification of flash flood reports, in order to evaluate the ability of US flash flood forecasting tools to predict such categories of impacts. It is organized as follows. Section 2 presents and reviews the NWS and SHAVE flash flood reports datasets. Section 3 describes the methodology for impact classification and a spatio-contextual analysis of impacts from SHAVE reports. Section 4 introduces an evaluation of the ability of US flash flood forecasting tools to predict impacts, for two extreme cases of flash flooding in Oklahoma. Finally, section 5 provides a summary and concluding remarks.

2 Presentation and review of the flash flood reports datasets

2.1 NWS reports

The NWS dataset contains flash flood reports in USA. Reports are collected throughout the year and since 2006 they are digitised and stored either as latitude/longitude points (from 2006 to 2007) or polygons (from 2007 to 2010). See Figure 1 for a presentation of SHAVE and NWS spatial coverage. The principal aim of this dataset is to verify NWS flash flood warnings, issued by Weather Forecast Offices (e.g., the Flash Flood Guidance [FFG]).

The sampling method is based on calling of trained spotters and businesses within the warned areas. Then forecasters define polygons (formerly, points) that delineate the regions impacted by flash flood. Information about event timing, fatalities, injuries or damages are also gathered, as well as comments about the flood event and the meteorological context. The structure of this dataset is presented in Appendix A (Table 11).

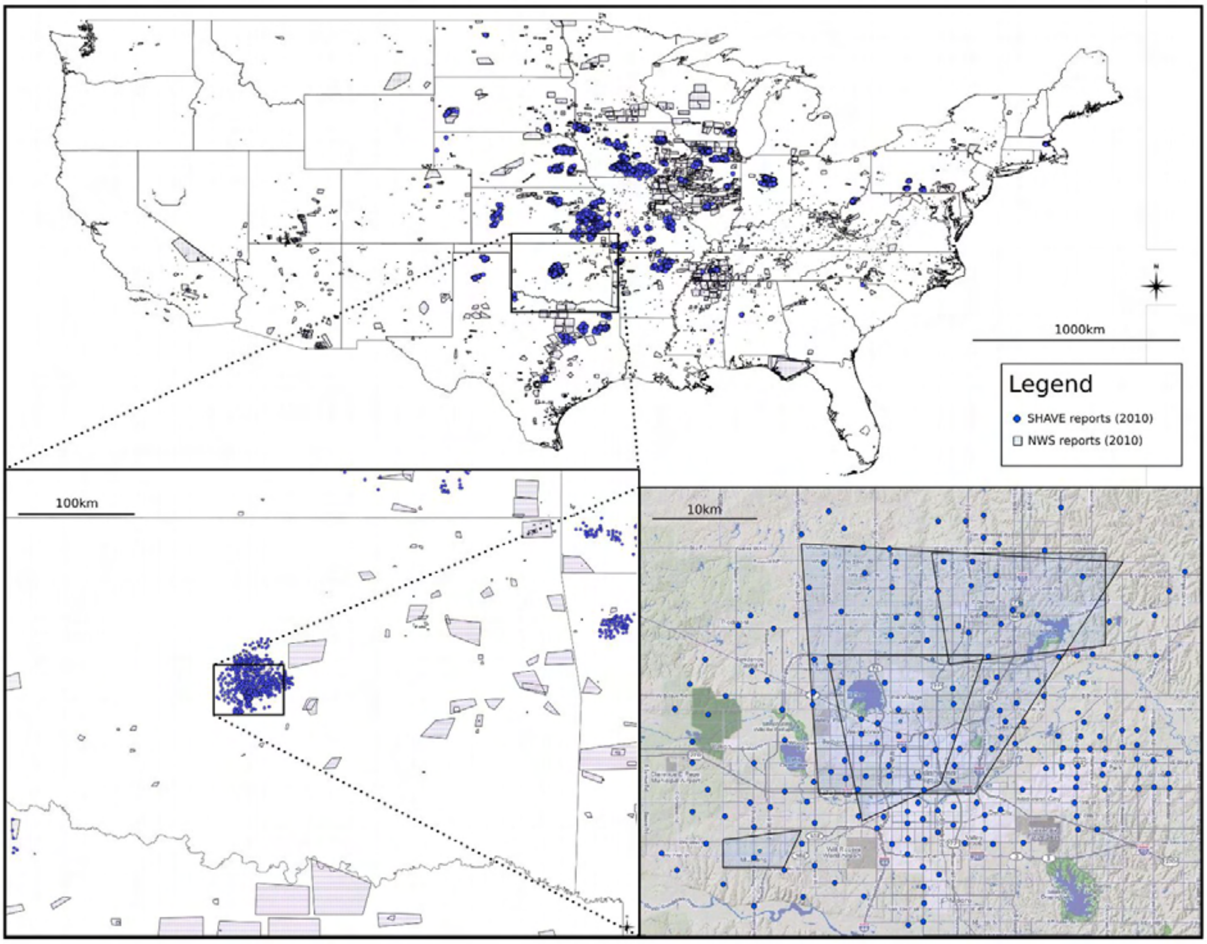


Figure 1: Spatial coverage of SHAVE flash flood observation and NWS flash flood event polygons in 2010.

NWS reports main advantages:

- NWS collects flash flood observations on all US.
- Reports sources are considered as reliable (i.e., emergency managers, trained spotters). Although sources also include newspapers and public.
- The database is developed by forecasters who are very familiar with their area of responsibility, can provide immediate quality control of reports, and there is personnel in the office at all hours of the day throughout the year.

NWS reports main drawbacks:

- Does not include reports of no flooding in warned regions (i.e., false alarms).
- Often does not report flood events that occurred without warning (i.e., missed events).
- Poor spatial accurracy of polygon reports.
- Poor timing accurracy (meteorological event start/end times are often taken as flood event timing).

2.2 SHAVE reports

The SHAVE dataset was set up at the National Severe Storms Laboratory in Norman, OK [Gourley et al., 2010] and includes flash flood reports in USA, from 2008 to 2010. Reports were collected during summer months (June > August) by undergraduate students, using questionnaires. This dataset is point-based and was originally designed to complement hail observational data (starting in 2006). But subsenquently wind damage, tornadoes and flash flood reports were added to the experiment, in order to create higher-resolution datasets for model verification.

The sampling method is storm-targeted, which means that the survey is only initiated if at least one of the following conditions is met:

1. The NWS issued a flash flood warning or urban/small stream advisory.
2. Quantitative Precipitation Estimation (QPE) from the US weather radar network approached or exceeded FFG for any of 1-,3-, or 6-h rainfall accumulation periods.
3. A call for another targeted hazard (e.g., hail) suggested flooding was a problem.

Note that the study being focused on flash flood, larger scale, fluvial floods (within basin areas $> 300\text{km}^2$) were avoided, using DEM-derived catchments and river maps.
Phone calls are then made to the public and geolocated via Google Earth^{TM} using the following strategy:

1. Identification of areas potentially impacted by flash flood (as described above).
2. Poll the public in regularly spaced clusters through the potentially impacted region. Distances between cluster points vary from 3 to 10km, depending on population density and habitat/road network configuration.
3. Adapt the spatial coverage of each cluster so the range of magnitudes varied from no impact to severe impacts.

Reports are classified as null (i.e. non impacted location), minor and severe events using information collected through the questionnaire : flood type, water movement, depth, extent, eventual evacuation or rescue, flood frequency. Localisation and timing are also included, as well as textual comments about the flood event. See Appendix A (Table 12) for a description of the database structure and details about how questionnaire answers and report severity were classified.

SHAVE reports main advantages:

- Higher spatial resolution (dense point sampling).
- Estimation of flash floods spatial magnitude (reports range from no impact to severe for each event).
- Additional information about the event (flood type, depth, extent, frequency, textual comments).
- Includes reports of no flooding in warned regions (i.e., false alarms) and flood events that occurred without warning (i.e., missed events).

SHAVE reports main drawbacks:

- Data collected only during summer months.
- Does not include information about interviewee's age, gender or profession.
- As a public-based survey, uncertainty and bias may occur :

- About timing : people might not be able to estimate start event time (e.g., if started at night) or only give rough estimations. Even worse, when timing is unknown, the recorded event start/end time is simply the time of phone call, which can be the next day. Also, because it is a near-real time poll, the event is often still ongoing, so the end time is not known. Therefore, as for NWS reports, timing must be considered extremely cautiously.

- About spatial accurracy : even if reports are precisely georeferenced using GoogleEarth^{TM}, the event described by the public may occur from one meter to a few kilometers around the report point, depending on people's perception or knowledge. For instance, people living in urban areas would be aware of their neighborhood's flooding, whereas a farmer will know if flooding occurred on his land, which might represent a much larger area. *Uncertainty* buffers must then be considered around SHAVE reports points.

Furthermore, as SHAVE is based on public survey, reports will be more represented in urban areas than in remote or rural places.

Before impact classification, a first clean up of the SHAVE dataset has been done. Fields for internal use or redundant ones were removed, i.e.: 'Event type' (only floods), 'Mags units' (meters), 'Revision time', 'Revision number' and 'Contact phone'. Also, as SHAVE reports are georeferenced in GIS format and because it was not recorded in 2010, 'City', 'County', 'State' and 'CWA' fields were removed as well.

Moreover, start/end timing fields, originally separated in 'Year', 'Month', 'Day', 'Hours (UTC)' and 'time (UNIX)' were put in two single fields : a date/time in UTC and a UNIX time. After analysing histograms of UTC hours for the three SHAVE years, a shift of 12hours was observed for the 2010 dataset. UNIX time was then found to be the most accurrate, so the new date/time UTC were computed from UNIX fields. Unfortunately, for the reasons listed above (see SHAVE drawbacks), timing inaccuracies will not permit a temporal analysis. Nevertheless, analysing them raises questions that could help to set up more efficient polls in the future :

- As SHAVE end times are often missing (or even equal to start times) and because poll times are considered inaccurate, could one single event time be sufficient, with its own uncertainties?

- Adding a field with the event start in local time might be useful if the social event has to be analysed. Nevertheless, such a field can easily be added *a posteriori*.

Further clean up of the SHAVE dataset was undertaken through impact classification, presented in the following section.

3 Flash flood impacts classification and analysis

3.1 Creation of an impact typology from flash flood reports

An impact typology was created, based on the information available in each dataset. These impacts classes were ranked from the less to the most severe, based on *a priori* judgment. As more than one impact can occur for a single flash flood report, three impact fields were created, in order to keep record of the first, second and third most severe impacts. This way, multi-impact reports can be handled.

Impact classes were made in order to be extracted from both SHAVE and NWS datasets and to allow comparison. Nevertheless, there is a slight difference between both datasets classification.

3.1.1 SHAVE impacts

Impact classification for the SHAVE dataset is based on information contained in the following fields : 'Flood type', 'Flood nature other', 'Flood evac', 'Flood rescue', 'Comments' and 'Metr comments'. As information about evacuation and rescue was recorded in two distinct fields, two impact classes could be created. Along with the presence of a 'No impact' class, this is the only difference from the NWS classification.

SHAVE impact typology is presented in Table 1.

Code	**SHAVE impact classes**	**Description**
1	No impact	SHAVE 'null reports'.
2	Other	Unclassified or unknown impact.
3	Overflow	Streams out of their banks.
4	Greenlands	Flooded cropland, pasture, yard or grassland.
5	Street/Road	Flooded street or road
6	Road closure	Also includes impassible roads.
7	Inundation	Floodwaters in buildings or homes, including basements.
8	Evacuation	
9	Stranded cars	e.g.: moved by floodwaters, stalled in ditches,...
10	Rescue/Fatality/Injury	

Table 1: SHAVE impact classification.

3.1.2 NWS impacts

Impact classification for the NWS dataset is based on information contained in the following fields : 'Direct injuries', 'Indirect injuries', 'Direct fatalities', 'Indirect fatalities' and 'Event narrative'. For the NWS dataset, impact classification is almost only based on textual comment (the 'Event narrative' field), which makes the task more challenging (there are about 3000 reports a year). The impact classification was then limited to the Arkansas Red River Basin, for which forecasting products are readily available.

Recall that NWS does not include null reports, then the 'No impact' class is not used. Also, as for SHAVE dataset, redundant fields were removed. NWS impact typology is presented in Table 2.

Code	**NWS impact classes**	**Description**
2	Other	Unclassified or unknown impact.
3	Overflow	Streams out of their banks.
4	Greenlands	Flooded cropland, pasture, yards or grassland.
5	Street/Road	Flooded street or road
6	Road closure	Also includes impassible roads.
7	Inundation	Floodwaters in buildings or homes, including basements.
8	Stranded cars	e.g.: move by floodwaters, stalled in ditches,...
9	Evacuation/Rescue/Fatality/Injury	

Table 2: NWS impact classification.

3.2 Spatial and contextual analysis of SHAVE flash flood impacts

An analysis was done by crossing SHAVE flash flood impacts with the interviewee perception of the flooding context and characteristics (already included in the dataset) and with spatial attributes, sampled using Geographic Information System.

3.2.1 Perceived attributes

Some information about interviewees perceptions of flash floods are readily available for each report in the SHAVE dataset. These *perceived attributes*, chosen to be compared with flash flood impacts are the following :

Water Movement ; nominal variable

Water Depth ; scaled variable [$meters$]

Flood Frequency (return period) ; nominal variable

3.2.2 GIS-sampled spatial attributes

Spatial attributes were added to the SHAVE dataset by sampling raster data in a GIS :

Land Use ; nominal variable

Population Density ; scaled variable $[inhab./km^2]$

Local Upslope ; scaled variable $[degrees]$

Drainage Area ; scaled variable $[km^2]$

Wetness index (CTI) ; scaled variable : $\mathrm{CTI} = \ln[\textit{Drainage area}/\tan(\textit{Local upslope})]$

The Wetness index, or Compound Topographic Index (CTI) is a value combining upslope drainage area and local slope. It is commonly used to quantify topographic control on hydrological processes : it describes water accumulation in soils [Beven and Kirkby, 1979]. Low CTI values describe places with smaller drainage areas and steeper slopes, whereas high values are associated to larger drainage areas and more gentle slopes.

Attributes were retrieved by point-sampling, i.e., by assigning the pixel value located right under each SHAVE report. All these raster layers are at 1km resolution.
Population density is derived from the US 2000 census. Land Use is taken from USGS' Land Use and Land Cover (LULC) database. Local Upslope, Drainage Area and Wetness index rasters are taken from the USGS' HYDRO1k dataset, derived from the 30 arc-second digital elevation model of the world (GTOPO30). Please refer to the Earth Resources Observation and Science (EROS) website for additional information : http://eros.usgs.gov/.
See Appendix B (Tables 13 and 14) for a description of the updated *impact-foccused* SHAVE and NWS datasets.

3.2.3 Methodology for attribute categorisation and cross-tabulation analysis

Because flash flood impacts and many attributes consist of categorised variables, a cross-tabulation approach was chosen to analyse the relationship between these variables. So, in order to deal only with nominal variables, continuous attributes (i.e., Water Depth, Population Density, Local Upslope, Drainage Area and Wetness index) had to be categorised. Attribute categories were chosen manually, in order to be both meaningful and sufficiently sampled (see Table 3).

- The Water Depth perceived attribute has been split into three categories: $\leq$10cm (corresponding to ankle-deep waters and below), 10-30cm (between ankle-deep and knee-deep waters) and $>$30cm (above knee-deep waters).

- The 13 Land Use classes (see Table 14) have been grouped to make five new categories : Water, Forest (includes all Forest, Woodland and Shrubland sub-categories), Grassland, Cropland and Urban. Note that the Bare Ground class has not been sampled.

Impact classes	OVER.	GREEN.	ROAD	CLOS.	INUN.	EVAC.	CARS	RESC.
n =	471	1019	237	291	388	71	40	31
Water Movement	Moving	Standing	(unknown)					
n =	1131	776	641					
Water Depth [cm]	≤ 10	]10-30]	> 30	(unknown)				
n =	815	608	905	220				
Flood Return Period	$\leq$ 1year	]10-30y.]	Never Seen	(unknown)				
n =	1440	295	312	501				
Land Use	Water	Forest	Grassland	Cropland	Urban			
n =	15	906	160	1225	242			
Pop. Density [i./km^2]	≤ 4	]4-70]	]70-500]	> 500				
n =	1017	866	339	326				
Local Upslope [°]	≤ 0.2	]0.2-0.6]	> 0.6					
n =	946	1182	420					
Drainage Area [km^2]	≤ 1	]1-20]	> 20					
n =	1562	642	343					
Wetness index (CTI)	≤ 5	]5-7]	> 7					
n =	503	1244	801					

Table 3: Distribution of impacts and spatial attributes classes.

- Population Density has been divided into four classes, in order to account for sparsely populated ($\leq$4 inhab./km^2), low density (]4-70] inhab./km^2), high density (]70-500] inhab./km^2) and very high density areas ($>$500 inhab./km^2).

- Drainage Area has been split into three classes. The first class ($\leq$1km^2) corresponding to points sampled in a grid cell which has only one or no adjacent draining cell. The two other classes has been chosen to make the distinction between drainage areas being below and above 20km^2. This limit was chosen in accordance with Ruin et al. [2008], who studied the hydro-meteorological circumstances of fatal accidents during the 2002 flash flood event in the Gard region (France). They found that fatalities in catchments $<$20km^2 occured outdoor, with and average age of 43 and were mainly males, whereas in larger catchments ($>$1000km2), fatalities occured at home and concerned older people (average age of 76).

- Finally, Local Upslope and Wetness index were first split into five classes using quantiles, but it happened that there were too many classes to give robust cross-tabulation results, then these 5 classes were grouped into three larger categories.

This classification leads to unevenly distributed categories of spatial/perceived attributes (the independent variables). The impact classes (the dependent variables) being also strongly unevenly sampled (for instance, n=1019 for 'Greenland' and n=71 for 'Evacuation'), a somewhat *double* standardisation has been done. One one hand, to take into account the uneven impact classes distribution, percentages have been computed within each impact classes (see Table 4). On the other hand, to take into account the non-uniform attribute classes distribution, a *deviation from the attribute total percentage* was computed. It is simply the subtraction between impact percentages and the total percentage of each attribute. These values will be used in the next section to illustrate cross-tabulation results.

Attributes	**Impacts**								
(Water Move.)	OVER.	GREEN.	ROAD	CLOS.	INUN.	EVAC.	CARS	RESC.	Total
MOVING									
Count	293	340	128	166	111	46	23	24	1131
% within Impacts	80.9%	45.9%	74.4%	75.8%	38.7%	69.7%	76.7%	80.0%	59%
Deviation from attribute Total %	21.6%	-13.4%	15.1%	16.5%	-20.6%	10.4%	17.4%	20.7%	
STANDING									
Count	69	401	44	53	176	20	7	6	776
% within Impacts	19.1%	54.1%	25.6%	24.2%	61.3%	30.3%	23.3%	20.0%	41%
Deviation from attribute Total %	-21.6%	13.4%	-15.1%	-16.5%	20.6%	-10.4%	-17.4%	-20.7%	
Total									
% of Total	19.0%	38.9%	9.0%	11.5%	15.1%	3.5%	1.6%	1.6%	100%
Count	362	741	172	219	287	66	30	30	1907

Table 4: Example of cross-tabulation : SHAVE Impacts crossed with Water Movement.

Note that for impacts, the 'second' and 'third' impact fields were added to the analysis, which allows larger sample sizes.

3.2.4 Results and discussion : cross tabulation analysis

Before analysing cross-tabulation results, two statistics have been computed to account for independence and relationship's strength of each impact-attribute crossing : the Pearson Chi^2 and Cramer's V tests. The Pearson's Chi^2 is a test for independence (i.e., independence between tested variables is the null hypothesis, H0). So if H0 is significantly rejected (when the Chi^2 asymptotic significance value [the p-value] is below the significance level, at α=0.05), there is a statistically significant relationship between the variables. The Cramer's V is a test that evaluates the strength of a relationship between variables. High Cramer's V values indicate strong relationship, with the maximum being 1 and the minimum zero. In the following table, these statistics are presented for each cross-tabulation (Table 5).

Chi^2 values indicate significant relationships (at α=0.05) between impacts and attributes, apart from Drainage Area, which appears to be significantly independent of impacts. However, Cramer's V values are relatively low, indicating weak relationship between impacts and attributes, especially for Local Upslope, Drainage Area and Wetness index (< 0.1). But these statistics account for the whole cross-tabulation table, so individual relationships between each classes of impacts and attributes should be analysed case by case in order to retrieve as much information as possible.

Impact vs :	Chi^2 2-sided p-value	Cramer's V
Water Movement	0	0.22
Water Depth	0	0.18
Flood Return Period	0	0.15
Land Use	0	0.13
Pop. Density	0	0.19
Local Upslope	0.006	0.08
Drainage Area	0.063	0.07
Wetness index (CTI)	0.004	0.08

Table 5: Summary of statistical tests (Cramer's V and Pearson's Chi^2 2-sided asymptotic significance [p-value]) for each Impact vs Attritube cross-tabulation.

Bar charts representing the *deviation from attribute total %* have been choosen to illustrate the relationship between impacts and spatial attribute classes. Note that the three first cross tabulation analyses (Fig. 2, 3 and 4) are those conducted with the SHAVE data themselves.

The Impact versus Water Movement chart (see Fig. 2).
Strong positive signals (>10% dev.) show that Moving Water is related with Overflow, Street/Road, Road Closure, Evacuation, Stranded Cars and Rescue. On the other hand, strong signals (>10% dev.) indicate that Standing Water is associated with Greenlands and Inundation. These relationships seem consistent with the impacts context, i.e. standing water for *flat, low lying* impacts and running water for impacts associated to overflow/runoff and most severe ones. As these results make sense with impact contexts, this is a first element showing that impacts were correctly classified.

These moving/standing results will be complemented wih the next cross-tabulation : Impact versus Flood Depth (see Fig. 3).

The Overflow, Road Closure, Evacuation, Stranded Cars and Rescue impact classes show strong signals for the $>$ 30cm bin, with deviations over $>$ 10%. The three most severe impacts are then related by the interviewees to high waters, as well as rivers out of their banks and road closures (which are often associated with overflows on nearby roads or low-water crossings). These associations then make sense with these impact contexts. And if previous results are added, these impacts are associated to high *and* moving waters, representing severe hazards. Note that Overflow is not a severe impact, but there may be little vulnerability (see next associations to low density/rural areas).

Furthermore, Inundation is strongly not associated ($\leq$ -10% dev.) with high flood waters. This can be explained by the fact that this classification also includes basement flooding. Also, recall the previous association with standing waters. Street/Road shows a positive

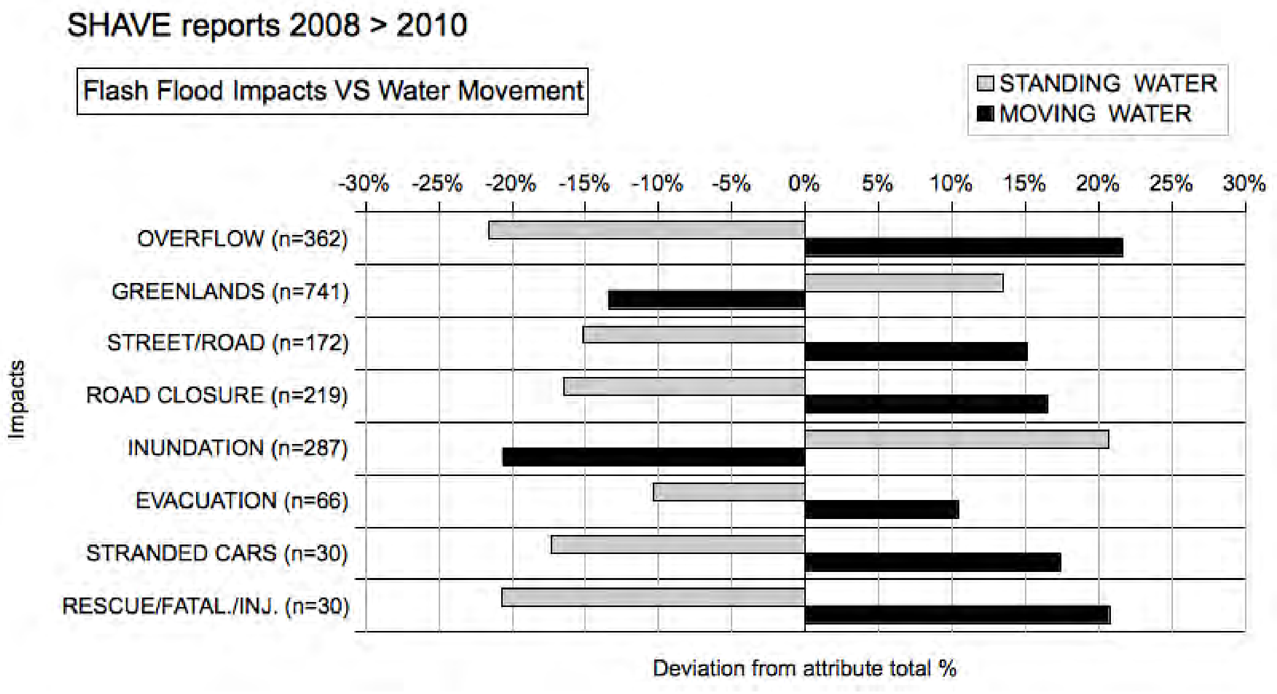


Figure 2: SHAVE impacts versus Water Movement cross-tabulation. Bar chart representing deviation from attribute total %.

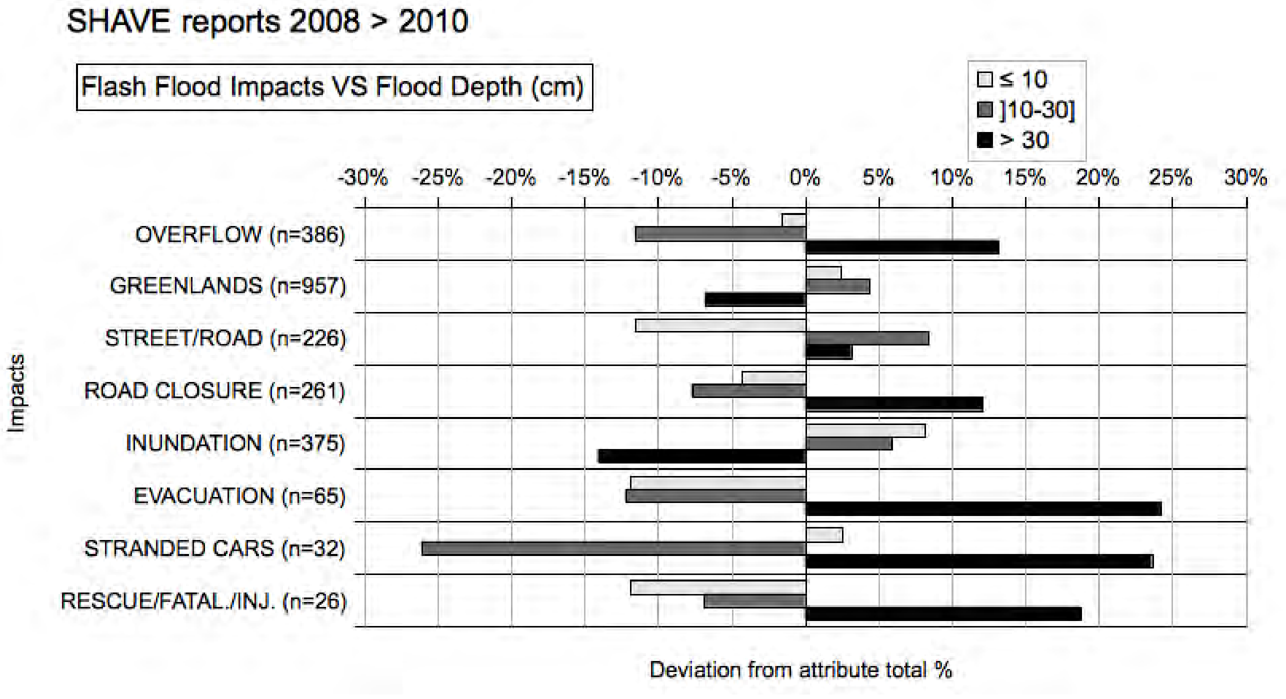


Figure 3: SHAVE impacts versus Water Depth cross-tabulation. Bar chart representing deviation from attribute total %.

signal (> 5 % dev.) for the 10-30cm bin and strongly negative (< -10% dev.) for shallow floodwaters. This association with intermediate water depths (in movement) could be possibly related to a runoff context. Finally, Greenlands shows a weak positive signal for the two ≤ 30cm bins (< 5% dev.), but is clearly not linked (< -5% dev.) to high waters. This result makes sense with the previous standing water association, if we imagine a situation with inundating waters in croplands.

The last perceived attribute chart is Impact versus Flood Return Period (see Fig. 4). Overflow impacts are associated with 1-10years return period (> 5% dev.), whereas Greenlands and Street/Roads are linked (> 5% dev.) to frequent events (≤ 1year). However, Road Closure and Inundation show weak signals (< 5% dev.), so interviewees equaly associate these impacts to frequent and rare events. Finally, Evacuation, Stranded Cars, and Rescue are mostly associated with rare events (note the contradictory signal for Stranded Cars). For this SHAVE attribute, signals are sometimes weak, or unexpected (Stranded cars). It may be due to people's perceptions, knowledge or age. These results must also be taken cautiously because of the smaller sample size for severe impacts. In the end, they show that people are moderately able to evaluate flood frequency, but there is still a general tendency showing that the most severe the impact is, the rarest the event.

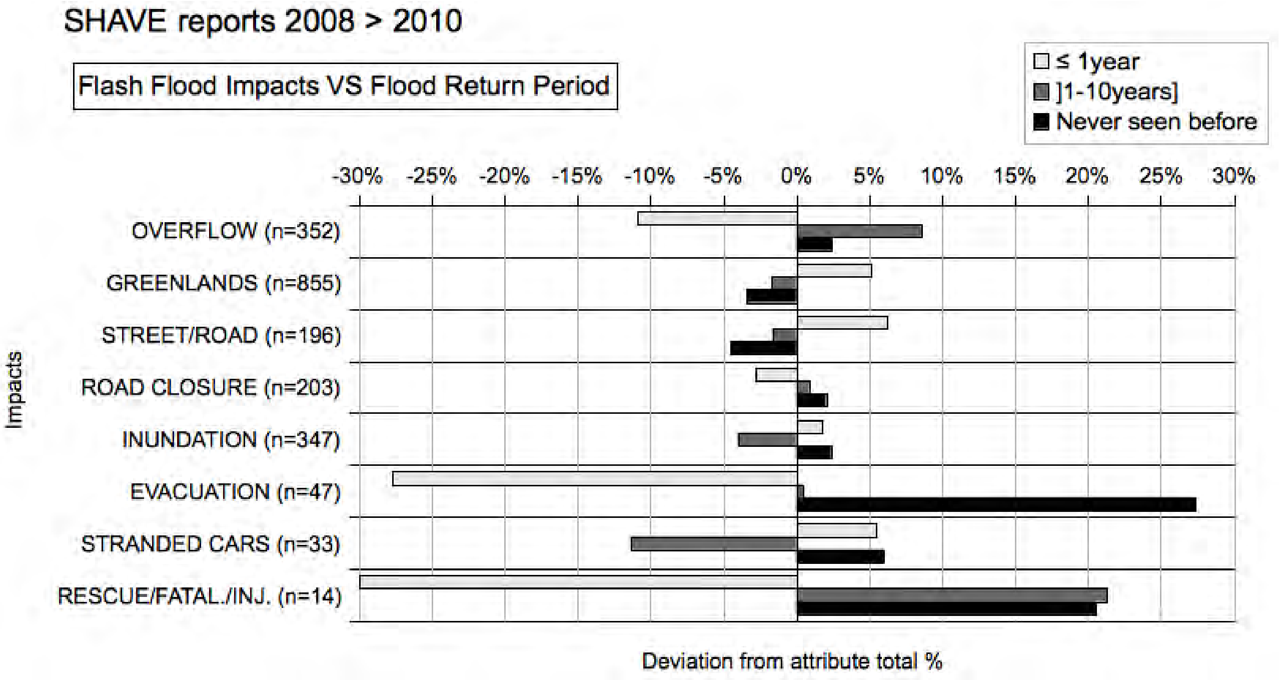


Figure 4: SHAVE impacts versus Flood Return Period cross-tabulation. Bar chart representing deviation from attribute total %.

The following cross tabulation analyses (Fig. 5, 6, 7 and 9) are now those concerning the independent, GIS-sampled dataset :

The Impact vs Land Use chart (see Fig. 5).

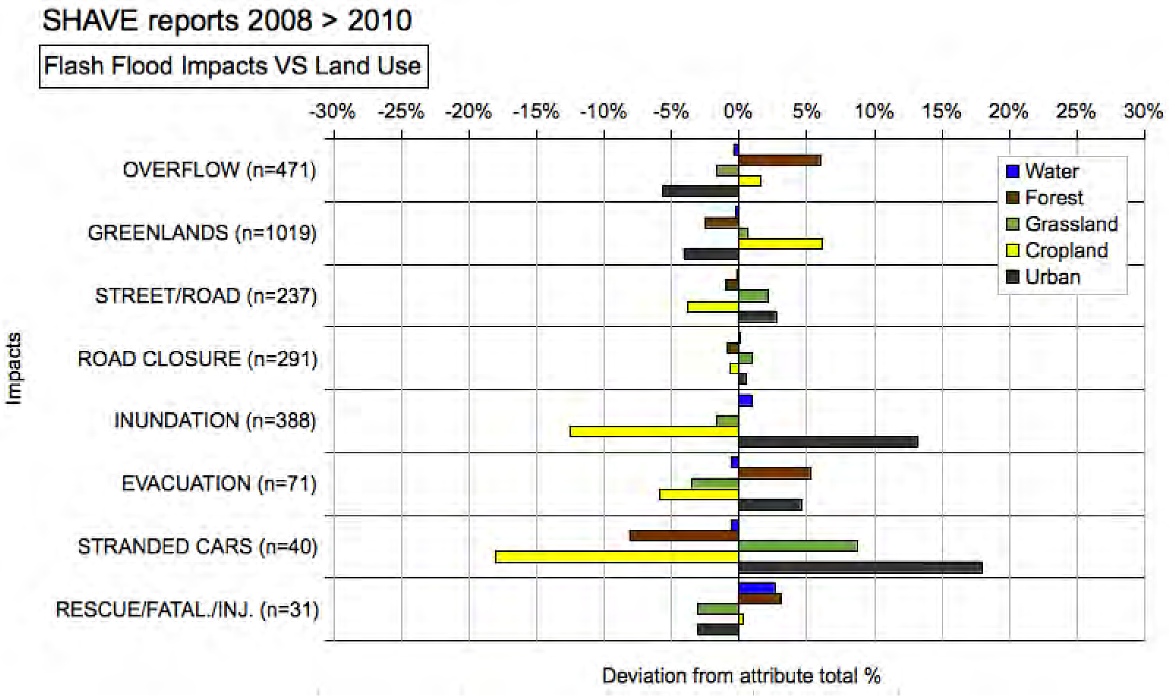


Figure 5: SHAVE impacts versus Land Use cross-tabulation. Bar chart representing deviation from attribute total %.

Overflow impacts are mostly associated with Forest (> 5% dev.) and show a negative signal for the Urban bin (< -5% dev.). Overflow then appears to be more of a rural impact. Furthermore, Greenlands flooding is associated with Cropland (> 5% dev.), which makes sense. Street/Roads and Road Closure show weak signals (both below 5% dev.). These impacts seem to happen both in rural/urbanized areas. Nevertheless, there is a slightly negative association with Cropland. For the Inundation impact, there is a strong positive signal for Urban (> 10% dev.). This result (along with Greenlands associated to Cropland) confirms that the impact classification is consistent. The Evacuation impact shows a negative signal for Cropland (< -5% dev.) and positive signals (but relatively weak : around 5% dev.) for both Forest and then Urban. Evacuation is then not especially linked to urbanised areas, as we would expect, but also occurs in rural zones. The Stranded cars impact is strongly associated to Urban (> 15% dev.) and also linked to Grassland (> 8% dev.). This result reflects well the classification of this impact, which includes cars stalled in parking lots (Urban) or ditches (Grassland). Finally, the Rescue impact does not show any clear tendency (all bins bellow 5% dev.), which indicates that these observed Rescues, Fatalities and Injuries are not linked to a particular type of land use.

The Impact vs Population Density chart (see Fig. 6).

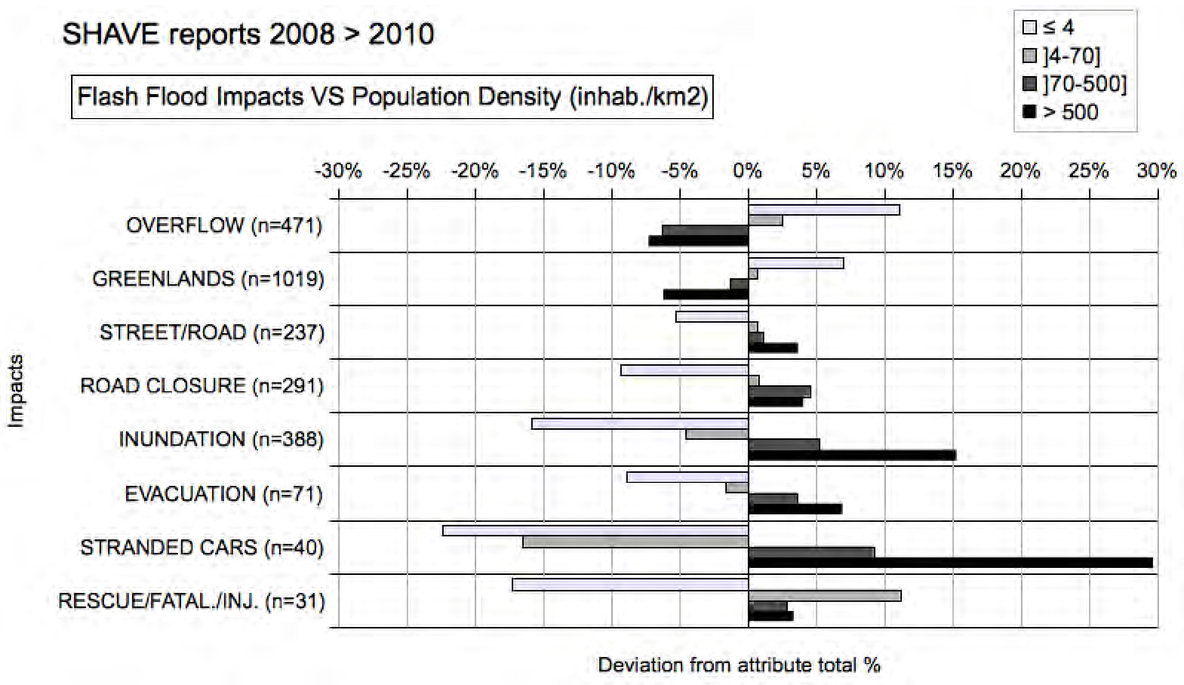


Figure 6: SHAVE impacts versus Population Density cross-tabulation. Bar chart representing deviation from attribute total %.

Before going into details, note that this chart presents clear trends for almost every impact category : going from the lowest to the highest population density bins, deviations evolve progressively from minimum to maximum values. Moreover, recall that this population density information was collected independently from the SHAVE reports.

Overflow and Greenlands classes are associated to sparsely populated areas (strong positive signals ($>$ 10% and $>$ 5% dev.) for densities $\leq$ 4 inhab/km^2). This is consistent with the previous *rural* association. Street/Road and Road closure does not show strong positive signals ($<$ 5% dev.), but there is a really clear evolution from strongly negative on sparsely populated areas ($<$ -5% dev.) to positive towards dense areas. Street/Roads has a maximum postive signal for very dense, possibly urban areas ($>$ 500 inhab/km^2) whereas Road Closure has its maximum slightly more towards medium/heavily populated areas (70-500 inhab/km^2). Road closure and Street/Roads were not linked to a particular land use, but are now associated to denser inhabited zones ($>$70 inhab/km^2). This result seems logical if we consider that urbanisation is built along roads, or the other way round. Inundation and Stranded Cars show very strong association ($>$ 15% dev.) with heavily populated areas ($>$ 500 inhab/km^2). This is consistent with their association to urban land use. Evacuation is also associated to heavily populated areas, but there is a weaker signal than for the two previous impacts (even if > 5% dev.). This less strong association can be correlated with its previous correlation with both Forest and Urban land uses. This result shows that Evacuation might happen in populated areas, but not necessarily in very dense cities. Finally, Rescue is mostly related to low population density (> 10% dev.). This association is very interesting, as the most severe impact then seems to occur in less dense areas.

The Impact vs Local Upslope chart (see Fig. 7).

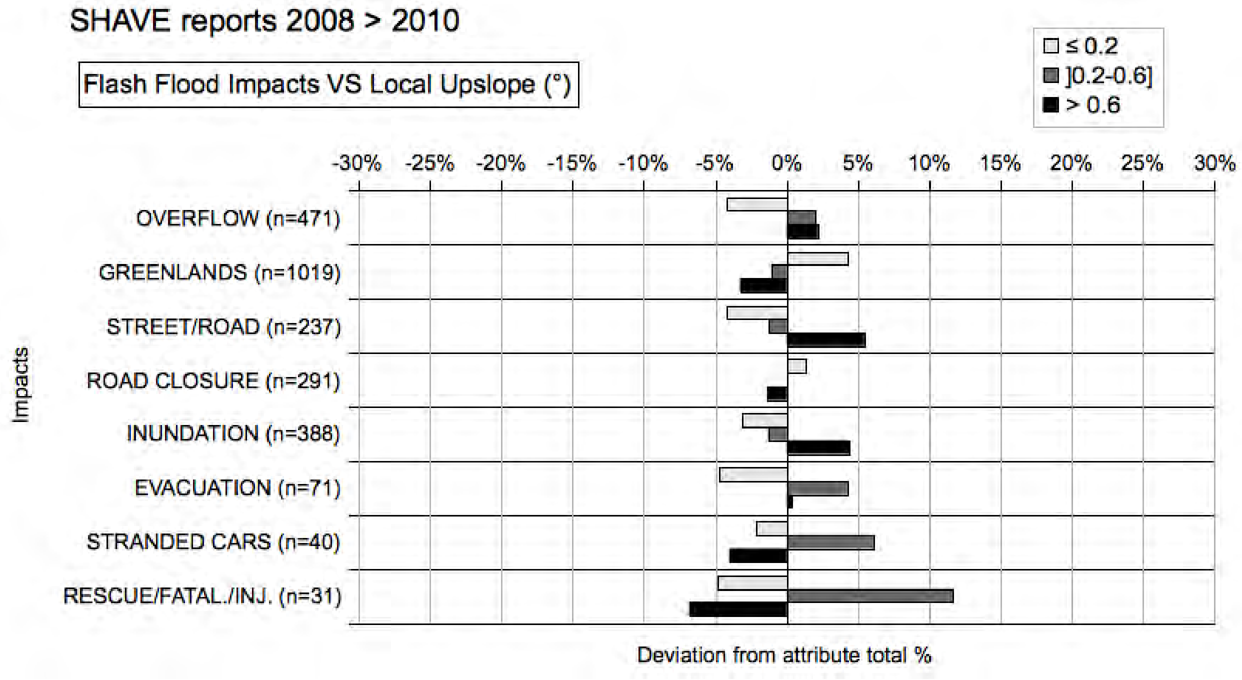


Figure 7: SHAVE impacts versus Local Upslope crosstabulation. Bar chart representing deviation from attribute total %.

In general Local Upslope associations show relatively weak signals. Moreover, note the overall low range of slope values (see Table 3). Note also that these values are computed from the maximum change in the elevations between each gridcell (i.e., on a km^2) and its eight neighbors.

Nevertheless, for Overflow, Street/Road and Inundation, there is a tendency towards higher slopes and for Greenlands, towards lower slopes. For Road closure, no signal at all. Finally, Evacuation, Stranded Cars and Rescue (the most severe impacts) are associated to the intermediate slope bin (0.2-0.6°).

The Impact vs Drainage Area chart (see Fig. 8).

Overflow, Greenlands, Road Closure and Inundation show weak signals, but note a positive tendency (even if weak : < 5% dev.) for Road Closure towards larger drainage areas.

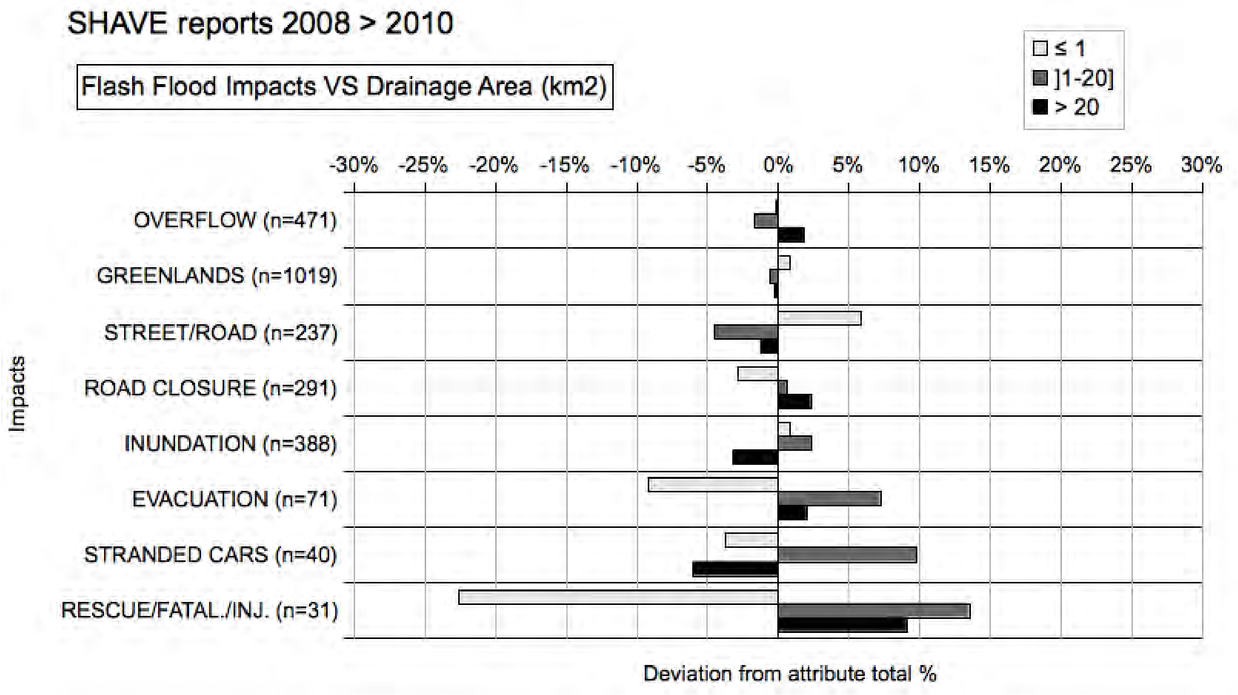


Figure 8: SHAVE impacts versus Drainage Area crosstabulation. Bar chart representing deviation from attribute total %.

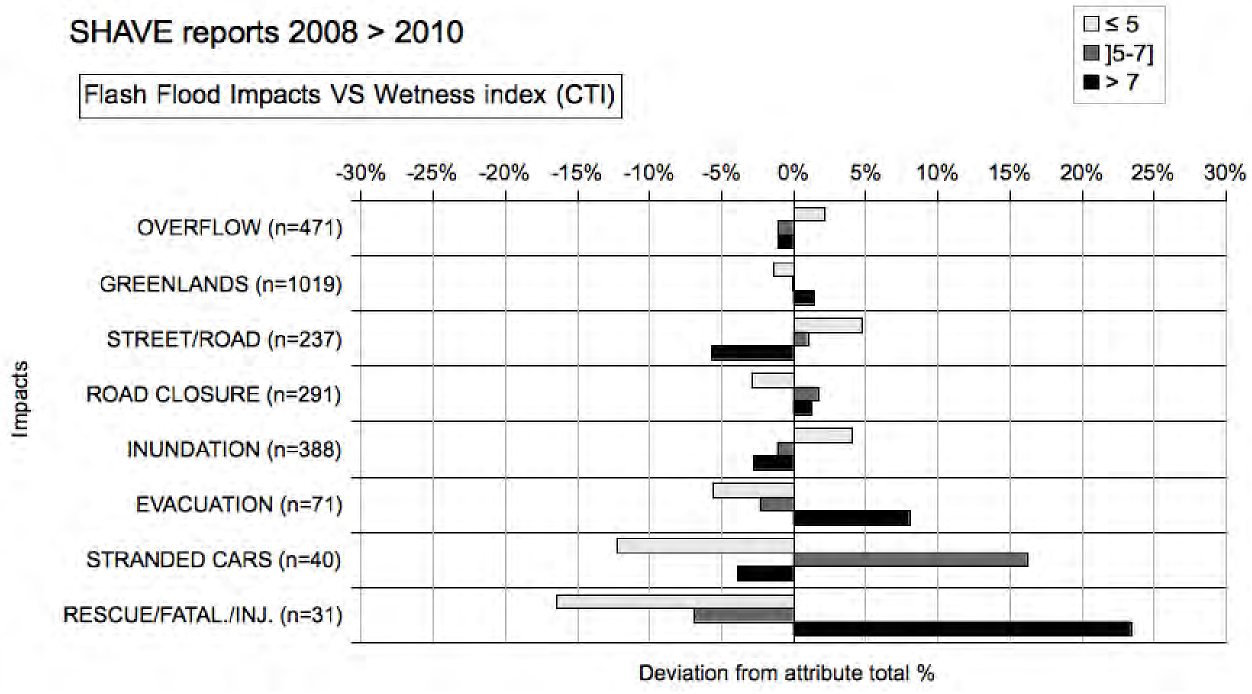


Figure 9: SHAVE impacts versus Wetness index (CTI) crosstabulation. Bar chart representing deviation from attribute total %.

However, Street/Roads flooding is associated (> 5% dev.) with smaller drainage areas ($\leq 1km^2$), possibly linked to runoff. Evacuation, Stranded Cars and Rescue (the most severe impacts) are associated to the intermediate bin (> 5% dev.) of drainage areas between 1 and 20 km^2, possibly related to overflows in upstream catchments.

The Impact vs Wetness index (CTI) chart (see Fig. 9).
First, Overflow, Greenlands, Road Closure and Inundation show weak signals (as for drainage area). But here, note a positive tendency (even if weak : < 5% dev.) for Inundation towards lower CTI values. Street/Roads is mostly associated (almost 5% dev.) with smaller CTI values (see low drainage areas) and has a negative signal (< -5% dev.) for higher values. Again, linked to runoff? On the other side, Evacuation and Rescue are associated to high CTI values (> 7), which represent larger drainage areas and more gentle slopes, leading to higher water accumulation. Finally, Stranded cars is associated with intermediate CTI values]5-7].

3.2.5 Summary table and conclusion

Table 6 is summarizing attributes mostly associated to each impact classes.

	OVER.	GREEN.	ROAD	CLOS.	INUN.	EVAC.	CARS	RESC.
Water Move.	move	stand	move	move	stand	move	move	move
Return Period	]1-10y]	≤1y	≤1y	-	-	n.seen	≤1/n.seen	]1y-n.seen]
Land Use	forest	crop	-	-	urban	forest/urban	urban/grass	-
Pop. Density	≤ 4	≤ 4	>500	>70	>500	>500	>500	]4-70]
Water Depth	>30cm	]10-30cm]	]10-30cm]	>30cm	≤10cm	>30cm	>30cm	>30cm
Local Upslope	-	≤0.2°	>0.6°	-	>0.6°	]0.2-0.6°]	]0.2-0.6°]	]0.2-0.6°]
Drainage Area	-	-	≤1km^2	-	-	]1-20km^2]	]1-20km^2]	]1-20km^2]
CTI	-	-	≤5	-	≤5	>7	]5-7]	>7

Table 6: Table summarizing attributes mostly associated to each impact classes.

A first significant result of this spatio-contextual analysis is that associations found using cross-tabulation are consistent with the impact classification. This is true for perceived attributes already included in the SHAVE dataset, as well as for spatial attributes (at 1km^2 resolution) sampled through GIS. However, topographic variables (i.e.: Local Upslope, Drainage Area and Wetness index) do not show significant results. The $1km^2$ grid resolution used for the present analysis might be to coarse to correctly characterise local topography, which controls hydrological processes. Nevertheless, meaningful results found for the other attributes show that the SHAVE dataset is a trustworthy tool for flash flood characterisation (at least at a $1km^2$ scale), even if it is based on public polls.

A second important finding is that, apart from trivial associations, interesting results emerge from the spatio-contextual analysis :

- Evacuations are not necessarily only observed in urban zones, but also in rural areas.

- Rescues, Fatality or Injuries mostly take place in low population density areas. Moreover, these impacts are not perceived by the interviewees as extreme, rare events (as it is the case for Evacuation) because associated return periods vary from one year to 'never seen before'. This result indicates how people may have difficulties to estimates an event frequency.

4 Evaluation of the ability of US flash flood forecasting tools to predict impacts

Launched in the mid-eighties, the operational flash flood prediction tool in the US is radar-based and relies on the concept of flash flood guidance (FFG). Besides this, alternative approaches to FFG have been recently developed, using spatially distributed land surface and soil characteristics maps (the *Gridded* FFG) as well as distributed hydrological models.

In this section, the impact-classified NWS and SHAVE datasets will be used to evaluate the ability of three of these prediction tools (FFG, GFFG and the *Distributed Hydrological Model - Threshold Frequency* [DHM-TF]) to predict flash flood impacts.

4.1 Presentation of the flash flood forecasting tools

4.1.1 Flash Flood Guidance

The concept of flash flood guidance (FFG) is the threshold rainfall over nominal accumulation periods of one, three, and six hours required to initiate flooding on small streams that respond to rainfall within a few hours [Georgakakos, 1986]. In other words, FFG is the basin-averaged rainfall required over a basin to produce flooding at its outlet. One to three times a day, FFG is derived using a hydrologic model taking into account initial soil moisture and stream states. These values, when overlaid with radar's Quantitative Precipitation Estimates (QPE), are used by the forecaster to issue flash flood warnings when observed or forecast rainfall rates exceed the thresholds. FFG is computed in two steps :

First, determining the threshold runoff $[L]$ required to cause flooding (bankfull conditions) at the basin outlet. In the NWS, this value is derived by dividing the estimated 2-year return period flow $[L^3/T]$ by the unit hydrograph peak flow $[L^2/T]$. Threshold runoff values are computed once offline at a resolution down to $5km^2$ basins and are considered static.

Then, a lumped-parameter hydrological model is run under differing basin-averaged rainfall scenarios to yield rainfall-runoff curves over 1-, 3-, and 6-hours accumulation periods, given initial soil moisture and stream states. The method employed at the NWS uses the Sacramento Soil Moisture Accounting model (SAC-SMA) and includes contributing processes such as snowmelt, interception, infiltration, interflow, soil water storage and evapotranspiration. These rainfall-runoff curves are then used in reverse to look up the rainfall rates that correspond to the static threshold runoff values; this is FFG [Gourley et al., 2011b].

Because FFG values are computed at basin scale, a recent development has been made to create a tool at higher spatial resolution : the Gridded FFG.

4.1.2 Gridded Flash Flood Guidance

The general GFFG methodology, proposed by Schmidt et al. [2007], follows that of FFG in that static values of threshold-runoff are first derived to estimate bankfull discharge and are subsequently used to derive rainfall thresholds, which change in response to modelled soil saturation [Gourley et al., 2011b]. The difference here is that threshold-runoff values and rainfall-runoff curves are computed at a grid cell scale, taking into account variability in the land surface and soil types, as well as slope. The nominal resolution of GFFG products is 4km (see Figure 10 for a comparison between maps of FFG and GFFG). Note that GFFG is progressively replacing FFG as operational tool in several US River Forecast Centers, but FFG is still in use for the major part of the US. For this reason, FFG and GFFG products are hardly available simultaneously for a particular area.

4.1.3 Distributed Hydrological Model - Threshold Frequency

Developed by Reed et al. [2007], the Distributed Hydrological Model - Threshold Frequency (DHM-TF) deviates from FFG in that it uses observed or forecast rainfall as direct forcing to a hydrological model, rather than determining the rainfall thresholds in scenario mode [Gourley et al., 2011b]. The method consists of running a distributed hydrologic model at each grid point using historical rainfall historic data. This way, simulated runoff can be assigned to grid cells where discharge observations are not available. Then a flood frequency analysis (assuming a log-Pearson Type III distribution) is used to compute flows that correspond to return periods of one, two, five yr, etc. In forecast mode, DHM-TF is forced with real-time, radar-based rainfall. Exceedance of simulated flows over the threshold return period flows (in this study, a 2-year return period flow) is the basis for alerting on an impending flash flood [Gourley et al., 2011b].

4.2 Presentation of the flash flood case studies

In this analysis, two flash flood case studies (considered as extreme events) were chosen, for which at least two of the three forecasting tools were available. They are: the flash floods

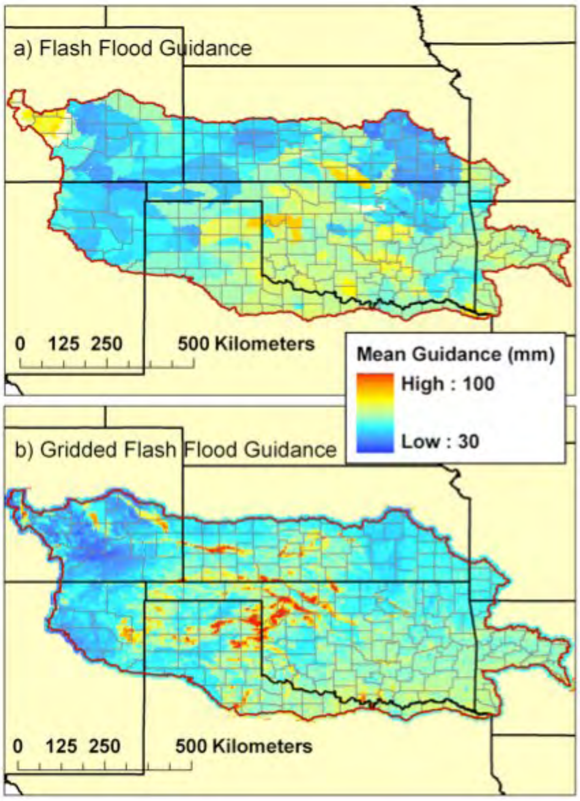


Figure 10: Average values of a) FFG and b) GFFG corresponding to 1-h accumulation period over Arkansas-Red River Basin from 01 September 2006 to 22 August 2008 [Gourley et al., 2011a].

caused by the Erin storm over the state of Oklahoma in 2007 and the Oklahoma City flash flood event of 2010. These events occurred at different spatio-temporal scales (see Fig. 11):

- The first flash flood case study was caused by the remains of the tropical storm Erin, which crossed the state of Oklahoma from West to East over two days (August 18th to 20th) in 2007. Rainfall rates of over three inches (76mm) per hour were common, with significant flash flooding reported in numerous counties. Rainfall amounts exceeded five inches (127mm) over a large area, with some locations receiving eight to ten inches (203 to 254mm).

- The second flash flood case is at a smaller spatio-temporal scale. It was more of an urban event, occurring June 14th 2010 over the Oklahoma City area. A first round of significant rain impacted central Oklahoma around 3 am. This round moved east before another, longer lived, thunderstorm complex developed over the Oklahoma City metro area. Rainfall rates averaged one to two inches (25-50mm) per hour, with some thunderstorm bands producing rates near three inches (76mm) per hour. A total of 5-9inches (127-228mm) was reported over the area, with up to 12inches (305mm) over the north-central portion of Oklahoma City.

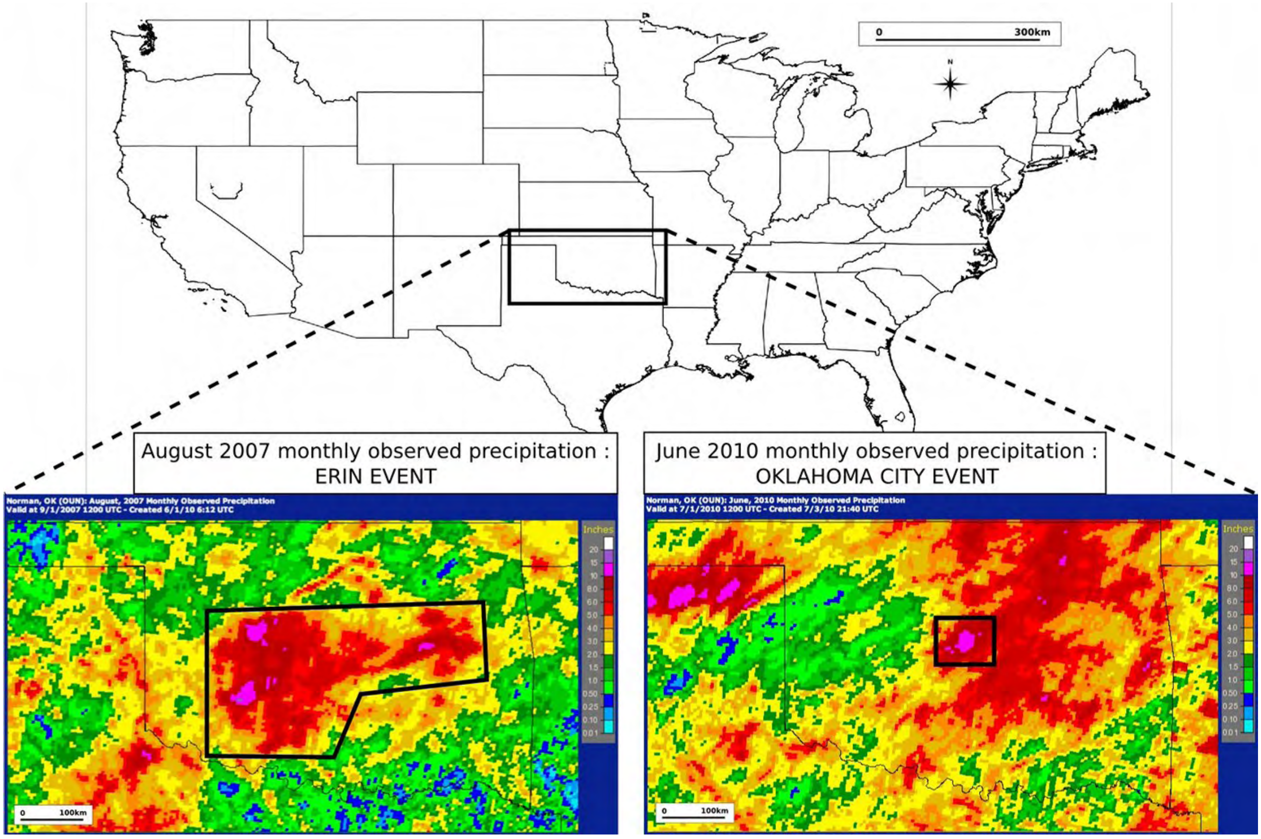

0 300km
N
August 2007 monthly observed precipitation :
ERIN EVENT
June 2010 monthly observed precipitation :
OKLAHOMA CITY EVENT
Norman, OK (OUN): August, 2007 Monthly Observed Precipitation
Valid at 9/1/2007 1200 UTC - Created 6/1/10 6:12 UTC
Norman, OK (OUN): June, 2010 Monthly Observed Precipitation
Valid at 7/1/2010 1200 UTC - Created 7/3/10 21:40 UTC
Inches
20
15
10
8.0
6.0
5.0
4.0
3.0
2.0
1.5
1.0
0.50
0.25
0.10
0.01
0 100km

Figure 11: Presentation of the two case studies : maps of observed monthly precipitation. Source : National Weather Service, http://water.weather.gov/precip/.

A description of the spatio-temporal domain of the meteorological events associated to these flash floods as well as the available impacts and forecasting datasets is included in Table 7 :

	Erin event	**Oklahoma City event**
Approximated start time	August 18th 2007 17:00 UTC	June 14th 2010 14:00 UTC
Approximated end time	August 20th 2007 7:00 UTC	June 14th 2010 21:00 UTC
Spatial extent	300x200km	50x60km
Available impact datasets	NWS (points)	NWS (polygons) / SHAVE
Available forecasting tools	FFG / GFFG / DHM-TF	GFFG / DHM-TF

Table 7: Table summarizing spatio-temporal domains and available impacts and forecasting products for both case studies.

4.3 Preparation of the flash flood forecasting tools datasets

In this analysis, 1-hour accumulation FFG and GFFG were chosen, rather than 3- or 6hours, as they showed better skill when using observed NWS data (see Gourley et al. [2011a]). To be sure not to miss the events, these hourly FFG, GFFG and DHM-TF products were collected over the whole Arkansas-Red River Basin (see Fig. 10) and over a time window extending from 8 hours prior to the meteorological event to 2 hours after. Quantitative Precipitation Estimates (QPE), taken from the hourly multi-sensor Stage IV product (mosaicked from US radars and rain gauges, see: http://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/stage4/) are used to calculate QPE/FFG and QPE/GFFG ratios for every 1-hour grid. Then, grids of maximum 1-h QPE/FFG, QPE/GFFG and DHM-TF return periods were computed, encompassing the whole event duration. Finally, these grids of maximum forecasting tool values were sampled for each impact point, using circular point clusters with a radius of 7.5km for the Erin case and 1.5km for the Oklahoma City case (see Fig. 12 for an illustration of this sampling cluster). The selected tool value associated to the impact is the maximum value sampled by the cluster.

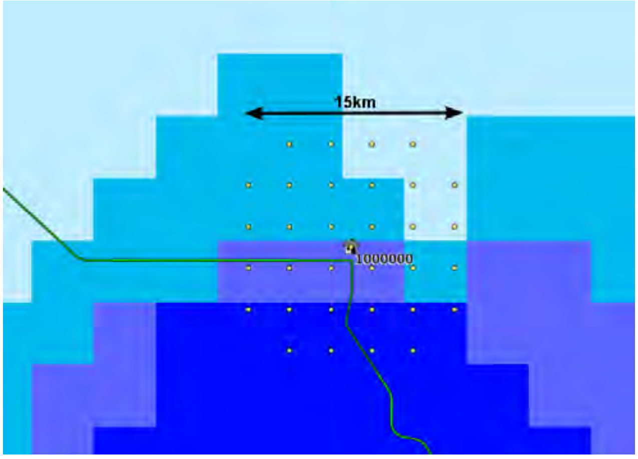


Figure 12: Sampling cluster used for the Erin case.

Note that bank-full conditions are met starting from QPE/FFG-GFFG ratios of 1, or a DHM-TF return period of 2 years. These limits will be used to define if a flash flood event is forecast by the tools or not, in order to populate contingency tables (see Fig. 8).

	Forecast	Not forecast
Observed	*hit*	*miss*
Not observed	*false alarm*	*correct negative*

Table 8: Contengency table, for forecasting.

Three statistics were then computed (when possible) from the hits, misses, and false alarms in each of the contingency tables:

- The *Probability Of Detection* (POD) describes the fraction of observed flash floods that were correctly forecast (eq. 1).

$$POD = \frac{hits}{hits + misses} \tag{1}$$

A POD of 1 indicates all flash floods were correctly forecast while 0 indicates no flash floods were detected by the forecast tools.

- The *False Alarm Ratio* (FAR), which describes the fraction of forecast events that were not associated with observed events (eq. 2).

$$FAR = \frac{false\ alarms}{hits + false\ alarms} \tag{2}$$

Similar to POD, FAR ranges from 0 indicating no forecast events went unobserved to 1 indicating all forecast flash floods were not associated with an observed event.

- The *Critical Success Index* (CSI) combines both aspects of POD and FAR and thus describes the skill of a forecast system (eq. 10).

$$CSI = \frac{hits}{hits + misses + false\ alarms} \tag{3}$$

CSI ranges from 0, indicating no skill, to 1 for perfect skill.

4.4 Results and discussion

4.4.1 The Erin storm event

For the Erin case, only NWS, point-based impacts are available. A symbology was created for each impact. Furthermore, to better illustrate the multi impact aspect, the first and second impact are symbolised by white squares and circles, respectively. Property damage estimations are also included as labels. A first map compares flash flood impacts and Population Density (Fig. 13) and the next three compare impacts with 1-h QPE/FFG ratios (Fig. 14), 1-h QPE/GFFG ratios (Fig. 15) and DHM-TF return periods (Fig. 16). Primary roads and majors streams are also included in each figure.

The map representing impacts and population density (Fig. 13), was created with the idea to identify factors of possible vulnerability linked to impacts. The map shows that impacts are not necessarily located in areas of high population densities, but are clearly situated along primary roads and/or major streams.

A first analysis is done on a YES/NO event basis, by studying the forecasting tool maps and then by computing statistics. In general and before looking at statistics, we can see that all flash flood forecasting tools correctly locate the global area impacted by flash flood.

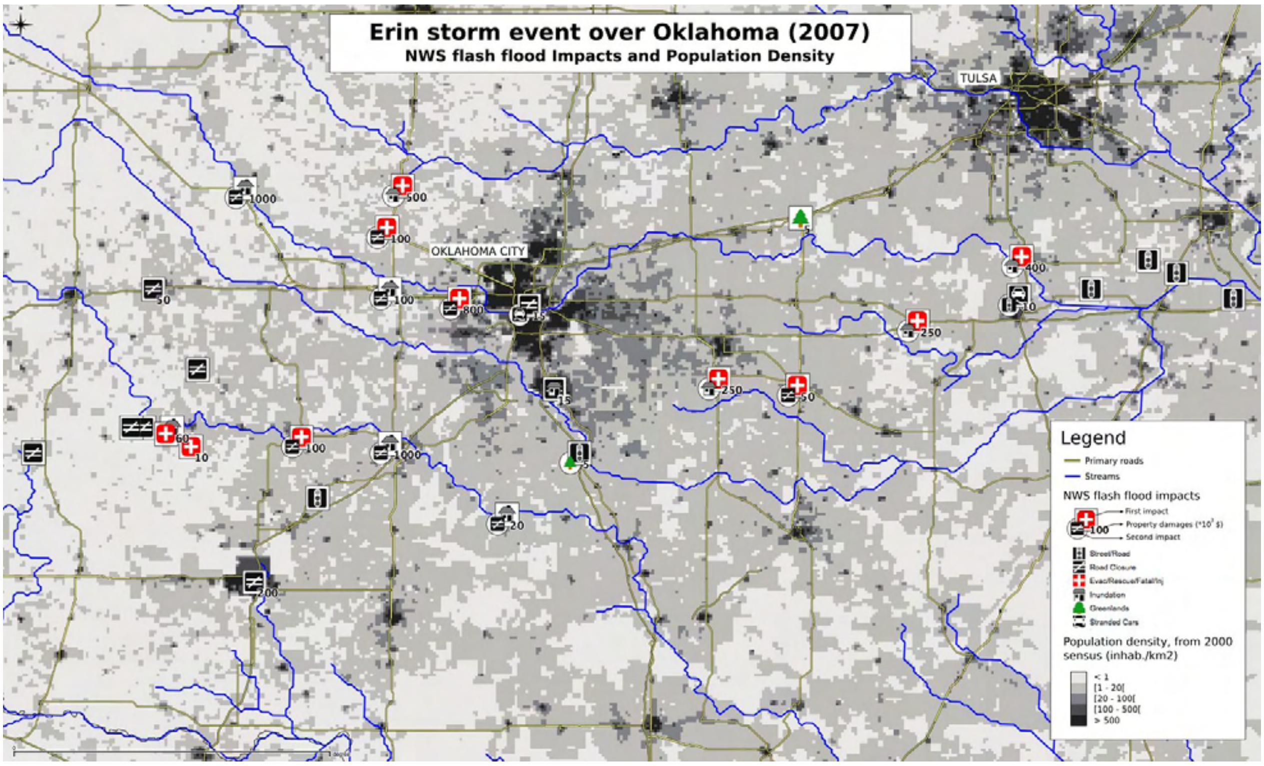


Figure 13: Map of NWS flash flood Impacts and Population Density for the Erin case.

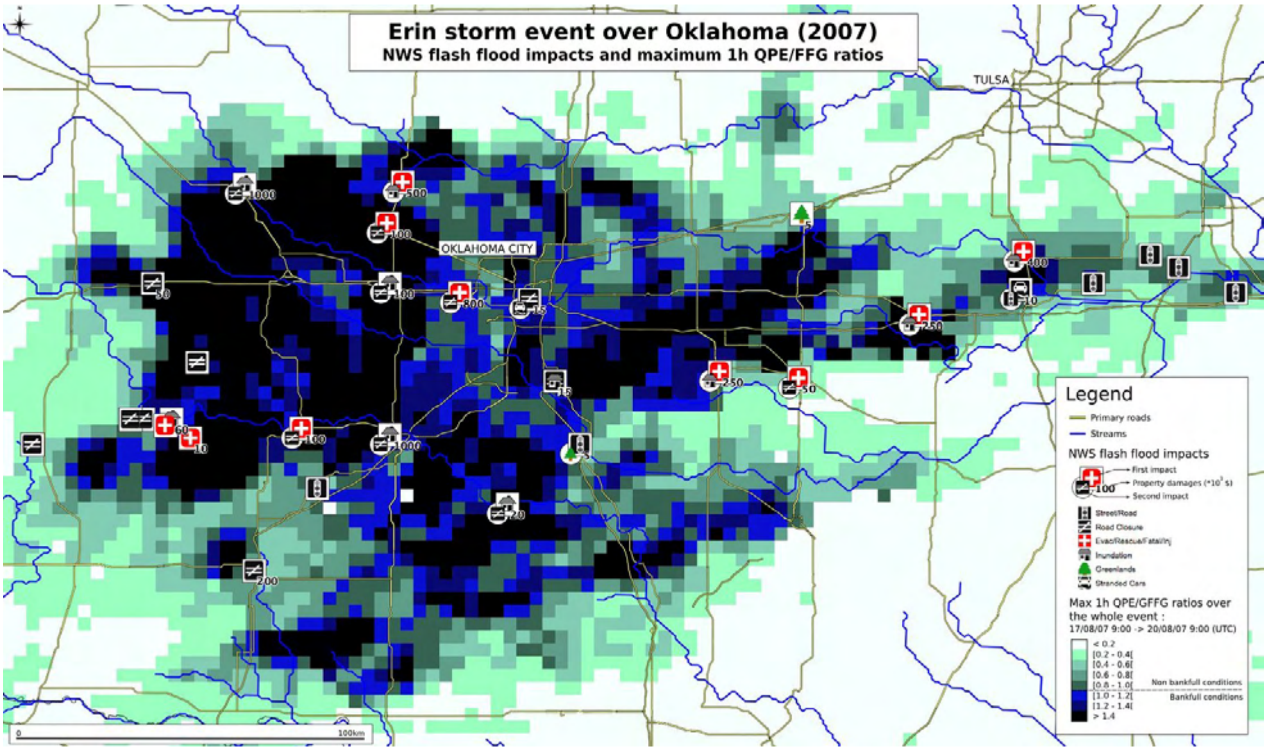


Figure 14: NWS flash flood Impacts and maximum 1h QPE/FFG ratios for the Erin case.

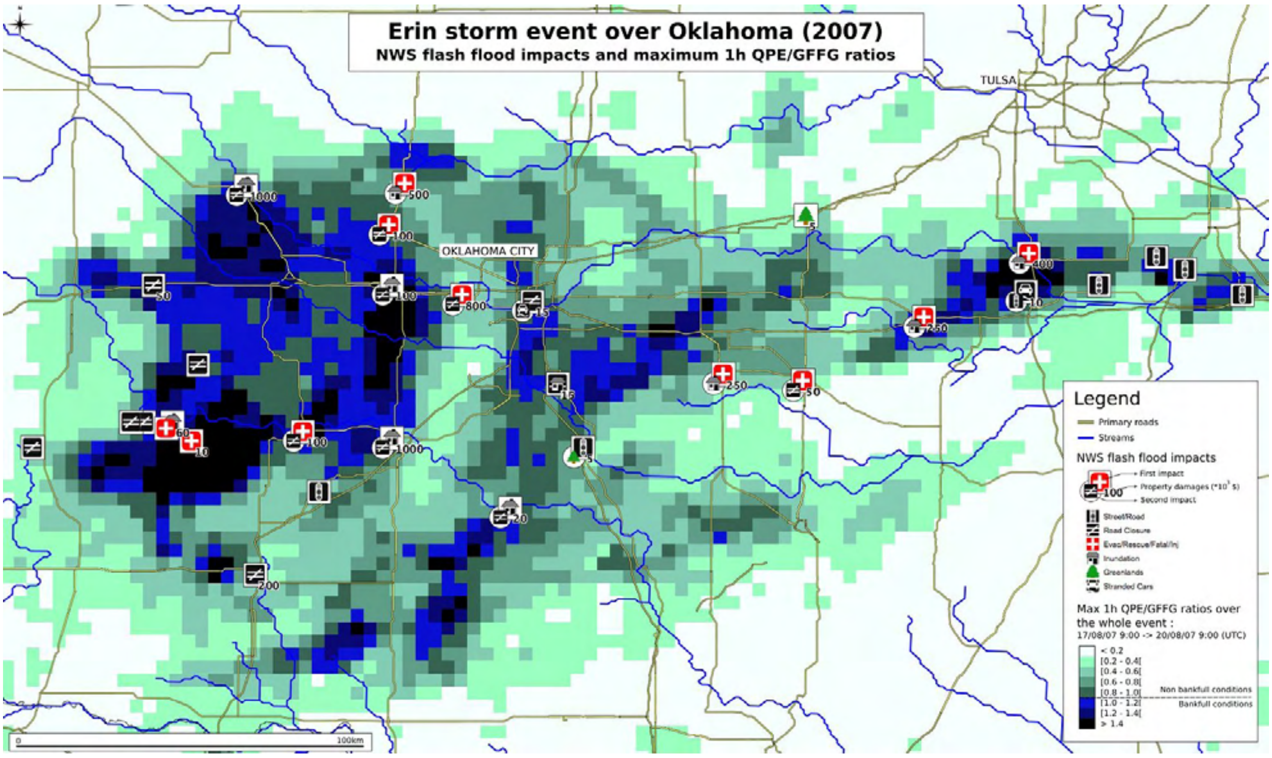


Figure 15: NWS flash flood Impacts and maximum 1h QPE/GFFG ratios for the Erin case.

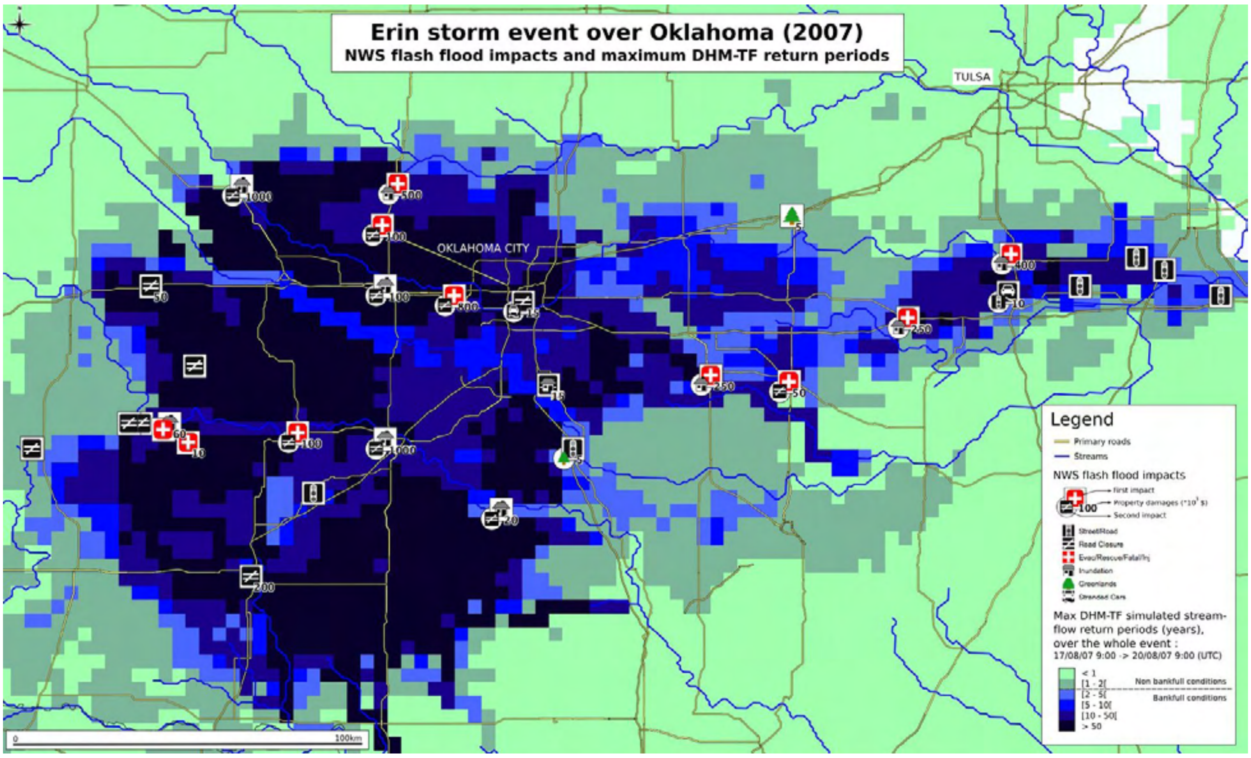


Figure 16: NWS flash flood Impacts and maximum DHM-TF return periods for the Erin case.

Note that GFFG is the tool identifying the smaller impacted areas (in blue colours) compared to the two other tools, and it is missing a few impacts. But for FFG and DHM-TF, even if they detect almost all impacts, they also forecast large areas where no NWS impacts are observed. This could mean that either:
- the tools are overestimating flash flood impacted zones, or
- these zones were impacted but no NWS report was collected, or
- these zone were experiencing flash floods but there was little vulnerability, so no impact.

Because the NWS dataset does not include null reports (so no information about false alarms), the Probability Of Detection was the only statistics that could be computed for the three flash flood forecasting tools (see Table 9). The tool having the best detection skill is DHM-TF, followed by FFG then GFFG. However, these high detection skills may go along with high False Alarm Ratio values, which unfortunately can not be estimated.

Erin event	POD
1-h QPE/FFG	0.94
1-h QPE/GFFG	0.78
DHM-TF	1

Table 9: POD results for Erin impacts sampling.

Recall also that because tools were sampled by taking the maximum of a circular point cluster, it artificially increases the POD.

A second analysis was done by comparing sample values for each tool, as function of impacts. A line delimiting the bank-full and non bank-full conditions (a detected flash flood or not) is added to the following graphs.

On the impact versus 1-h QPE/FFG graph (Fig. 17), the average ratio per impact (black squares) are on the 'detected' zone for all impacts. Note that the distribution of sampled tool values is very wide (see grey diamonds), but apart from two sampled impacts, all are over bank-full conditions. Also, average ratios per impact seem to be divided in two groups : the four most severe impacts have larger ratios (around 2) whereas the two less severe have lower average values (around 1.4), meaning that in this particular case, the FFG tool is able to make a distinction between non severe and more severe impacts.

On the impact versus 1-h QPE/GFFG graph (Fig. 18), more impacts remain non detected, compared to the FFG tool. And the Greenlands impacts are event not detected at all. Also, compared to FFG average ratios per impacts, the GFFG tool show globally lower values (all below 2). But while FFG seemed to distinguish two groups, for GFFG, there seems to be a increase of ratio values, going from the less severe to the most severe impact. But this must be taken with caution, as the Stranded Cars impact has only three sampled

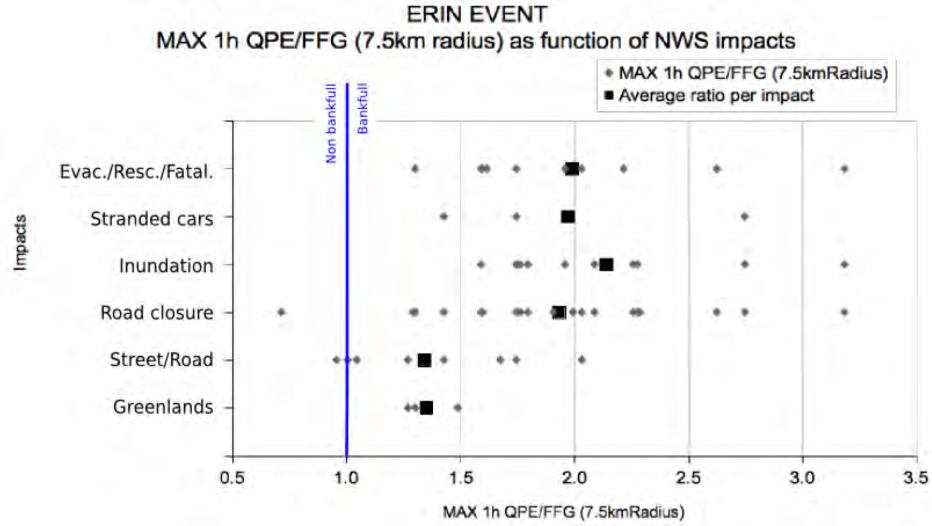


Figure 17: Sampled maximum 1h QPE/FFG ratios as function of impact classes for the Erin case.

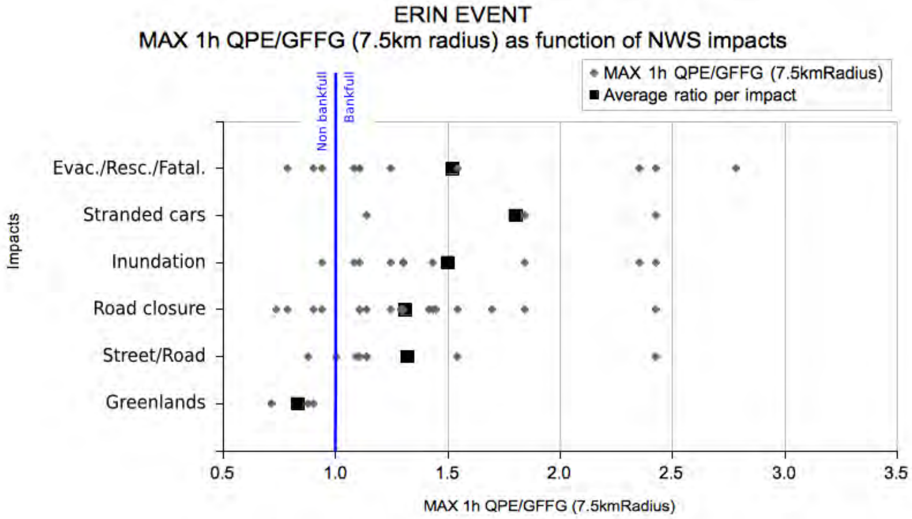


Figure 18: Sampled maximum 1h QPE/GFFG ratios as function of impact classes for the Erin case.

values, so the average value could be biased. Note also the highly spread distribution of sampled ratios (grey diamonds).

On the impact versus DHM-TF return periods graph (Fig. 19), all impact classes have very high return periods (recall that both Erin and Oklahoma City are extreme cases), with average values ranging from 100 to 200 years. Every single impact is sampled with values over the bank-full zone. However, no strong link between impact type and return period is observed, apart from the two less severe impacts (Street/Road and Greenlands), for which the average return period is about a hundred years lower (note the logarithmic scale). Again, note the very wide distributions of sampled tool values (grey diamonds).

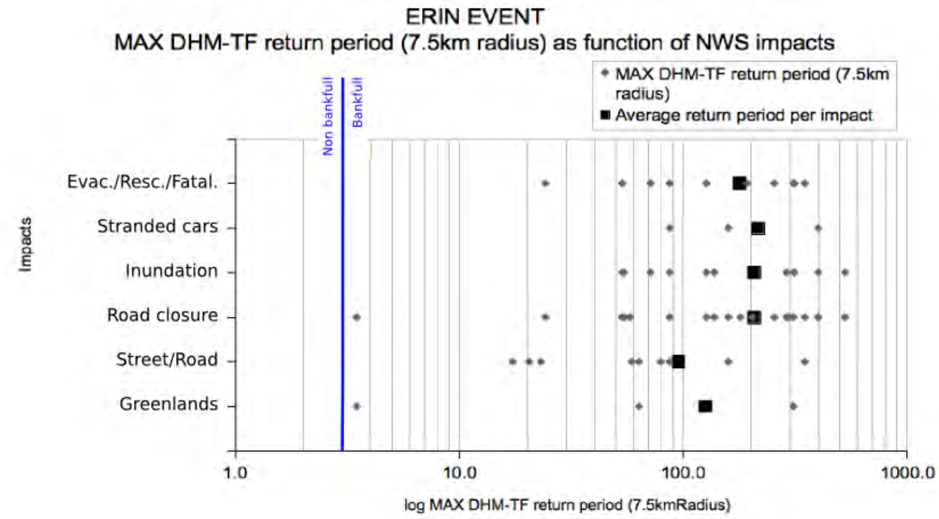


Figure 19: Sampled maximum DHM-TF return periods as function of impact classes for the Erin case.

4.4.2 The Oklahoma City event

For the Oklahoma City flash flood event, only SHAVE impacts (points) were used, because they include null reports (white points on the maps), so that FAR and CSI can be computed. Moreover, in 2010, NWS reports are represented by polygons, which appeared to be inconvenient, as they are often the size of the whole metro area.

As for the Erin case, a map of impacts and Population Density is presented before the two analysed tools, which are in this case 1-h QPE/GFFG ratios and DHM-TF return periods. Then skill analysis is done on a YES/NO event basis (computation of POD, FAR and CSI), before undergoing an impact-based analysis.

The map presenting SHAVE impacts versus Population Density (Fig. 20) shows that the SHAVE sampling covers the whole Oklahoma City metro area. Impacts are not necessarily located along primary roads, neither along major streams. Note that the major river crossing Oklahoma City is regulated.

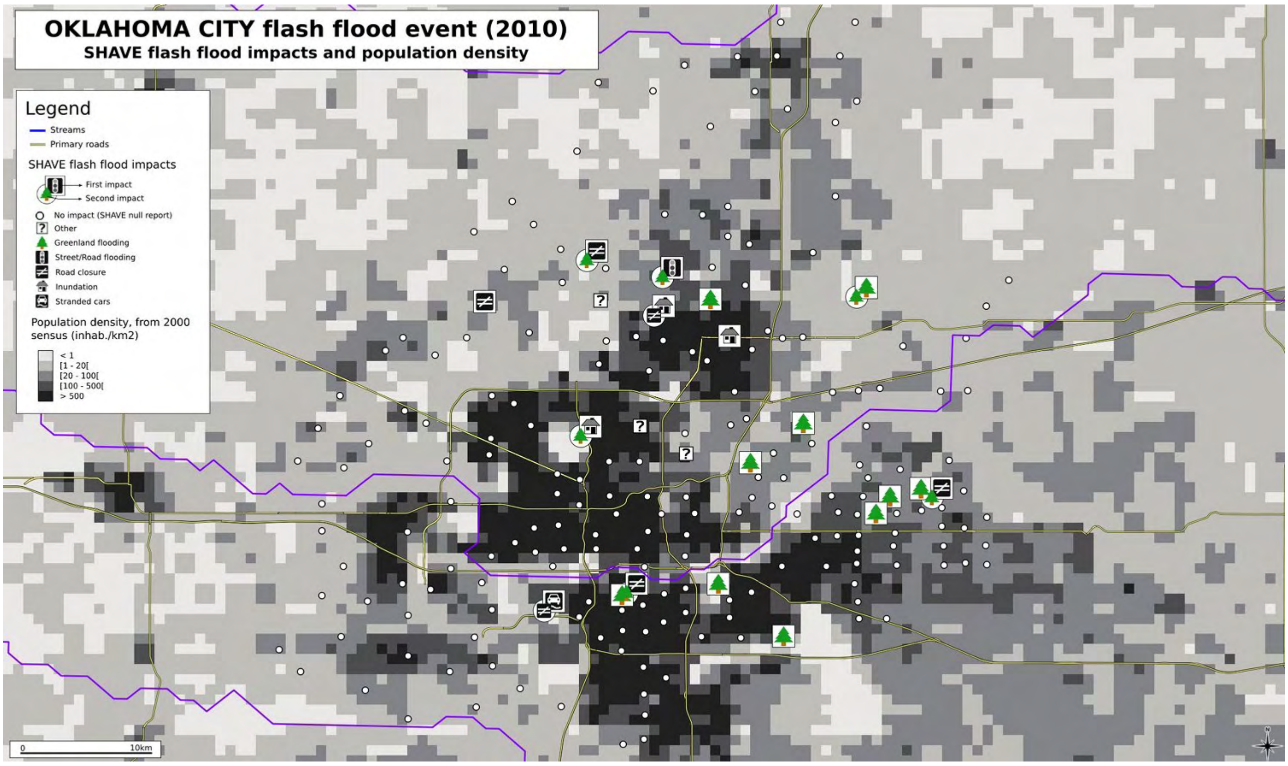

OKLAHOMA CITY flash flood event (2010)
SHAVE flash flood impacts and population density
Legend
Streams
Primary roads
SHAVE flash flood impacts
First impact
Second impact
No impact (SHAVE null report)
Other
Greenland flooding
Street/Road flooding
Road closure
Inundation
Stranded cars
Population density, from 2000 sensus (inhab./km2)
< 1
[1 - 20[
[20 - 100[
[100 - 500[
> 500
0
10km

Figure 20: Map of SHAVE flash flood Impacts and Population Density for the Oklahoma City case.

On the maps of impact versus 1-h QPE/GFFG (Fig. 21) and DHM-TF return periods (Fig. 22), flash flood forecast patterns (in blue) correctly match the global extension of impacts. But at the same time, there is a lot of null reports on forecast grid-cells, indicating numerous false alarms. In order to assess their skills on a YES/NO event basis, POD, FAR and CSI are computed for both tools in Table 10.

Oklahoma City event	POD	FAR	CSI
1-h QPE/GFFG	0.86	0.85	0.14
DHM-TF	1	0.88	0.12

Table 10: POD, FAR and CSI results for Oklahoma City impacts sampling.

Results show that DHM-TF has the highest POD (1), but also the highest FAR (0.88). In the end, it is 1-h QPE/GFFG that shows the best CSI, with a score of 0.14, despite a lower POD value, but thanks to a better FAR. Note that even if the GFFG tool has a better skill than DHM-TF for this particular case, both CSI values are still very low, even if they stay in the same order of magnitude as the highest value found by Gourley et al. [2011a] for the 1-h GFFG tool (0.12), using NWS reports over Arkansas-Red River Basin from 01 September 2006 to 22 August 2008.

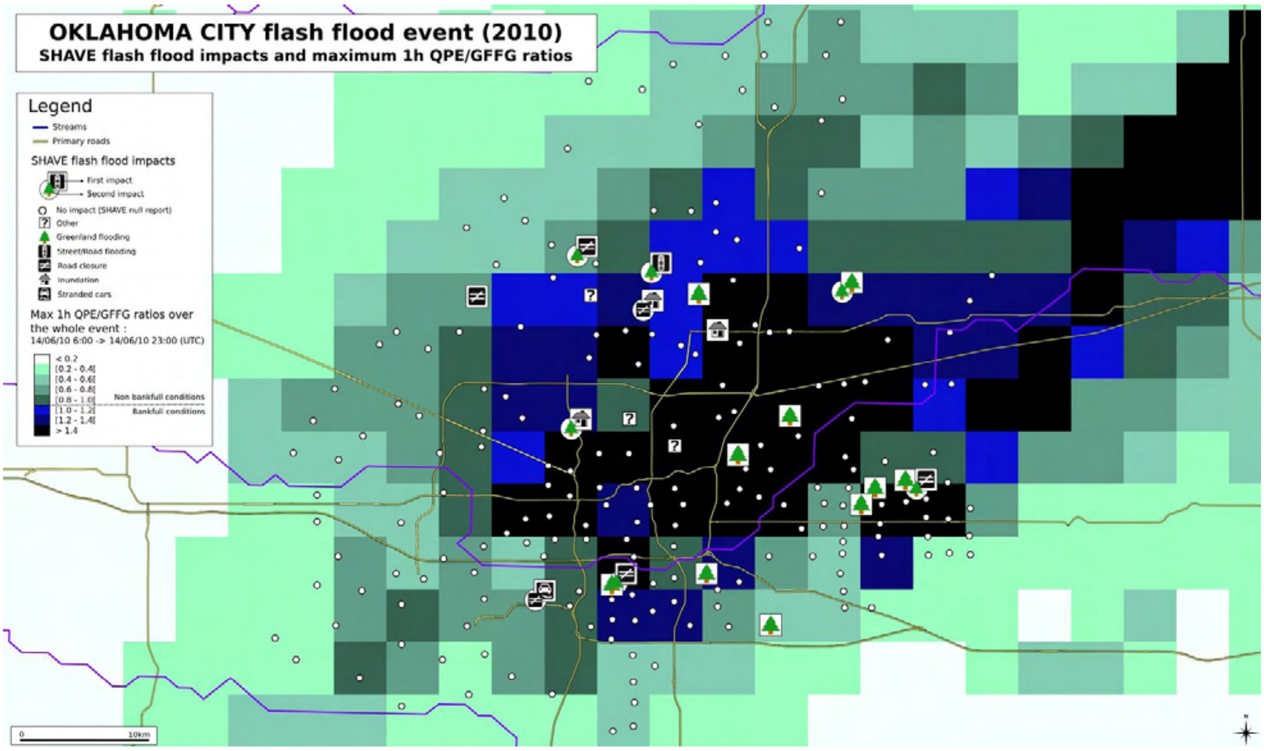


Figure 21: SHAVE impacts and max 1h QPE/GFFG ratios for the Oklahoma City case.

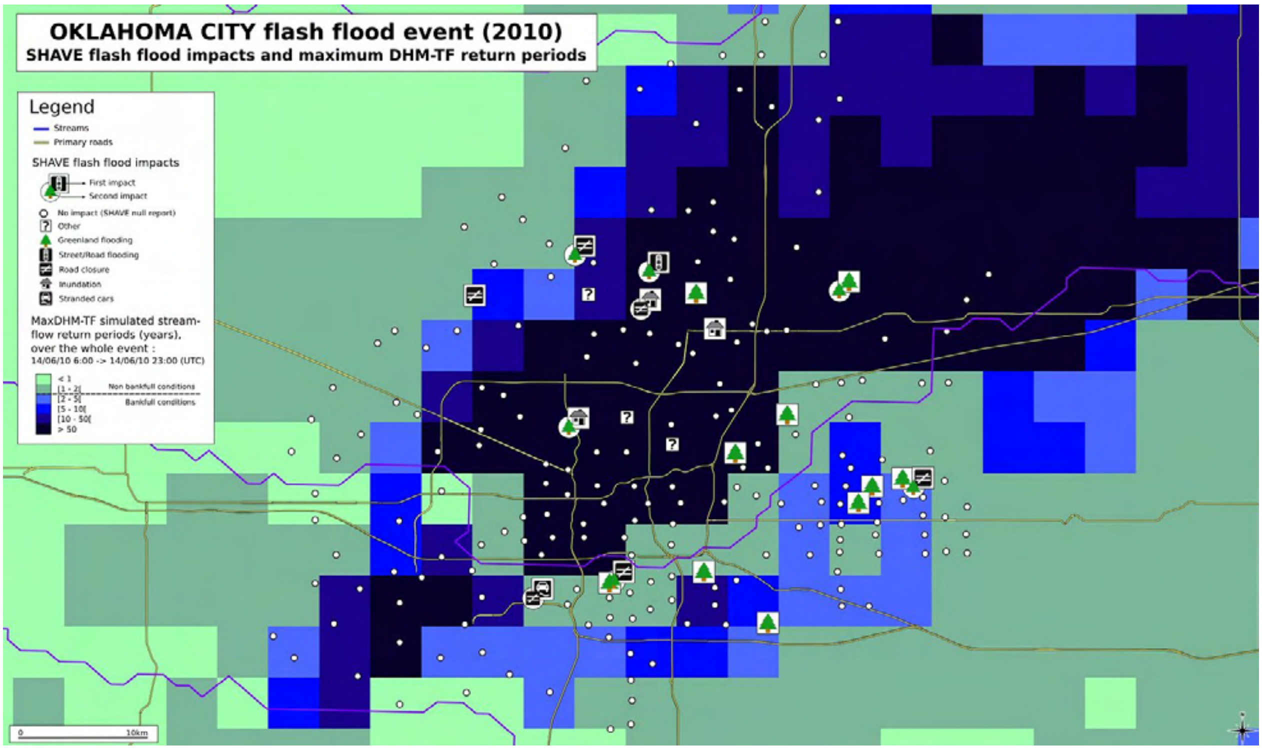

OKLAHOMA CITY flash flood event (2010)
SHAVE flash flood impacts and maximum DHM-TF return periods
Legend
Streams
Primary roads
SHAVE flash flood impacts
First impact
Second impact
No impact (SHAVE null report)
Other
Greenland flooding
Street/Road flooding
Road closure
Inundation
Stranded cars
MaxDHM-TF simulated stream-flow return periods (years), over the whole event :
14/06/10 6:00 -> 14/06/10 23:00 (UTC)
< 1
[1 - 2[
[2 - 5[
[5 - 10[
[10 - 50[
> 50
Non bankfull conditions
Bankfull conditions
0
10km

Figure 22: SHAVE impacts and max DHM-TF return periods for the Oklahoma City case.

The next analysis for this case consists of comparing tool values as function of impact classes. On average (see Fig. 23 and 24) both tools are detecting all kind of impacts (except Stranded Cars for 1-h QPE/GFFG, but this result must be taken with high caution as there is only one observation). But note that the GFFG tool shows values are just above bank-full conditions, whereas DHM-TF shows very high return periods, compared to bank-full conditions. For this particular smaller scaled, urban case, no clear link is found between impact type and tool values.

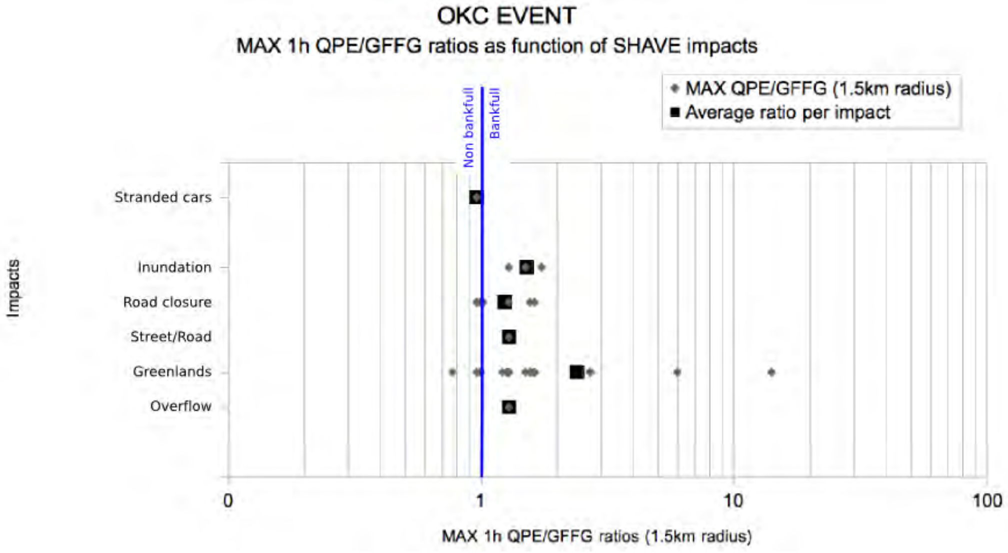


Figure 23: Max 1h QPE/GFFG ratios as function of impacts for the Oklahoma City case.

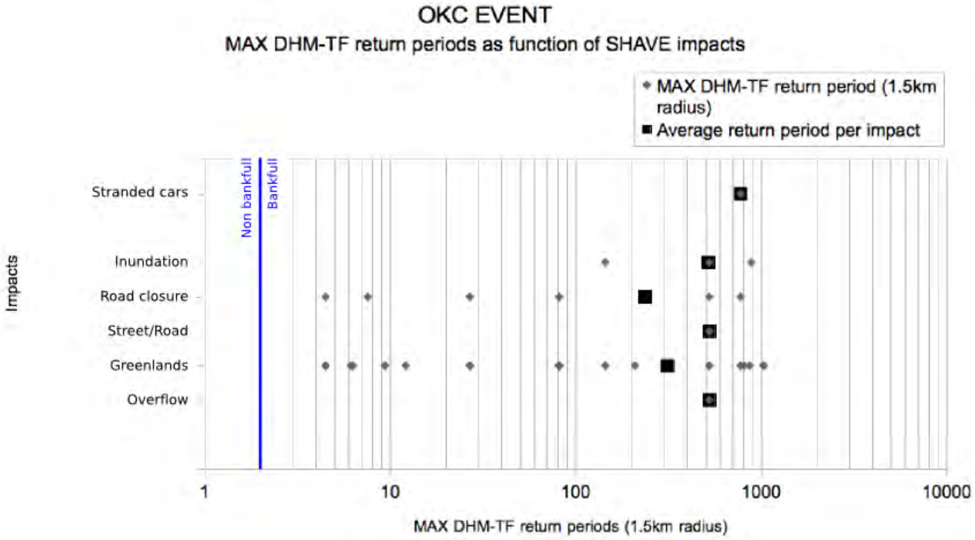


Figure 24: Max DHM-TF return periods as function of impacts for the Oklahoma City case.

5 Conclusion

This study provides an impact classification of flash flood reports datasets over the United States, in order to evaluate the ability of US flash flood forecasting tools to predict such categories of impacts.

After presenting the flash flood reports datasets (NWS and SHAVE), the method used for impact classification is described. That way, impact-enhanced datasets could be created. SHAVE impacts are then used in a spatio-contextual analysis, based on a cross-tabulation method using attributes already included in the SHAVE dataset (*perceived* attributes : Water Movement, Return Period and Water Depth) as well as GIS-sampled spatial attributes (Land Use, Population Density, Local Upslope, Drainage Area and Wetness index). This analysis showed consistent results (apart from topographic variables : Local Upslope, Drainage Area and Wetness index), indicating that on the one hand, impact classification was made correctly and on the other hand, that the SHAVE dataset (even if based on public polls) is a reliable tool for flash flood characterisation. Moreover, interesting results emerge from this analysis:

- Evacuations are not necessarily only observed in urban zones, but also in rural areas.
- Rescues, Fatality or Injuries mostly take place in low population density areas. Moreover, these impacts are not perceived by the interviewees as extreme, rare events (as it is the case for Evacuation) because associated return periods vary from one year to "never seen before". This result indicates how people may have difficulties to estimate an event frequency.

The second part of this study consisted in an evaluation of three US flash flood forecasting tools : FFG, GFFG and DHM-TF. After a brief presentation of the tools, two extreme cases of flash flood in Oklahoma (the Erin and Oklahoma City events) were chosen to evaluate the tools on a YES/NO-forecast basis (computing Probability Of Detection, False Alarm Ratio and Critical Success Index), but also as function of impacts.

For the Erin case (using NWS reports), all three tools are available, but the only statistic that can be computed is the POD, as there is no false alarm reports available in the NWS dataset. DHM-TF shows the maximum value of POD (1), followed by FFG (0.94) and GFFG (0.78). But on the maps, DHM-TF (and GFFG, to a lesser extent) show larger forecast areas, which may indicate more false alarms, even if FAR can not be computed. When looking at tools values as function of impacts, FFG and GFFG seem to show a relation (even if it must be taken with extreme care as some impacts are not sufficiently sampled) between average tool values and impacts, when ranked by severity, whereas DHM-TF does not show any relation between return period and impact severity. Moreover, for this larger scale case, on a map, impacts seem to be located mostly along primary roads and streams.

For the Oklahoma City case (using SHAVE reports), only two tools are available (GFFG and DHM-TF) but in this case, FAR and CSI can be computed. GFFG shows the best value of CSI (0.14), despite a lower POD value than DHM-TF, but thanks to a better FAR. These CSI values are very low, even if they stay in the same order of magnitude as the highest value found by Gourley et al. [2011a]. When looking at tools values as function of impacts, for this particular smaller scaled, urban case, no clear link is found between impact type and tool values. Moreover, on a map, impacts are not especially located along primary roads or streams.

In conclusion, only a weak link between FFG and GFFG ratios and impact types could be observed, for the Oklahoma City case. Indeed these tools not were not designed to take into account flash flood impacts, which is the combination of a hazard (in this study, quite well described by the tools) but also human and environmental vulnerability. This result demonstrates that these vulnerabilities must be assessed in more details. It can be done by a thinner analysis of topographic attributes (Drainage Area, Local Upslope, Wetness index), using nominal resolutions down to 30m (this can be derived from an Aster Digital Elevation Model for instance). Other methods could also be used to estimate the Drainage Area, like stream proximity or watershed area. Furthermore, it would be interesting to cross impacts with road network density.

Finally, of course, the analysis of these two particular flash flood cases should be completed with a study of the whole NWS and SHAVE dataset. This would provide much more samples and produce more robust statistics for tool evaluation.

References

Ashley, S., Ashley, W., 2008. Flood fatalities in the united states. J. Appl. Meteor. Climatol. 47, 806–818.

Beven, K., Kirkby, M., 1979. A physically-based variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 1, 43–69.

Demuth, J.L., Gruntfest, E., Morss, R.E., Drobot, S., Lazo, J.K., 2007. Was*is, building a community for integrating meteorology and social science. Bull. Am. Meteorol. Soc. , 1729–1737.

Gaume, E., Borga, M., 2008. Post-flood field investigations in upland catchments after major flash floods: proposal of a methodology and illustration. J. Flood Risk Manage. 1, 175–189.

Georgakakos, K.P., 1986. On the design of national, real time warning systems with capability for site-specific flash flood forecasts. Bull. Am. Meteorol. Soc. , 1233–1239.

Gourley, J.J., Erlingis, J., Smith, T., Ortega, K., Hong, Y., 2010. Remote collection and analysis of witness reports on flash floods. J. Hydrol. , 53–62.

Gourley, J.J., Erlingis, J.M., Hong, Y., Wells, E.B., 2011a. Evaluation of tools used for monitoring and forecasting flash floods in the united states. Wea. Forecasting (in press).

Gourley, J.J., Flaming, Z.L., Hong, Y., Howard, K.W., 2011b. Evaluation of past, present, and future tools for radar-based flash flood prediction. Hydro. Sci. J. (in review).

Hong, Y., Adhikari, P., Gourley, J.J., 2010. Flash flood. Encyclopedia of Natural Hazards, Springer, in press.

Reed, S., Schaake, J., Zhang, Z., 2007. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. J. Hydrol. , 402–420.

Ruin, I., Creutin, J.D., Anquetin, S., Lutoff, C., 2008. Human exposure to flash-floods, relation between flood parameters and human vulnerability during a storm of september 2002 in southern france. Journal of Hydrology , 199–213.

Schmidt, J.A., Anderson, A.J., Paul, J.H., 2007. Spatially-variable, physically-derived flash flood guidance., in: AMS 21st Conference on Hydrology, San Antonio, TX. p. 6B.2.

Versini, P.A., Gaume, E., Andrieu, H., 2010. Assessment of the susceptibility of roads to flooding based on geographical information - test in a flash flood prone area. Nat. Hazards and Earth Syst. Sci. 10, 793–803.

Appendix A : original flash flood reports datasets

	NWS fields	Comments
1.	ID number	
2.	Event Type	In our case, only flash floods.
3.	WFO	Weather Forecast Office
4.	Begin Time (UTC)	
5.	End Time (UTC)	
6.	Begin Time (Unix)	
7.	End Time (Unix)	
8.	Timezone	
9.	State	
10.	County	
11.	Region	
12.	Direct Injuries	
13.	Indirect Injuries	
14.	Direct Fatalities	
15.	Indirect Fatalities	
16.	Property Damage	Often estimated.
17.	Crop Damage	Often estimated.
18.	Flood Cause	For the great majority of the cases : Heavy Rain.
19.	Event Source	e.g.: Emergency manager, Trained Spotter, Public, News
20.	Event Narrative	of the flood event.
21.	Episode Narrative	of the meteorological event.
22.	Location	City, range and azimuth of the impacted area.
23.	Location1 (Lat)	Latitude in decimal degrees of the point or polygon first vertex.
24.	Location1 (Lon)	Longitude in decimal degrees of the point or polygon first vertex.

Table 11: Structure of the NWS dataset.

	SHAVE fields	Comments
1.	Event type	Denotes if the report is about flood, wind, or hail. In our case: flood only.
2.	Mag units	Units for flood depth and extent
3.	Id tag	Denotes the call time and caller (for internal use)
4.	Revision time	Time of file edit (for internal use)
5.	Revision number	Number of times the file was edited (for internal use)
6.	Start time	Event start time in Unix time
7.	Start year/month/day/hour	Event start (hour in UTC), separated in 4 fields.
11.	End time	Event end time in Unix time
12.	End year/month/day/hour	Event end (hour in UTC), separated in 4 fields.
16.	Start lat	Latitude of report
17.	Start lon	Longitude of report
18.	City	Relationship of report location to the nearest city
19.	County/State/CWA	County/State/County Warning Area where the report is located (3fields).
22.	Flood nearby location	The location of the report if not at lat/lon of residence
23.	Flood type	- poor drainage/street flooding - road/bridge closure - inundation of structure - pond/creek/stream overflow - farmland/pasture flooding - yard flooding - cropland flooding - other (see comments)
24.	Flood nature other	Comments about flooding marked "other" in the above field
25.	Flood ongoing	Denotes if the flooding was still occurring at the time of the call
26.	Flood move	Denotes if the floodwater was moving or standing.
27.	Flood water depth m	Depth of the floodwater (in meters)
28.	Flood lateral extent m	Lateral extent of pond/creek/stream overflow (in meters)
29.	Flood evac	Denotes if evacuations occurred due to flooding
30.	Flood evac location	Location of the evacuation(s)
31.	Flood rescue	Denotes if rescues occurred because of flooding
32.	Comments	Additional comments about the call (not pertaining to meteorological events)
33.	Metr comments	Additional comments about the call (pertaining to meteorological events)
34.	Contact phone	Resident's phone number [Removed prior to public consumption of data]
35.	Contact results	Marks the call as "questionable" if the resident could not provide an exact time or location of the event or if the data given were suspect
36.	Flood frequency	0 = no response 1 = every time it rains 2 = only during heavy rain 3 = once per year 4 = once every 5 years 5 = once every 10 years 6 = never had seen it before
37.	Report type	2 = severe / 1 = non-severe / 0 = null The following criteria were used for determining a severe flood: 0.5 ft (0.15 m) of moving water, 3.0 ft (0.91 m) of standing water, road/bridge closures, washed out roads/bridges, rescues, evacuations, water in an above-ground structure, or major creeks/rivers out of banks.

Table 12: Structure of the SHAVE dataset.

Appendix B : modified *impact-foccused* flash flood reports datasets

	Modified NWS fields	Comments
1.	ID number	
2.	Begin Time (UTC)	
3.	End Time (UTC)	
4.	Begin Time (Unix)	
5.	End Time (Unix)	
6.	Timezone	
7.	Property Damage	Often estimated.
8.	Crop Damage	Often estimated.
9.	Impact1	Most severe recorded impact
10.	Impact2	Second most severe recorded impact
11.	Impact3	Third most severe recorded impact
12.	Event Source	e.g.: Emergency manager, Trained Spotter, Public, News
13.	Event Narrative	of the flood event.
14.	Location	City, range and azimuth of the impacted area.
15.	Location1 (Lat)	Latitude in decimal degrees of the point or polygon first vertex.
16.	Location1 (Lon)	Longitude in decimal degrees of the point or polygon first vertex.

Table 13: Structure of the modified impact-foccused NWS dataset.

	Modified SHAVE fields	Comments
1.	Id tag	Report identification number
2.	Start UTC	Event start date and time in UTC
3.	Start UNIX	Event start time in UNIX
4.	End UTC	Event end date and time in UTC
5.	End UNIX	Event end time in UNIX
6.	Start lat	Latitude of report
7.	Start lon	Longitude of report
8.	Impact1	Most severe recorded impact
9.	Impact2	Second most severe recorded impact
10.	Impact3	Third most severe recorded impact
11.	Metr comments	Additional comments about the call (pertaining to meteorological events)
12.	Report type	2 = severe / 1 = non-severe / 0 = null The following criteria were used for determining a severe flood: 0.5 ft (0.15 m) of moving water, 3.0 ft (0.91 m) of standing water, road/bridge closures, washed out roads/bridges, rescues, evacuations, water in an above-ground structure, or major creeks/rivers out of banks.
13.	Flood ongoing	Denotes if the flooding was still occurring at the time of the call
14.	Flood move	Denotes if the floodwater was moving or standing.
15.	Flood water depth m	Depth of the floodwater (in meters)
16.	Flood lateral extent m	Lateral extent of pond/creek/stream overflow (in meters)
17.	Flood evac location	Location of the evacuation(s)
18.	Contact results	Marks the call as "questionable" if the resident could not provide an exact time or location of the event or if the data given were suspect
19.	Flood frequency	0 = no response 1 = every time it rains 2 = only during heavy rain 3 = once per year 4 = once every 5 years 5 = once every 10 years 6 = never had seen it before
20.	Land Use C	Land Use (code) 0 = Water 1 = Evergreen Needleleaf Forest 2 = Evergreen Broadleaf Forest 3 = Deciduous Needleleaf Forest 4 = Deciduous Broadleaf Forest 5 = Mixed Forest 6 = Woodland 7 = Wooded Grassland 8 = Closed Shrubland 9 = Open Shrubland 10 = Grassland 11 = Cropland 12 = Bare Ground 13 = Urban and Built-up
21.	Land Use	Land Use (text)
22.	Population	Population density (inhabitant/km2)
23.	Slope	Local maximum upslope (°)
24.	Flow Accu	Flow accumulation, i.e.: drainage area (km2)
25.	CTI	Compound Topographic Index, also called ’Wetness Index’

Table 14: Structure of the modified impact-foccused SHAVE dataset.