Journal of Hydrology 477 (2013) 1–16

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

# Supplementing flash flood reports with impact classifications

Martin Calianno [a], Isabelle Ruin [a,*], Jonathan J. Gourley [b]

[a] CNRS/Laboratoire d'étude des Transferts en Hydrologie et Environnement (LTHE), Grenoble cedex, France
[b] NOAA/National Severe Storms Laboratory (NSSL), Norman, OK, United States

ARTICLE INFO

*Article history:*
Received 25 June 2012
Received in revised form 7 September 2012
Accepted 15 September 2012
Available online 26 September 2012
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Efrat Morin, Associate Editor

*Keywords:*
Flash flood
Database
Impact classification
Forecast verification

SUMMARY

In recent years, there has been an increase in flash flood impacts, even as our ability to forecast events and warn areas at risk increases. This increase results from a combination of extreme events and the exposure of vulnerable populations. The issues of exposure and vulnerability to flash floods are not trivial because environmental circumstances in such events are specific and complex enough to challenge the general understanding of natural risks. Therefore, it seems essential to consider physical processes of flash floods concurrently with the impacts they trigger. This paper takes a first step in addressing this need by creating and testing the coherence of an impact-focused database based on two pre-existing public and expert-based survey datasets: the Severe Hazards Analysis and Verification Experiment (SHAVE) and the US National Weather Service (NWS) *Storm Data*. The SHAVE initiative proposes a new method for collecting near-real-time high-resolution observations on both environmental circumstances and their disastrous consequences (material and human losses) to evaluate radar-based forecasting tools. Forecast verification tools and methods are needed to pursue improving the spatial and temporal accuracy of forecasts. Nevertheless by enhancing SHAVE and NWS datasets with socially and spatially relevant information, we aim at improving future forecast ability to predict the amount and types of impacts.

This paper describes the procedures developed to classify and rank the impacts from the least to the most severe, then to verify the coherence and relevance of the impact-focused SHAVE dataset via cross-tabulation analysis of reported variables and GIS-sampled spatial characteristics. By crossing impact categories with socio-spatial characteristics, this analysis showed first benchmarks for the use of exposure layers in future flash flood impact forecasting models. The enhanced impact-focused datasets were used to test the capabilities of flash flood forecasting tools in predicting different categories of impacts for two extreme cases of flash flooding in Oklahoma, USA. Results showed a general tendency for the more severe impacts to be associated to higher mean exceedances over tool values. This means that, at least for these particular case studies, the tools were able to make a distinction between less severe and more severe impacts. Finally, a critical analysis of the NWS and SHAVE data collection methodologies was completed and challenges for future work were identified.



## 1. Introduction

Flash floods are rapid surface water responses to rainfall from intense thunderstorms or a sudden release of water from a dam or ice jam. Inundation occurs over normally dry land from within minutes to a few hours of the causative rainfall and can have devastating impacts on lives and infrastructure (Hong et al., 2012). In recent years, there has been an increase in flash flood impacts, even as our ability to forecast events and warn areas at risk increases (Montz and Gruntfest, 2002). For instance, in the US, flash flooding is considered one of the deadliest among weather-related hazards (Ashley and Ashley, 2008). This increase results from a combination of extreme events and the exposure of vulnerable populations. In addition to the need for understanding physical processes of flash floods, it has become more important to analyze the human impact of such disasters (World Bank, 2010). Efforts have been made on the side of hydrometeorological sciences to collect data on rainfall–runoff processes and spatio-temporal patterns of rainfall and runoff associated with flash floods. Current flood datasets include measurements from in situ stream gauges and acoustic Doppler profilers (Simpson and Oltmann, 1993), remote sensing of water surface extents (Brakenridge et al., 2005), post-event field investigations (Gaume and Borga, 2008) or rainfall–runoff modeling.

Social impacts from flash floods are also documented, as some studies and datasets differentiate rapid-onset flash flooding from slow-onset riverine flooding and other types of flooding. But those

* Corresponding author. Address: CNRS/UJF-Grenoble 1/G-INP/IRD, Laboratoire d'étude des Transferts en Hydrologie et Environnement (LTHE), UMR 5564, Bâtiment OSUG-B, Domaine Universitaire, BP 53, 38041 Grenoble cedex 09, France. Tel.: +33 4 76 82 50 56; fax: +33 4 76 82 50 14.
*E-mail address:* isabelle.ruin@ujf-grenoble.fr (I. Ruin).

0022-1694/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2012.09.036

studies, which may be driven by various research questions, differ in content, spatial and temporal extent as well as comprehensiveness. Case studies often adopt a qualitative approach in an attempt to understand the root causes of a specific catastrophe. For instance, several case studies focus on human vulnerability through the study of loss of life circumstances (Duclos et al., 1991; Gruntfest, 1977; Ruin et al., 2008; Staes et al., 1994; Vinet et al., 2011). In a similar attempt to characterize human vulnerability factors, several researchers also investigated national and international databases compiled from newspapers, historical accounts, government and scientific reports (French et al., 1983; Coates, 1999; Antoine et al., 2001; Few et al., 2004; Jonkman, 2005; Jonkman and Kelman, 2005; Sharif et al., 2010; Zahran et al., 2008). It is noteworthy that few of the existing large-scale databases readily accommodate both quantitative and qualitative analysis. The level of details associated with the event and impacts description is often conditioned by the territorial scale (commune, county, region, state, and nation) at which the data are collected (and sometimes subsequently lost). Similarly, reported events are often the ones classified as "catastrophic" based on criteria such as the number of fatalities, affected people, emergency declaration and call for international assistance. The criteria and classification used are not always consistent across databases. Therefore it is often difficult to use these various datasets in a complementary and seamless fashion.

The impacts of flash floods are diverse. Flooding may take the form of runoff, street/urban flooding, streams flowing out of their banks, or even mud/debris flow. Flash flood impacts are a bit different than other natural hazards (like hail or tornadoes) in that they are more strongly controlled by surface properties, infrastructure, and spatial and temporal distribution of societal exposure. All of these specific aspects require interdisciplinary efforts, integrating natural and human sciences to improve the understanding and ultimately, the prediction of flash floods. Therefore this type of event requires different tools, models and data collection strategies than those used for forecasting on larger, well-instrumented basins.

Few datasets include detailed information about flash flood events and/or their related impacts. In Europe a recent initiative supported by the HYDRATE (HYdrometeorological Data Resources And Technologies for Effective flash flood forecasting) project compiled detailed hydrometeorological data on flash flood events that have occurred since 1994 (Gaume et al., 2009). In the USA, the Storm Events database, a product of NOAA's National Weather Service consists of both meteorological and impact data collected by local forecast offices through spotter reports. The NWS dataset contains detailed narratives about events and often supplies damage estimates, but the times and spatial extents of the events can be imprecise, as the latter are designated with forecaster-drawn bounding polygons.

To address the lack of high-resolution datasets over vast regions for both flash flood forecasting verification and research on flash flood impacts, Gourley et al. (2010) proposed a near-real-time public-based survey within the Severe Hazards Analysis and Verification Experiment (SHAVE). This was a student-led and student-run experiment conducted at the National Severe Storms Laboratory (NSSL) in Norman, Oklahoma during the summer months from 2008 to 2010. The magnitude and instances of respondent-reported flash flooding has been used to evaluate new, gridded tools used operationally in the NWS for flash flood monitoring and prediction (Gourley et al., 2012a, in preparation-b). While the original intent of SHAVE was to collect high-resolution observations to evaluate these radar-based forecasting tools, it became clear that the details collected during the experiment were well suited for studying the specific impacts and characteristics of flash floods in a more general, comprehensive way. To date, the specific impacts of flash flooding collected during SHAVE have yet to be studied or utilized in any way.

This paper takes a first step in addressing the need to consider both flash flood physical characteristics and impacts by creating and testing the coherence of an impact-focused database based on two pre-existing public and expert-based survey datasets: the Severe Hazards Analysis and Verification Experiment (SHAVE) and the National Weather Service (NWS) *Storm Data*. By supplementing SHAVE and NWS datasets with socially and spatially relevant information, we aim at improving future forecast ability to predict the amount and types of impacts. The paper describes the dataset-enhancement process allowing to classify and rank the impacts from the least to the most severe. We also verify the coherence and relevance of the impact-focused SHAVE dataset via cross-tabulation analysis of reported variables and GIS-sampled spatial characteristics. Finally, we incorporate the enhanced impact-focused datasets, developed in this study, to test the capabilities of flash flood forecasting tools in predicting different categories of impacts for two extreme cases of flash flooding in Oklahoma.

The paper is organized as follows. Section 2 introduces the flash flood reports datasets and their impact-focused enhancement. Section 3 presents the cross-tabulation analysis of SHAVE impacts. In Section 4, two extreme cases of flash flooding in Oklahoma are used to test the capabilities of forecasting tools in predicting impacts. Section 5 reviews SHAVE and NWS data collection methodology and proposes ways for improvements. Finally, a summary of results and concluding remarks are given in Section 6.

## 2. Flash flood datasets

The SHAVE and NWS datasets are employed in this study to characterize flash flooding impacts. Both datasets are built from personal reports. NWS forecasters collect *Storm Data* reports from trained spotters, emergency management personnel, and the public. The SHAVE data are obtained from the general public through responses to a questionnaire. Both datasets are obtained shortly after (generally within hours up to 1 day) of the causative event. While personal accounts inherently introduce subjectivity, uncertainty, and occasional embellishments, they are presently the best information available for determining flash flood impacts. These reports have been combined into a consistent database available for display in Google Earth™ and for analysis using geographical information system tools. The database, described in more detail in Gourley et al. (in preparation-c), is freely available to the public.

### 2.1. NWS reports

The NWS *Storm Data* dataset contains flash flood reports across the entire USA. Local NWS forecast offices collect reports throughout the year and store them either as latitude/longitude points (from 2006 to 2007) or polygons (from 2007 to present). In total, 15,999 reports of flash floods have been gathered from 2006 to 2010, over the whole US territory (see Fig. 1). These reports are meant to be comprehensive for all flash flooding events occurring within each of the 122 local forecast offices' areas of responsibility. The principal aim of this dataset is to verify NWS flash flood warnings that have been issued by the local forecast offices. The sampling method is primarily based on calling of trained spotters, local law enforcement, and emergency management officials within the warned areas. Then forecasters define polygons (formerly, points) with as many as eight vertices that delineate the regions that are suspected of being affected by flash floods. Information about event timing, fatalities, and injuries or damages is also gathered. For larger events, damages are estimated (in US dollars) and

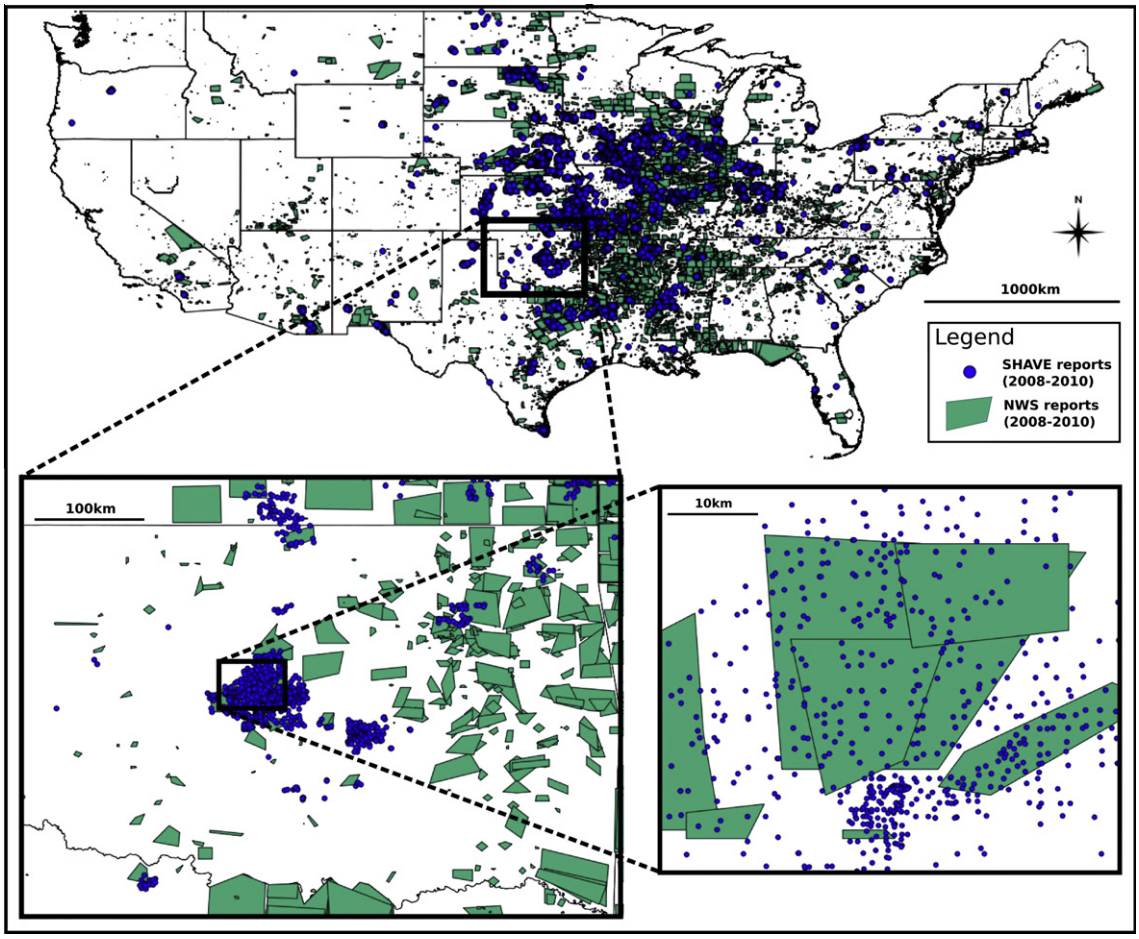


**Fig. 1.** Spatial coverage of SHAVE flash flood observations and NWS flash flood event polygons from 2008 to 2010.

provided in the report. Lastly, forecasters often include lengthy narratives about the flood event and the meteorological context causing the event.

The main advantage of this dataset is that reports are collected throughout the year and across the entire US. Moreover, the personnel providing the reports are very familiar with their area of responsibility so they can provide immediate quality control of reports. And the reporting is often part of their job requirement. Thus, a majority of the reports are quite reliable, although some reports include information obtained from newspaper articles and the general public. Nevertheless, the NWS dataset does not include reports of no flooding in warned regions (i.e. false alarms), and often does not report flood events that occurred without warning (i.e. missed events) or floods that happened in more sparsely populated areas where few people could have witnessed them (Gourley et al., 2010). Because reports mostly come from local officials they are somehow biased toward focusing on urban features, especially road disruptions that impact city functions. Prior to 2007, the point-based reports often represented flash flooding somewhere within a much larger political county. Nowadays, the reports are defined by bounding polygons but still have large uncertainties in their spatial accuracy. Also, information about event timing has poor accuracy, as the meteorological event start/end times are often taken as flood event timing. This lack of accuracy encumbers an analysis of the flash flood onset and evolution.

### 2.2. SHAVE reports

The SHAVE database was set up at the National Severe Storms Laboratory in Norman, OK (Gourley et al., 2010) and includes flash flood reports for the entire US from 2008 to 2010. Undergraduate students collected reports using landline telephone surveys to poll the general public, based on their residential address, during the summer months (June through August). This dataset is point-based and was originally designed to complement hail observational data. Subsequently, wind damage, tornadoes and flash flood reports were added to the experiment, in order to create higher-resolution datasets for model verification. The flash flood sampling method is storm-targeted for flash floods rather than larger scale, fluvial floods (with basin areas >300 $km^2$). Reports were classified in Gourley et al. (2010) as null (i.e., no impact), minor, or severe events, using the following information collected through the questionnaire: flood type, water movement, water depth and extent, and occurrence of evacuation/rescue and flood frequency. Localization and timing are included as well as textual comments about the flood event.

Compared to the NWS dataset, the main attributes of the SHAVE data are its higher spatial resolution (denser point sampling) and an estimation of the event's spatial extension and magnitude (reports range from no impact to severe for each event) based on responder's answers. Moreover, additional information about the characteristics of the event (i.e., flood type, depth, extent, frequency, and textual comments) is included, as well as, reports of no flooding in warned regions (i.e., false alarms) and flooding in unwarned areas (i.e., missed events). However, we acknowledge that this mode of data acquisition based on the collation of public perceptions of on-going environmental conditions through surveys may introduce uncertainty and perhaps bias, a topic we will revisit in the last section of the paper.

### 2.3. Development of impact-focused datasets

Contextual comments and information already included in SHAVE and NWS flash flood reports were used to develop a comprehensive flash flood impact typology. It includes information grouped using: (1) multiple fields originally included in the datasets, for instance 'flood type', 'evacuation', 'rescue' (from SHAVE) and 'fatalities', 'injuries' (from NWS), and (2) textual comments (i.e., flood event narrative and meteorological event description). Ten different impact categories are created from SHAVE (Fig. 2). These impact classes are subjectively ranked from the least to the most severe, based on *a priori* judgment: '*no impact*' (=SHAVE null report), '*other*' (unclassified or unknown impact), '*overflow*' (streams out of their banks), '*green lands*' (flooded cropland, pasture, yard or grassland), '*street/road flooding*', '*road closure*' (roads closed by the authorities, or impassible), '*inundation*' (floodwaters in an above-ground residence), '*evacuation*', '*stranded cars*' (moved by floodwaters, stalled in ditches) and finally, '*rescues/fatalities/injuries*'.

For NWS, no 'null report' is included, but the first categories are similar to SHAVE ('*overflow*', '*green lands*', '*street/road flooding*', '*road closure*', '*inundation*' and '*stranded cars*'). Because not enough information was available about evacuations to create a single category, it was grouped into a wider, most severe class: '*evacuation/rescue/fatalities/injuries*'. Since more than one impact may 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. With this system, multi-impact reports can be handled.

## 3. Are SHAVE data reliable even when based on a public survey?

Based on the aforementioned impact typology, we now evaluate the coherence and relevance of the SHAVE impact-focused dataset. To accomplish this task, a cross tabulation analysis has been set up, where impact classes are crossed with interviewees' perceptions of the flood event characteristics and with spatial raster maps using GIS.

### 3.1. Selection and sampling of attributes to be crossed with impact classes

Information about interviewees' perceptions of flash flood characteristics is readily available for each report in the SHAVE dataset, herein referred to as "perceived attributes", and includes water movement, depth and flood frequency. In addition, attributes were added to the dataset by sampling raster data. These SHAVE-independent variables, called "spatial attributes" were retrieved by point sampling, which assigns the pixel value located right under each SHAVE report. In the case of multiple flash flood impacts, the second and third most severe impacts were copied as additional report points in order to maximize the data sample. Spatial attributes added to the database are land use, population density, and drainage area (see Table 1 for a presentation of the attributes). Land use is extracted from the USGS National Land Cover Database 2006 raster map at a spatial resolution of 30 m (Fry et al., 2011) and population density distribution from the US 2000 census grid at a 1 km$^2$ grid cell resolution (Owen and Gallo, 2000). The drainage area raster is computed from the USGS global elevation model GMTED2010, at a resolution of 7.5 arc sec (Danielson and Gesch, 2011) using the *r.watershed* algorithm and the Single Flow Direction (SFD) method (Ehlschlaeger, 1989). In this drainage area raster, the absolute value of each cell is the number of upstream cells that drain to it. This value was then converted into square kilometers by knowing the area of each pixel. For this attribute, a sampling cluster was used to select the maximum value of flow accumulation within a 300 m radius around each SHAVE report. The aim is to sample the nearest '*stream segment*' within 300 m around the interviewee's location. This arbitrary 300-m buffer was chosen to represent the mean distance to be sighted by the average person from their home.


**Fig. 2.** SHAVE impact categories ranked by severity and their corresponding symbology.

### 3.2. Attributes distribution and categorization

Because the majority of attributes (i.e., '*water movement*', '*return period*' and '*land use*') describing flash flood impacts consists of categorized variables, a cross tabulation approach was chosen to analyze the relationship between these variables and impact classes. The continuous attributes (i.e., '*water depth*', '*population density*', '*local upslope*' and '*drainage area*') were categorized as follows (Table 2):

- The *water depth* perceived attribute was split into three categories, according to the interviewees' perception: ⩽10 cm (corresponding to ankle-deep water and shallower), 10–30 cm (between ankle-deep and knee-deep water) and >30 cm (above knee-deep water).
- The original SHAVE '*flood frequency*' field contained six categories. It has been simplified into three new categories of perceived return periods: 'more often than every year', 'every year to every ten years' and 'never seen before'.
- The twenty NLCD2006 land cover classes originally contained in the raster layer were grouped to make five global categories. These are '*natural vegetation*' (including forest, shrubland and herbaceous land covers), '*pasture/crops*' (planted and cultivated lands) and three levels of urbanized covers: '*developed-open space*' (containing <20% of impervious surface), '*developed-low intensity*' (20–49% impervious surface) and '*developed-high intensity*' (>49% impervious surface). This classification allows a gradation from natural to more anthropogenic areas.
- Population density were divided into four classes, in order to account for sparsely-populated (⩽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$^2$).
- The '*drainage area*' distribution was split into five classes. The first classes were created using quantiles: ⩽0.25 km$^2$, 0.25–0.75 km$^2$ and 0.75–2 km$^2$. The last two classes were chosen to make the distinction between drainage areas below and above 20 km$^2$: 2–20 km$^2$ and >20 km$^2$. This limit was chosen in accordance with Ruin et al. (2008), which studied the hydrometeorological circumstances of fatal accidents during the 2002 flash flood event in the Gard region of France. The study found that fatal accidents in catchments <20 km$^2$ occurred

**Table 1**
Presentation and statistical description of the perceived and spatial attributes.

| Attributes presentation and statistics | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Perceived attributes | | | Spatial attributes | | |
| | Water movement | Water depth | Flood frequency | Population density | Land use | Drainage area |
| Unit | Three categories | (meters) | Six categories | (inhabitant/km$^2$) | Five categories | (km$^2$) |
| Resolution | Point-based | Point-based | Point-based | 1 km raster | 30 m raster | 200 m raster |
| Source | SHAVE | SHAVE | SHAVE | US 2000 census | USGS NLCD2006 | UGS GMTED2010 |
| Sample size | $n$ = 1907 | $n$ = 2328 | $n$ = 2047 | $n$ = 2548 | $n$ = 2682 | $n$ = 2697 |
| Min | | 0.00 | | 0.0 | | 0.036 |
| Max | | 6.10 | | 8138.6 | | 65586.08 |
| Average | | 0.40 | | 198.2 | | 142.10 |
| Std dev. | | 0.56 | | 516.0 | | 2123.22 |
| Skewness | | 3.8 | | 5.5 | | 21.21 |
| Kurtosis | | 24.1 | | 48.4 | | 502.91 |

**Table 2**
Categorization of impacts and attributes.

| Attributes categorization | | | | | | |
|---|---|---|---|---|---|---|
| Impacts | Perceived attributes | | | Spatial attributes | | |
| SHAVE impacts | Water movement | Water depth (meters) | Flood frequency | Pop. density (inhabitant/km$^2$) | Land use | Drainage area (km$^2$) |
| Overflow | Moving | ⩽0.1 | ⩽1 year | ⩽4 | Natural vegetation | ⩽0.25 |
| $n$ = 471 | $n$ = 1131 | $n$ = 815 | $n$ = 1440 | $n$ = 1017 | $n$ = 359 | $n$ = 669 |
| (18.5%) | (44.4%) | (32.0%) | (56.5%) | (39.9%) | (13.4%) | (24.8%) |
| Greenlands | Standing | [0.1–0.3] | [10–30 years] | [4–70] | Pasture/crops | [0.25–0.75] |
| $n$ = 1019 | $n$ = 776 | $n$ = 608 | $n$ = 295 | $n$ = 866 | $n$ = 642 | $n$ = 672 |
| (40.0%) | (30.5%) | (23.9%) | (11.6%) | (34.0%) | (24.0%) | (24.9%) |
| Road Flooding | -unknown- | >0.3 | Never seen | [70–500] | Developed – open space | [0.75–2] |
| $n$ = 237 | $n$ = 641 | $n$ = 905 | $n$ = 312 | $n$ = 339 | $n$ = 897 | $n$ = 455 |
| (9.3%) | (25.1%) | (35.5%) | (12.2) | (13.3%) | (33.5%) | (16.9%) |
| Road closure | | -unknown- | -unknown- | >500 | Developed – low intensity | [2–20] |
| $n$ = 291 | | $n$ = 220 | $n$ = 501 | $n$ = 326 | $n$ = 583 | $n$ = 577 |
| (11.4%) | | (8.6%) | (19.7) | (12.8%) | (21.7%) | (21.4%) |
| Inundation | | | | | Developed – high intensity | >20 |
| $n$ = 388 | | | | | $n$ = 201 | $n$ = 324 |
| (15.2%) | | | | | (7.5%) | (12.0%) |
| Evacuation | | | | | | |
| $n$ = 71 | | | | | | |
| (2.8%) | | | | | | |
| Stranded cars | | | | | | |
| $n$ = 40 | | | | | | |
| (1.6%) | | | | | | |
| Rescue | | | | | | |
| $n$ = 31 | | | | | | |
| (1.2%) | | | | | | |

mainly outside in the open to middle aged males (43 years old, on average); whereas in larger catchments (>1000 km$^2$), fatalities occurred at home and concerned older people (average age of 76).

### 3.3. Results and discussion of the cross tabulation analysis

To measure the degree of independence and relationship strength between each impact type and spatial or perceived attributes, we performed the Pearson's Chi-squared and Cramer's $V$ tests. The Pearson's Chi-squared is a test for independence (i.e., independence between tested variables is the null hypothesis, $H_0$). If $H_0$ is rejected (i.e., when the Chi-squared asymptotic significance value is below the significance level taking alpha = 0.05), there is a statistically significant relationship between the variables. The Cramer's $V$ test evaluates the strength of a relationship between variables. High Cramer's $V$ values indicate strong relationships, with a maximum of 1 and a minimum of 0. These statistics are presented for each cross-tabulation (Table 3). The Chi-squared values indicate highly significant relationships between impacts and attributes classes, whereas the Cramer's $V$ values are relatively low, indicating moderate to weak relationships, especially for '*drainage area*' (Cramer's $V < 0.1$). The first attribute to be crossed with flash flood impacts is '*water depth*' (Fig. 3), a perceived characteristic collected through SHAVE. This flood description, combined with '*water movement*', will provide a first evaluation for the severity ranking of the SHAVE impact categories.

Significant positive deviations (above 5%) indicate that the '*overflow*', '*road closure*', '*evacuation*', '*stranded cars*' and '*rescue/fatalities*' impacts are mostly related to high waters (the >30 cm bin). Additionally, cross-tabulation results from the '*water movement*' perceived attribute show that all of the aforementioned categories are also associated with '*moving water*'. This combination of running, high waters represents a severe hazard and is thus consistent with its association to the three highest impacts categories ('*evacuation*', '*stranded cars*' and '*rescue/fatalities*'). It also makes sense in the context of '*overflow*' (rivers out of their banks) and '*road closure*' (which is often associated with overflows on nearby roads or low-water crossings). On the other hand, significant negative deviations indicate that '*green lands*' and '*inundation*' are not related to high floodwaters. Moreover, they are associated to '*standing water*'. Finally, the '*street/road flooding*' impact deviates most in the intermediate water depth bin (10–30 cm) and is strongly unassociated with shallow waters (<10 cm). This impact is also linked to '*moving water*', and thus, may indicate a dangerous runoff situation.

**Table 3**
Statistical tests (Cramer's *V* and Pearson's Chi2 2-sided asymptotic significance [*p*-value]) for each impact type versus attribute cross-tabulation.

| | SHAVE impacts, VS: | Chi-squared two-sided asymptotic value (alpha = 0.05) If <0.01: highly significant | Cramer's *V* |
|---|---|---|---|
| Perceived attributes | Water movement | 0.000 | 0.22 |
| | Water depth | 0.000 | 0.18 |
| | Flood frequency | 0.000 | 0.15 |
| Spatial attributes | Land use | 0.000 | 0.14 |
| | Population density | 0.000 | 0.19 |
| | Drainage area | 0.000 | 0.08 |

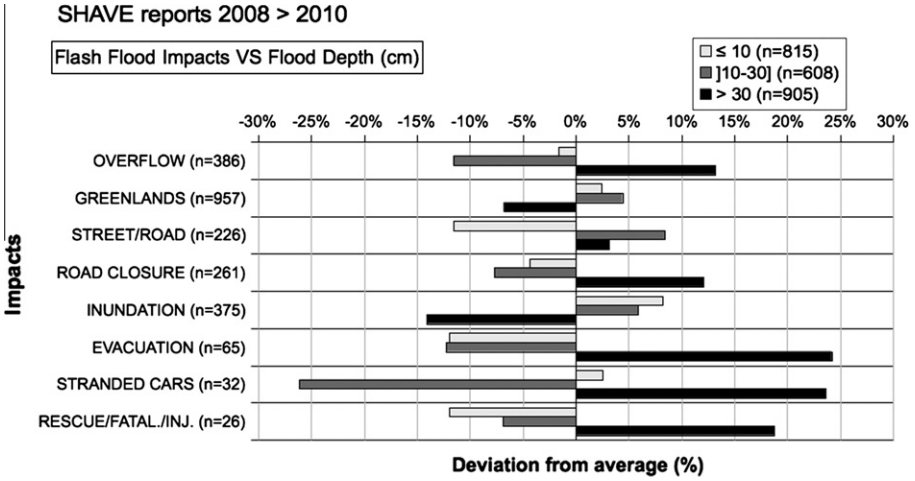


**Fig. 3.** Results of the crossing between SHAVE impacts and flood water depth: bar chart representing deviation from average (%).

'*Flood frequency*' estimated by SHAVE interviewees is the next perceived attribute to be crossed with SHAVE flash flood impacts (Fig. 4). Recall the respondents were asked during SHAVE to provide an estimate on the frequency (in years) at which the reported event occurs. '*Overflow*' impacts are most significantly associated with 1–10-year return periods (>5% deviation), but not to short return periods ($\leqslant$1 year), whereas, '*green land*' and '*street*/*road*' are mostly linked to such frequent events. Note that '*road closure*' and '*inundation*' show no significant deviation from the average (<5%), indicating that interviewees equally associate these impacts to frequent and rare events. Finally, '*evacuation*' and '*rescue*' are mostly associated with rare events, whereas, '*stranded cars*' shows positive deviations towards rare event, but also for the most frequent events, which is quite contradictory. It may be due to people's perceptions, knowledge, or age. Results must also be taken cautiously because of the small sample size for severe impacts. Nevertheless, they show that people are moderately able to evaluate flood frequency, indicating a general tendency showing that more severe the impact, the rarer the event.

The next cross tabulation analysis concerns the independent, GIS-sampled parameters. SHAVE impacts crossed with population density classes (Fig. 5) 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. '*Overflow*' and '*green land*' classes show clear positive deviations towards the lowest populations densities

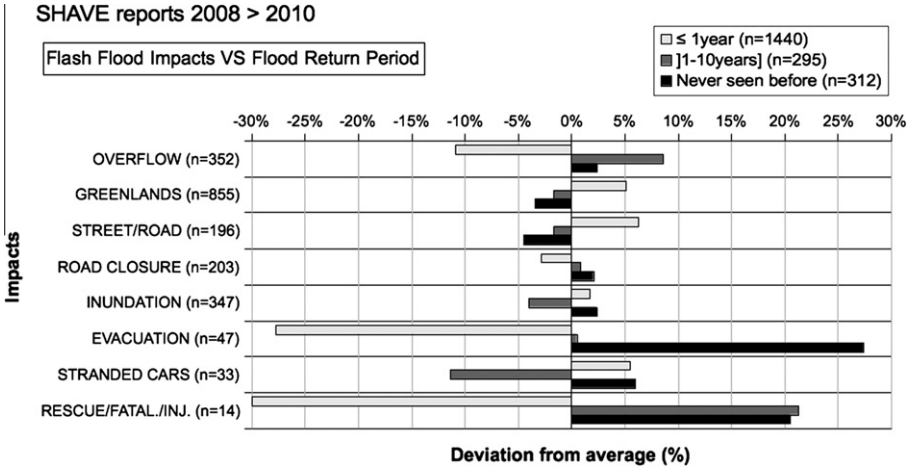


**Fig. 4.** Results of the crossing between SHAVE impacts and flood return period: bar chart representing deviation from average (%).

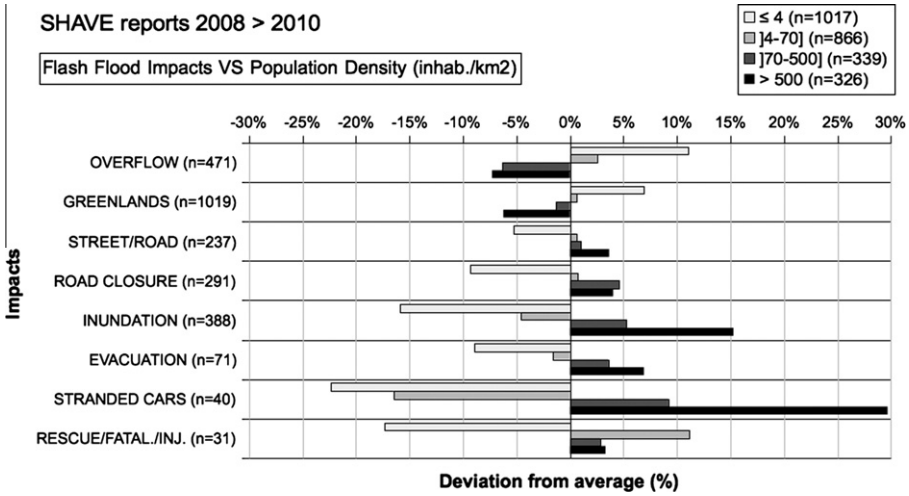


**Fig. 5.** Results of the crossing between SHAVE impacts and population density: bar chart representing deviation from average (%).

($\leqslant$4 inhab./km$^2$). This association with sparsely populated areas agrees with the 'non-urban' aspect of these impacts and will be confirmed by the following crossing with land use categories. On the other hand, '*street*/*road*' and '*road closure*' show weakly negative deviations in the lowest population density class. This result suggests these impacts are dependent on road networks, which tend to be more ubiquitous in populated areas as opposed to sparsely inhabited zones. '*Inundation*', '*evacuation*' and '*stranded cars*' categories show very strong association with densely populated areas (>500 inhab./km$^2$), confirming that such impacts are most likely to occur in developed areas. Finally, 'rescue/fatalities' is strongly unassociated with the least dense category but is strongly related to the second population density bin (4–70 inhab./km$^2$), while we would have expected a signal towards very high densities. This association is very interesting, as it shows that our most severe impact category does not necessarily occur in the most heavily populated areas. Again, some caution is warranted in interpreting results in the most extreme categories due to the small sample size.

Land use is the second GIS-sampled characteristic to be crossed with SHAVE impacts (Fig. 6). The deviation chart also shows clear trends, from natural to more and more developed areas. '*Overflow*' and '*green land*' have weak deviations but show a slight tendency towards natural, rural and open space land cover classes, confirming the previous association with sparsely populated regions. Also, the strongest deviation for the '*green land*' category is with '*pasture*/*crops*', which correctly matches our classification. '*Street*/*roads*' does not show strong deviations either, but this impact class may have combined events that occurred in both urban (streets) and rural zones (roads, highways). It is intriguing that the most severe impacts (i.e. '*road closure*', '*inundation*', '*evacuation*', '*rescue*/*fatalities*' and especially '*stranded cars*') are all strongly associated with the most developed areas, with high percentages of impervious surfaces, while they are strongly unassociated with natural and less developed classes. These results indicate that these impacts, which by definition should be linked to urbanized areas, were correctly classified. But most of all, *it shows that such a land cover classification*, *at high resolution*, *can be a useful tool to define exposure to flash flood*, *and ultimately*, *help to predict the locations of such impacts.* Moreover, note that the most severe impact, 'rescue/fatalities', which was not associated to the highest population density bins, is here related to the most developed land use class. This contradiction is certainly due to the difference of spatial resolution between grids of population density and land use. For instance, a

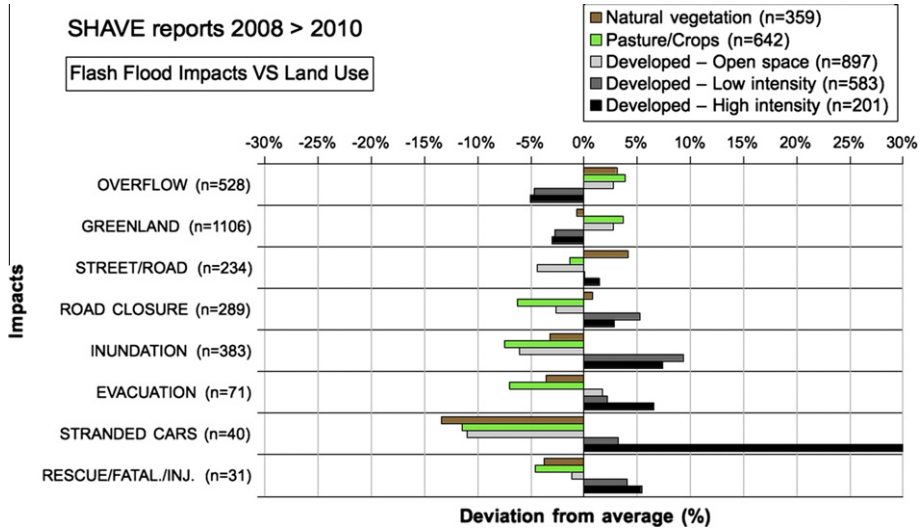


**Fig. 6.** Results of the crossing between SHAVE impacts and land use: bar chart representing deviation from average (%).

rescue can occur on a road (represented by a line of 'highly developed' 30-m land use grid-cells), but situated in a sparsely populated area (inside a 1-km population density grid-cell).

The last GIS-sampled attribute crossed with impacts classes is the maximum flow accumulation value ('*drainage area*'), sampled within a radius of 300 m of the SHAVE report (Fig. 7). In general, the chart shows weaker deviations than for the previous ones (recall the low Cramer's *V* value), especially for '*green land*', '*street/roads and road closure*'. But for the other impacts, significant trends can still be seen. First, the '*overflow*' impact (i.e., rivers out of their banks) is mostly associated with the larger drainage bin (>20 km$^2$). This is a consistent result, as streams are, by definition, the representation of flow convergence. Secondly, '*inundation*' shows a positive signal towards smaller drainage areas ($\leqslant$0.75 km$^2$) and a negative signal for higher flow accumulations (>2 km$^2$). This result is quite contradictory, as we would expect inundations (previously associated with standing waters) to be related to larger flow accumulations. This may be due more to poor infiltration in urban areas, as inundations are associated with highly developed, impervious areas. Finally, the three most severe impacts ('*evacuation*', '*stranded cars*' and '*rescue/fatalities*') have positive deviation in drainage areas between 0.75 and 20 km$^2$.

## 4. Example of use of NWS and SHAVE impact-focused dataset: two case studies

Launched in the mid-1980s, the operational flash flood prediction tools in the US are radar-based and rely on the concept of Flash Flood Guidance (FFG) (Georgakakos, 1986). Alternative approaches to FFG have been recently developed using spatially-distributed land surface and soil characteristics maps (the Gridded FFG, herein called GFFG), 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)) (Reed et al., 2007) to predict flash flood impacts for two extreme cases of flash flooding in Oklahoma, USA.

### 4.1. Flash Flood Guidance (FFG)

The concept of Flash Flood Guidance (FFG) is the threshold rainfall over nominal accumulation periods of 1, 3, and 6 h (and sometimes 12 and 24 h) required to initiate flooding in small streams that respond to rainfall within a few hours. 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 (QPEs) or forecasts, are used by forecasters to issue flash flood warnings when observed or forecast rainfall rates exceed the thresholds. FFG is computed in two steps. First, the threshold runoff (L) required to cause flooding (bankfull conditions) at the basin outlet is computed. In the NWS, this value is derived by dividing the estimated 2-years 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 5 km$^2$ 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-h accumulation periods, given initial soil moisture and stream states. The method employed in 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., in preparation-b). 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.

### 4.2. Gridded Flash Flood Guidance (GFFG)

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 modeled soil saturation (Gourley et al., in preparation-b). 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 4 km. Note that GFFG is progressively replacing FFG as an operational tool in several US River Forecast Centers (RFCs), but FFG is still in use for some RFCs. For this reason, FFG and GFFG products are hardly available simultaneously for a particular area.

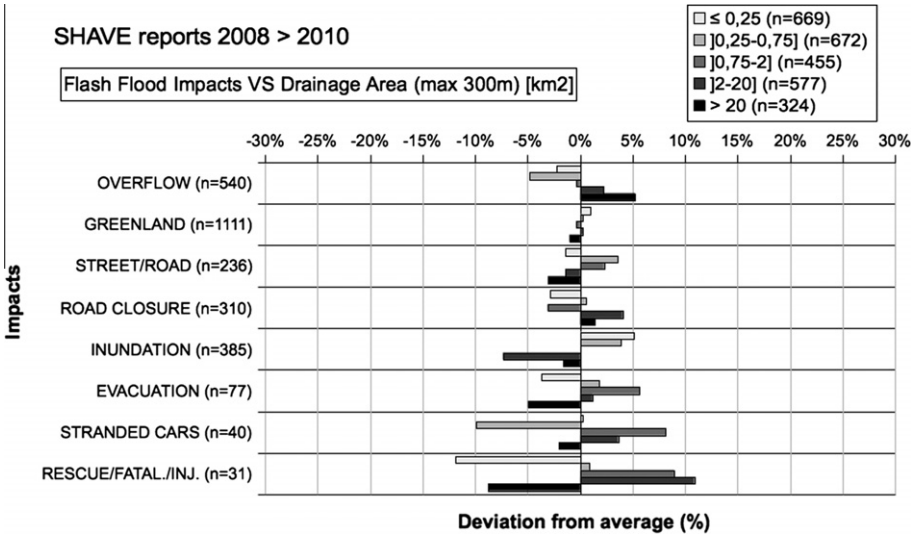


**Fig. 7.** Results of the crossing between SHAVE impacts and land use: bar chart representing deviation from average (%).

### 4.3. Distributed Hydrological Model – Threshold Frequency (DHM-TF)

The Distributed Hydrological Model – Threshold Frequency (DHM-TF) deviates from FFG in that it uses observed or forecast rainfall as a direct forcing to the hydrological model, rather than determining the rainfall thresholds in scenario mode (Gourley et al., in preparation-b). The method consists of running a distributed hydrologic model at each grid point using historical rainfall historic data; this allows simulated runoff to 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 1, 2, 5 years, etc. In forecast mode, DHM-TF is forced with real-time, radar-based rainfall or model forecast rainfall. Exceedance of simulated flows over the threshold return period flows (in this study, a 2-years return period flow) is the basis for alerting on an impending flash flood (Gourley et al., in preparation-b).

### 4.4. Results

In this analysis, two flash flood case studies (considered extreme events) were chosen, for which at least two of the three forecasting tools were available: the flash floods caused by Tropical Storm (TS) Erin over the state of Oklahoma in 2007, and the Oklahoma City flash flood event of 2010. These events occurred at different spatio-temporal scales (Fig. 8). The first flash flood case study was caused by the remains of Tropical Storm Erin, which crossed Oklahoma from west to east from 18 to 20 August 2007. Rainfall rates of over 76 mm/h were common, with significant flash flooding reported in numerous counties. Rainfall amounts exceeded 127 mm over a large area, with some locations receiving 203–254 mm. The second flash flood case was at a smaller spatio-temporal scale. It was mainly an urban event, occurring 14 June 2010 over the Oklahoma City metro area. The thunderstorm lasted about 7 h and rainfall rates averaged 25–50 mm/h, with some thunderstorm bands producing rates near 76 mm/h. A total of 127–228 mm was reported over the area, with up to 305 mm over the north-central portion of Oklahoma City.

One-hour accumulation periods for FFG and GFFG were chosen, rather than 3- or 6-h, as they showed better skill compared to flash flood observations (see Gourley et al., 2012a). The hourly FFG, GFFG, and DHM-TF products were collected over the whole Arkansas-Red River Basin and over a timeframe of 8 h prior to the meteorological event to 2 h after. Because FFG and GFFG represent rainfall thresholds triggering flash flooding, these values can be directly compared with precipitation estimates by computing ratios. As soon as the Rainfall-to-Guidance ratio is equal to or higher than one, forecasters at local NWS offices consider issuing flash flood warnings. QPEs were taken from the hourly multi-sensor Stage IV product, a mosaic of US radar-based rainfall rates and rain gauges (for more information, see http://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/stage4/). These QPE values were then used to calculate ratios of QPE-to-FFG and QPE-to-GFFG for every hour in the event. Then, in order to map the entire flash flood event, the maximum values of these hourly ratios and DHM-TF return periods grids were extracted to create a single grid of maximum values. Finally, NWS- and SHAVE-enhanced impact reports developed from this study were overlain on the gridded forecast products to determine the association between the forecast and the impact. To take into account the uncertainty of impact location with the reports, maximum ratios were searched within 7.5-km radii surrounding each impact for the TS Erin case and 1.5-km radii for the Oklahoma City case. Different search radii correspond to the NWS reports that were used for the TS Erin case and higher-density

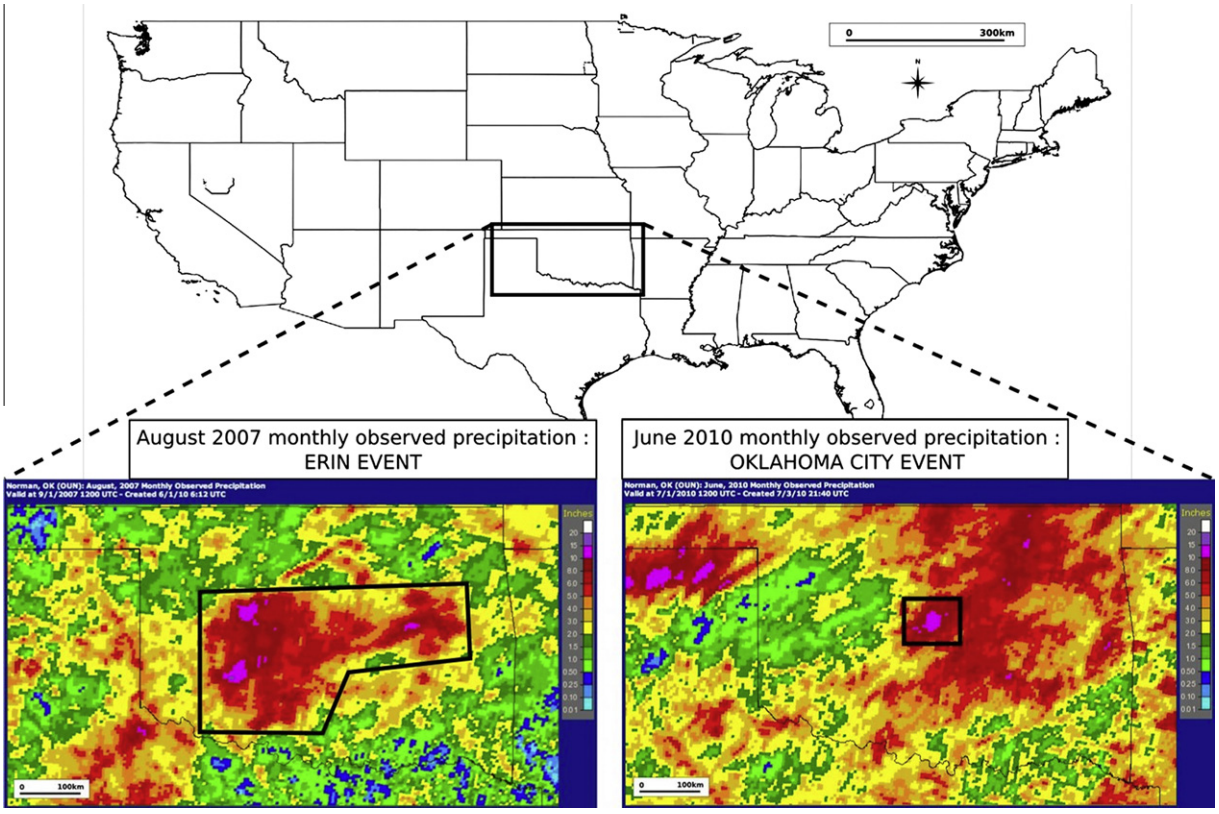


**Fig. 8.** Presentation of the TS Erin and Oklahoma City case studies: maps of observed monthly precipitation. Source: National Weather Service, http://water.weather.gov/precip/.

**Table 4**
Forecasting contingency table.

| | Forecast | No forecast |
|---|---|---|
| Observed | Hit | Miss |
| Not observed | False alarm | Correct negative |

SHAVE reports for the Oklahoma City case. Recall that bankfull conditions are expected when QPE-to-FFG and QPE-to-GFFG ratios exceed 1, or a DHM-TF return period exceeds 2 years. These limits were used to define if the tools forecasted a flash flood event in order to populate contingency tables (Table 4).

Three statistics were then computed from the hits, misses, and false alarms (only available with the SHAVE reports in the Oklahoma City case) in each of the contingency tables. The Probability Of Detection (POD) describes the fraction of observed flash floods that were correctly forecasted:

$$\mathrm{POD} = \mathrm{hits}/(\mathrm{hits} + \mathrm{misses}) \quad (1)$$

A POD of 1 indicates all flash floods were correctly forecasted while 0 indicates the forecast tools detected no flash floods. The False Alarm Ratio (FAR) describes the fraction of forecasted events that were not associated with observed events:

$$\mathrm{FAR} = \mathrm{false\ alarms}/(\mathrm{hits} + \mathrm{false\ alarms}) \quad (2)$$

Similar to POD, FAR ranges from 0 to 1, with zero indicating no forecasted events were unobserved and 1 indicating all forecasted 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:

$$\mathrm{CSI} = \mathrm{hits}/(\mathrm{hits} + \mathrm{misses} + \mathrm{false\ alarms}) \quad (3)$$

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

#### 4.4.1. Tropical Storm Erin

For the TS Erin case, only NWS point-based impacts were available to evaluate the forecast tools. Three maps compare impacts with hourly maximum QPE-to-FFG (Fig. 9), QPE-to-GFFG ratios (Fig. 10), and hourly maximum DHM-TF return periods (Fig. 11). A symbology was created for each impact (see Fig. 2). Furthermore, to better illustrate the multi-impact aspect, the most severe impacts are symbolized by white squares and the second most severe by white circles. In each map, property damage estimations are labeled for each report, and primary roads and major streams are also included as overlays. A first analysis was done on a YES/NO event basis by studying the forecasting tool maps and computing skill statistics. From Figs. 9–11, we can see that all flash flood forecasting tools correctly located the global area impacted by the flash flood. Note that GFFG identified the smaller extent of forecast flooding (ratios > 1) compared to the other two tools, yet it is still missing a few impacts. Although FFG and DHM-TF may detect almost every impact, they also forecast large areas where no impacts were reported by the NWS. This could mean that either: (1) the tools are overestimating flash flood impacted zones, (2) these zones were impacted but no NWS report was collected, or (3) these zones experienced flash flooding but there was little vulnerability or exposure, so no impact. Because the NWS dataset does not include reports of no flooding, the POD was the only skill statistic that we computed (see Table 5). The tool with the best detection skill was DHM-TF, followed by FFG, and then, GFFG. However, these high detection skills may also be associated with high FAR values, which unfortunately cannot be estimated. It should also be noted that because ratios were sampled by taking the maximum within 7.5-km of the impact, it artificially increases the POD.

A second analysis was done by comparing sample ratio values for each tool as a function of impacts in order to assess the ability of these tools to distinguish impact categories (Figs. 12–14). Grey diamonds represent sampled ratio values, black squares are the average value per impact category, and a vertical line delimits the forecast of bankfull and non-bankfull conditions (a detected flash flood or not). It is important to note the very high tool values for this flash flood case (with average FFG and GFFG ratios up to 2, and DHM-TF return periods up to 200 years), indicating the tools confirmed the extreme nature of the TS Erin event. Also, there is considerable spread in the distribution of values for each impact

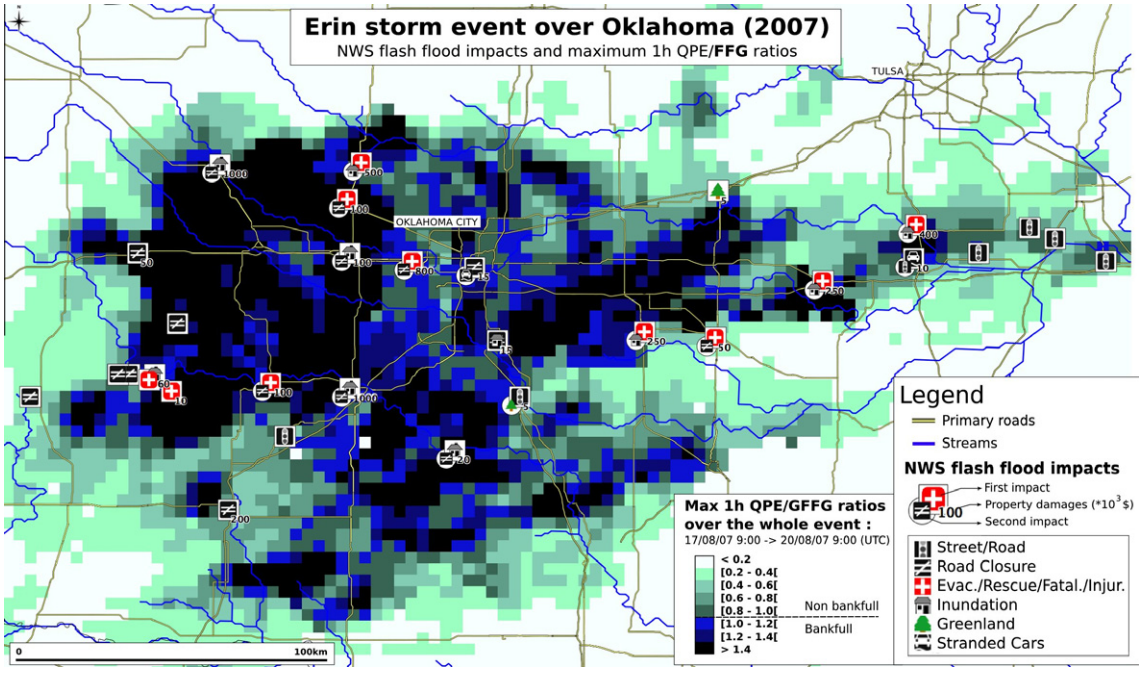


**Fig. 9.** NWS flash flood impacts and maximum 1 h QPE-to-FFG ratios for the TS Erin event.

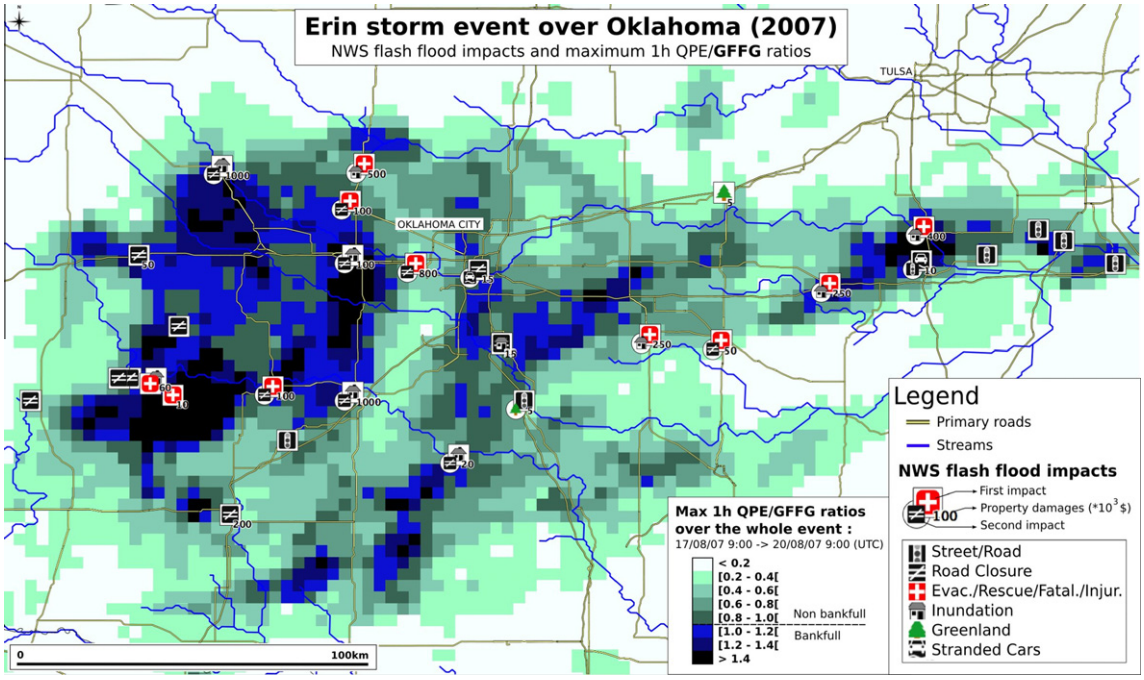


**Fig. 10.** NWS flash flood impacts and maximum 1 h QPE-to-GFFG ratios for the TS Erin event.

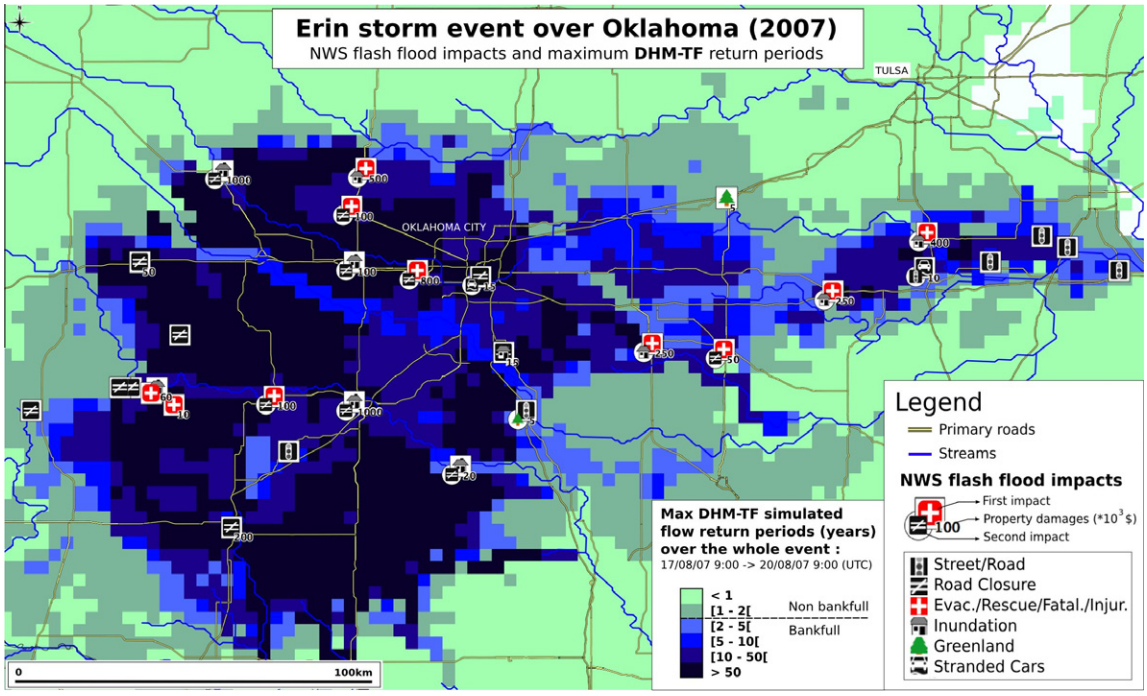


**Fig. 11.** NWS flash flood impacts and maximum DHM-TF return periods for the TS Erin event.

class (grey diamonds) and small sample sizes (for instance, only three '*green land*' impact occurrences). Therefore, these results must be taken cautiously until new case studies can be included for future analysis. Nevertheless, these plots confirm that a great majority of impacts is detected (in the bankfull zone values) by FFG and DHM-TF, whereas, GFFG shows more undetected impacts ('*green land*' is not detected at all, but this tendency might be due

**Table 5**
Probability Of Detection results for the TS Erin impacts sampling.

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

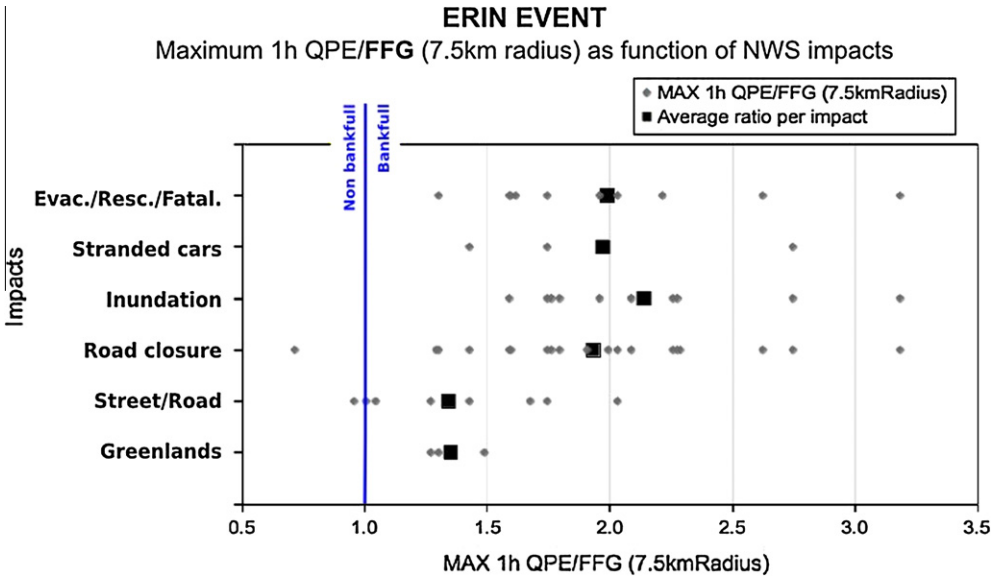


**Fig. 12.** Sampled maximum 1 h QPE-to-FFG ratios as function of impact classes for the TS Erin event.

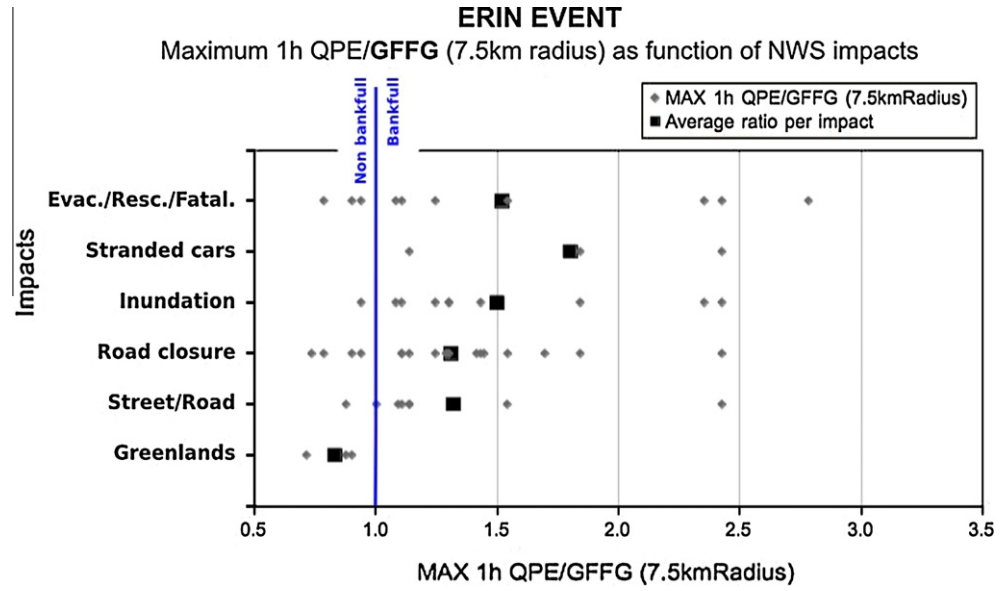


**Fig. 13.** Sampled maximum 1 h QPE-to-GFFG ratios as function of impact classes for the TS Erin event.

to the small sample size). Regarding average ratio values for the impact classes, there is a general tendency for higher ratio values being associated to the more severe impacts. This result indicates that, for this particular case study, all three tools are able to make a distinction between less severe and more severe impacts.

*4.4.2. Oklahoma City event*

For this smaller-scale urban event, only SHAVE impacts were used because they include reports of no impact (white points on the maps), so that the FAR and CSI can be readily computed. Moreover, NWS reports are represented by polygons in 2010, which was inconvenient for this study, because they are often the size of the whole metro area. In this case, we evaluated the 1-h QPE-to-GFFG ratios (Fig. 15) and DHM-TF return periods (Fig. 16); the FFG method had been replaced by GFFG at the operational River Forecast Center by this time and was unavailable. As with the previous case, skill was first analyzed on a YES/NO event basis (computation of POD, FAR and CSI). Recall that the sampling of the maximum values is made within a radius of 1.5 km from the report point due to the small-scale nature of the event and high-density SHAVE reports.

Maps of enhanced flash flooding impacts versus forecasting tools show that flash flood forecast patterns correctly match the global extension of impacts (Figs. 15 and 16). Yet, there are many forecast grid cells associated with null reports, which seems to indicate numerous false alarms. Additionally, as the reports are point-based, the assessment of such false alarms for gridded models was found to be problematic. For example, when an observed impact easily validates a forecast pixel, a null report cannot invalidate a whole $4 \times 4$ km forecast pixel, in which impacted zones may not be sampled by SHAVE point clusters. Moreover, an impact

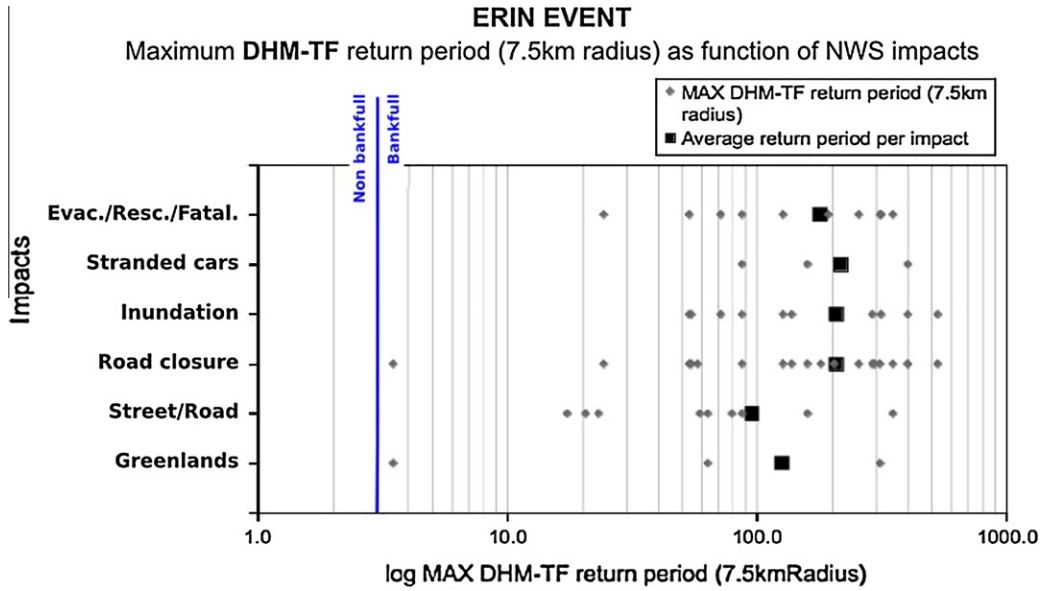


**Fig. 14.** Sampled maximum DHM-TF return periods (log scale) as function of impact classes for the TS Erin event.

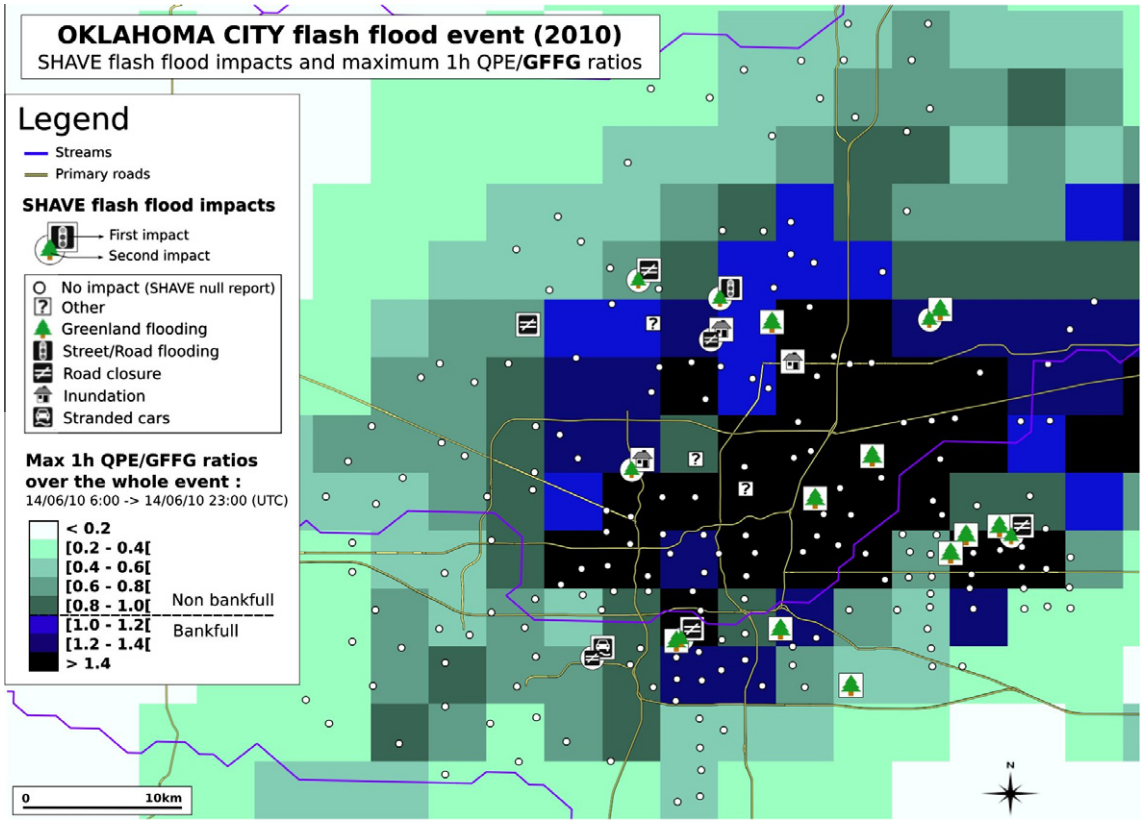


**Fig. 15.** SHAVE flash flood impacts and maximum 1-h QPE-to-GFFG ratios for the Oklahoma City event.

might be located next to the report point, but not seen by the interviewee. Nevertheless, to assess forecasting tools skill on a YES/NO event basis, POD, FAR and CSI were computed for both tools in (Table 6). Results show that DHM-TF has the highest POD (1), but also the highest FAR (0.88). GFFG has the best CSI, with a score of 0.14, despite a lower POD value (0.86), but it also has a better FAR (0.85). While the GFFG tool apparently has better skill than DHM-TF for this particular case, both CSI values are still quite low. The CSI values also agree with the highest value found by Gourley et al. (2012a) for the 1-h GFFG tool (0.12) using NWS reports over the entire Arkansas-Red River Basin from September 1, 2006 to August 22, 2008.

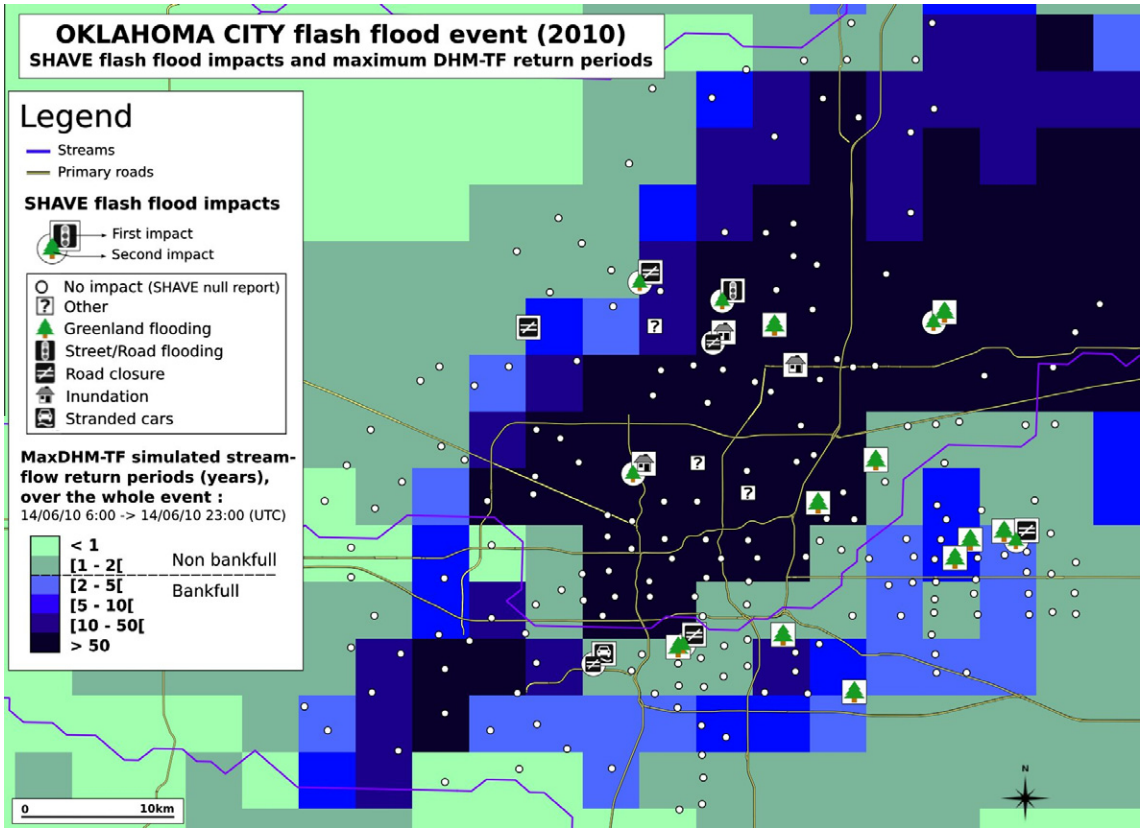


**Fig. 16.** SHAVE flash flood impacts and maximum DHM-TF return periods for the Oklahoma City event.

**Table 6**
Probability Of Detection, False Alarm Ratio and Critical Success Index results for the Oklahoma City impacts sampling.

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

## 5. Flash flood reports: critical analysis and ways of improvement

What are the strengths and weaknesses of the NWS and SHAVE observational datasets when it comes to the evaluation of flash flood forecasting models? How could they be improved? This study highlighted remaining challenges for forecast evaluation, in the particular case of flash flooding, and provides specific recommendations for improving the data collection methodology in this section.

### *5.1. Recommendation 1: estimate the timing of sudden events*

Both datasets were found to provide poor event timing estimates. In NWS reports, the meteorological event timing is often taken as flash flood timing; whereas for SHAVE reports, the general public might not be able to estimate the event start time (e.g., if it started overnight) or only give rough estimations. Even worse, when the timing is unknown, the recorded event start/end time is simply the time of the 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 often associated with significant uncertainty. To permit a meaningful temporal analysis, accurate timing (hourly to minute timescales) associated with the flash flood event must be recorded, including an estimation of the range of timing error. If temporal information is not available, an '*unknown*' category needs to be added.

Together with the location, the issue of timing is particularly important in the case of flash flooding because it allows connecting environmental circumstances with social activities determining the level of human exposure to a particular event (Creutin et al., 2009; Ruin et al., 2008). Unfortunately, this definitely appears as a weakness of most data collection strategies and new methods must be developed to address this need. The use of existing traffic or security cameras could be experimented with as well as the information provided through social networks (e.g. YouTube videos, Twitter posts). Therefore a deep understanding of what those tools could bring, what their limits are, and how they could be coupled for impact-based database enhancement is needed.

### *5.2. Recommendation 2: improve the spatial delineation of events*

This study showed that neither the NWS event polygons, nor the SHAVE high-resolution poll-based points were entirely appropriate to correctly delineate flash flood patterns to be compared with gridded forecasts. The current NWS polygons have poor spatial accuracy, even if they were quality controlled and drawn by professionals familiar with their area of responsibility. While for SHAVE, even with precisely geolocated reports, events are described by the general public, who are most likely untrained to provide accurate descriptions. Therefore, the described flash flood may occur from 1 m to a few kilometers around the report point, depending on people's perception or knowledge. For instance, people living in urban areas might be more aware of their immediate neighborhood's flooding, whereas, a farmer may be aware of what is happening on his entire property, which could represent a much larger area. Therefore, uncertainty buffers must be considered

around SHAVE report points. Also, SHAVE and NWS sampling strategies are more storm-targeted rather than having hydrologic relevance. In order to better delineate the diffuse and small-scale patterns of flash floods, report datasets should target their sampling to probe each small basin (<20 $km^2$), to be sure to report the state of small streams. This could be made by creating lists of potential interviewees living close to streams and notifying them in advance of the possibility to be called about heavy rain (i.e., using people as stream gauges). Though, in the particular case of urban flooding, the laws of natural hydrology are hardly valid, so the sampling could stay randomly distributed. Lastly, to facilitate improved spatial analysis of various types of impacts, we recommend the use of GIS tools in the collection of flash flood reports. Polygons are useful for contouring rainfall patterns, for example, but flash flood impacts are often more diffuse and may be associated to difficult-to-contour features such as road networks.

### 5.3. Recommendation 3: estimate random and systematic errors for human reports

The SHAVE dataset is largely based on survey responses by the untrained public. It is quite likely that perceptions influence responses and introduce bias. In the case of the Oklahoma City event, student callers fortuitously contacted local emergency management officials, and obtained a wealth of high-quality, unbiased data. But, this was the exception rather than the norm. SHAVE was initially designed to collect physical data rather than information about the interviewees themselves. But, this information is needed. Future questionnaires should include information about the interviewee's age, gender, profession, level of education, income, etc. Indeed, these parameters are likely to influence people's perception, and therefore, their description of the event (Brilly and Polic, 2005). As for spatial representation of their reports, the public should be asked how far they can see or how large their property bounds are. This information could be used along with GIS buffers to essentially assign a perimeter associated to each report.

### 5.4. Recommendation 4: measure false alarms

By collecting only positive reports of flooding, NWS reports do not readily allow an estimation of false alarms. Also, even though SHAVE provides reports of no flooding, the analysis of the Oklahoma City case study showed that this point-based information was not entirely appropriate to evaluate false alarms for distributed forecasting tools. Using polygon-delineated null reports should be more convenient to assess false alarms in the context of gridded forecasting models. The area of null reports included in forecast grid cells could then give the metric needed for the assessment of false alarms.

## 6. Conclusion

This paper is targeted to researchers and practitioners interested in deepening their understanding of flash flood impacts. Nevertheless it also provides insights and ideas in advancing ways of using spatial and temporal datasets for other hydrologic and non-hydrologic hazards. This study provides an impact classification of flash flood report datasets over the United States to evaluate the ability of US flash flood forecasting tools to predict such categories of impacts, and subsequently, to identify the problems and improvements that can be made to these flash flood report methodologies. After presenting the flash flood report datasets (NWS and SHAVE), the method chosen for impact classification was described and impact-enhanced datasets were created. SHAVE impacts were then used in a spatio-contextual analysis, via a cross tabulation method based on perceived attributes already included in the SHAVE dataset ('*water movement*', '*return period*' and '*water depth*'), as well as, GIS-sampled spatial attributes ('*land use*', '*population density*', '*local upslope*' and '*drainage area*'). The first result of this analysis is that associations found using cross tabulation are consistent with the impact classification. This is true for perceived attributes already included in SHAVE, as well as for independent spatial attributes (from 30 m to 1 km resolution) sampled through GIS. These meaningful results also show that the SHAVE dataset is a trustworthy tool for flash flood characterization, even if it is based on public polls. Moreover, by crossing impact categories with socio-spatial characteristics, this analysis showed first benchmarks for the use of exposure layers in future flash flood impact forecasting models: the NLCD2006 land use raster at a spatial resolution of 30 m appears to be a simple yet effective tool for flash flood exposure characterization, and a typical drainage area range for severe flash flood impacts was identified: 0.75–20 $km^2$.

The second part of this study consisted of 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 flooding in Oklahoma (Tropical Storm Erin in 2007 and the Oklahoma City urban flash flood in 2010) were chosen to evaluate the tools on a YES/NO-forecast basis (i.e., computing POD, FAR, and CSI), but also as a function of the impacts. For the YES/NO event analysis, FFG and DHM-TF detected a great majority of impacts, whereas GFFG showed more undetected impacts. There was a general tendency for the more severe impacts to be associated to higher mean exceedances over FFG and GFFG. This means that, at least for these particular case studies, the tools were able to make a distinction between less severe and more severe impacts. Of course, the analysis of these two particular cases should be supplemented with a study of the whole NWS and SHAVE dataset. This would provide more samples and produce more robust statistics for tool evaluation. It should be noted that these tools were not designed to take into account flash flood impacts, which are the combination of a hazard (in this study, quite well described by the tools), but also exposure and vulnerability. This result demonstrates that these two terms of the risk equation must be assessed in more detail.

Finally, a critical analysis of the NWS and SHAVE data collection methodologies was completed and specific recommendations are now provided for future datasets designed to collect details on flash flooding. These main challenges include: (1) the need for more accurate estimates of the event onset and recession, (2) A refined, hydrologically relevant delineation of impacted zones using GIS tools, (3) An estimation of systematic and random errors associated to public polls and (4) a more reliable method for false alarm quantification. Future work will expand the impact-focused evaluation of flash flood predictability across a larger study domain in space and time. We will also use the impact classifications developed in this study to refine the forecast tools so that they incorporate specific information about social vulnerability and exposure.

## Acknowledgements

This work was supported by the French National Research Agency (ANR) through the Project ADAPTflood funded by the programme "Retour Post-Doctorant" (ANR-09-RPDOC-001-01). The authors would like to acknowledge Zachary Flamig, Race Clark (National Severe Storm Laboratory, Norman, OK) for data processing and Jill Hardy for proofreading this manuscript.

## References

Antoine, J.-M., Desailly, B., Gazelle, F., 2001. Les crues meurtrières, du Roussillon aux Cévennes: «Deadly floods, from Roussillon to Cévennes» (France). Ann. Geogr. 622, 597–623.

Ashley, S.T., Ashley, W.S., 2008. Flood fatalities in the United States. J. Appl. Meteor. Climatol. 47, 806–818.
Brakenridge, G.R., Nghiem, S.V., Anderson, E., Chien, S., 2005. Space-based measurement of river runoff. EOS, Trans. Am. Geophys. Union 86, 185–188.
Brilly, M., Polic, M., 2005. Public perception of flood risks, flood forecasting and mitigation. Nat. Hazards Earth Syst. Sci. 5, 345–355.
Coates, L., 1999. Flood fatalities in Australia, 1788–1996. Aust. Geogr. 30, 391–408.
Creutin, J.D., Borga, M., Lutoff, C., Scolobig, A., Ruin, I., Creton-Cazanave, L., 2009. Catchment dynamics and social response during flash floods: the potential of radar rainfall monitoring for warning procedures. Meteorol. Appl. 16, 115–125.
Danielson, J.J., Gesch, D.B., 2011. Global multi-resolution terrain elevation data 2010 (GMTED2010): U.S. Geological Survey Open-File Report 2011-1073, 26 p.
Duclos, P., Vidonne, O., Beuf, P., Perray, P., Stoebner, A., 1991. Flash flood disaster-Nîmes, France, 1988. Eur. J. Epidemiol. 7, 365–371.
Ehlschlaeger, C., 1989. Using the AT search algorithm to develop hydrologic models from digital elevation data. In: Proceedings of International Geographic Information Systems (IGIS) Symposium '89, 275–281 (Baltimore, MD, 18–19 March 1989).
Few, R., Ahern, M., Matthies, F., Kovats, S., 2004. Floods, health and climatic change, a strategic review, Tyndall Centre for Climate Change Research Working Paper 63.
French, J.G., Ing, R., Von Allmen, R., Wood, R., 1983. Mortality from flash floods: a review of the National Weather Service reports, 1969–1981. Pub. Health Rep. 98, 584–588.
Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., et al., 2011. Completion of the 2006 National Land Cover Database for the Conterminous United States. PE&RS 77, 858–864.
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.
Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., et al., 2009. A compilation of data on European flash floods. J. Hydrol. 367, 70–78.
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. 67, 1233–1239.
Gourley, J.J., Erlingis, J.M., Smith, T.M., Ortega, K.L., Hong, Y., 2010. Remote collection and analysis of witness reports on flash floods. J. Hydrol. 394, 53–62.
Gourley, J.J., Erlingis, J.M., Hong, Y., Wells, E., 2012a. Evaluation of tools used for monitoring and forecasting flash floods in the United States. Wea. Forecast. 27, 158–173.
Gourley, J.J., Flamig, Z.L., Hong, Y., Howard, K.W., in preparation-b. Evaluation of past, present, and future tools for radar-based flash flood prediction. Hydrol. Sci. J.
Gourley, J.J., Hong Y., Flamig, Z.L., Arthur, A., Clark, R., Calianno, M., Ruin, I., Ortel, T., Wieczorek, M.E., Clark, E., Kirstetter, P.-E., Krajewski, W.F., in preparation-c. A unified flash flood database over the US. Bull. Amer. Meteorol. Soc.
Gruntfest, E., 1977. What people did during the big Thompson flood. Working Paper #32, Institute of Behavioral Science, University of Colorado, Boulder, CO 1977, 35pp.
Hong, Y., Adhikari, P., Gourley, J.J., 2012. Flash flood. In: Bobrowsky, Peter (Ed.), Encyclopedia of Nat. Hazards. Springer.
Jonkman, S.N., 2005. Global perspectives on loss of human life caused by floods. Nat. Hazards 34, 151–175.
Jonkman, S.N., Kelman, I., 2005. An analysis of the causes and circumstances of flood disaster deaths. Disasters 29, 75–97.
Montz, B.E., Gruntfest, E., 2002. Flash flood mitigation: recommendations for research and applications. Environ. Hazards 4, 15–22.
Owen, T.W., Gallo, K.P., 2000. Updated population metadata for the United States Historical Climatology Network stations. J. Climate 13, 4028–4033.
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. 337, 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. J. Hydrol. 361, 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, pp. 6B.2.
Sharif, H.O., Jackson, T., Hossain, M., Bin-Shafique, Sazzad, Zane, D., 2010. Motor vehicle-related flood fatalities in texas, 1959–2008. J. Trans. Saf. Secur. 2, 325–335.
Simpson, M.R., Oltmann, R.N., 1993. Discharge-measurement system using an acoustic Doppler current profiler with application to large rivers and estuaries. US Geol. Survey Water-Supply Paper 2395.
Staes, C., Orengo, J.C., Malilay, J., Rullan, J., Noji, E., 1994. Deaths due to flash floods in Puerto-Rico, January 1992: implications for prevention. Int. J. Epidemiol. 23, 968–975.
Vinet, F., Lumbrosso, D., Defossez, S., Boissier, L., 2011. A comparative analysis of the loss of life during two récent floods in France: the sea surge caused by the Storm Xynthia and the flash flood in Var. Nat. Hazards 61, 1179–1201.
World Bank, 2010. Natural Hazards, Unnatural Disasters: The Economics of Effective Prevention. Washington.
Zahran, S., Brody, S.D., Peacock, W.G., Vedlitz, A., Gover, H., 2008. Social vulnerability and the natural and built environment: a model of flood casualties in Texas. Disasters 32, 537–560.