Nat Hazards (2015) 75:71–94
DOI 10.1007/s11069-014-1305-2

ORIGINAL PAPER

Deriving information on disasters caused by natural hazards from limited data: a Guatemalan case study

Agnes Soto

Received: 30 August 2013 / Accepted: 14 June 2014 / Published online: 29 June 2014

Abstract This work proposes a method to overcome data limitations encountered when analyzing disasters at a local scale in disaster-prone areas. Research is required to understand the processes underlying the disasters in disaster-prone areas. However, many such areas lack sufficient data for the statistically significant studies that would strongly support disaster risk-reduction measures. Disasters are studied extensively at the national scale, but local-scale disaster research is greatly needed, specifically because the hazard exposures and vulnerabilities of populations are mainly site specific. The proposed method compiles data using two databases: the System of Information Management in case of Emergency or Disaster (SISMICEDE) and the Disaster Inventory System (DesInventar). SISMICEDE has a short time span and high spatial resolution, while DesInventar has a longer time span but low spatial resolution. SISMICEDE's spatial distribution was used to sort DesInventar disaster data, analyzing them spatially and temporally at a local scale. The Samala River basin in Guatemala was selected to exemplify a disaster-prone area for which there are insufficient disaster data. The results indicate that it was useful to combine the two databases to optimally describe disasters over time and space in the studied area. The refinement of the disaster data highlighted the discrepancies between administrative boundaries and local particularities. The results indicate that the municipal scale is too sparse for spatial analyses and that specific location details are needed. According to the limited data available, disasters, during the rainy season, are increasing over time in the study area. This work demonstrates a way to perform local-scale disaster studies of areas for which data are not readily available. These local-scale studies would enable research and actions intended to improve disaster risk-reduction management and measures. This study could also help promote an improved information system in Guatemala that includes complete information useful for emergency response and post-disaster analyses.

Keywords Disaster data analyses · Disaster temporal and spatial analyses · Guatemala · Disaster data compilation · Local scale

A. Soto (✉)
Earth Sciences Department, Uppsala University, Villavägen 16, 752 36 Uppsala, Sweden
e-mail: Agnes.soto@geo.uu.se

1 Introduction

Natural geological and hydrometeorological processes can be hazardous to human populations and cause disasters that affect people's lives and systems; disaster science identifies such processes as natural hazards. Disasters occur when natural hazards coincide in time and space with vulnerable people, negatively affecting their lives. The negative impacts caused by natural hazards necessitate actions to mitigate their effects. The damage caused by high-magnitude natural hazardous events, such as Hurricane Katrina or the Haiti earthquake, tends to attract considerable attention due to the magnitude of the events although they generally occur infrequently. In contrast, the damage caused by frequent but small hazardous events, such as seasonal floods in Pakistan or Guatemala, could be overlooked or underestimated because their cumulative impact may be less obvious (Alexander 1993), and geographical location can intensify this unbalanced attention (Adams 1986). However, small and frequent events can negatively affect the life quality of many people and therefore merit attention and research. Disaster risk reduction (DRR) has become an important issue because of the importance of the effects of disasters (UN/ISDR 2004; ISDR 2009). It is necessary to incorporate DRR as a standard tool in development management (OAS 1991; Sperling and Szekely 2005). Land-use planning, which usually involves actions that result in laws, regulations, and codes (Burby and Dalton 1994), has been identified as a key tool for implementing DRR (UN/ISDR 2004). However, DRR is not yet a default component of development planning by many national or local governments.

Spatial planning, as a technical tool within the planning process, should therefore incorporate DRR in its analyses and models.

Disaster research is partly general and partly site specific. Disasters due to natural hazards are caused by natural processes occurring at particular places and depend on the characteristics of the people ultimately affected by them. General research into disasters is important, and application of the resulting general understanding to the local level is required to render it practical. Scaling research into disasters to the local level is especially important in disaster-prone areas. It is vital to understand the processes by which specific local conditions define the potential incidents, spatial distribution, mechanisms, and risks leading to disasters.

Unfortunately, many disaster-prone areas are not adequately examined by disaster research. Local-scale disaster research is urgently needed to support spatial and development planning. It is at the local scale that operational decisions are made, and understanding disasters and their processes could provide stronger foundations for plans and actions intended to reduce the negative effects of natural hazards.

Disaster studies require data appropriate for their purposes, applications, and, consequently, the spatial and temporal scales at which they are to be used. There is no generally accepted methodology for collecting and organizing disaster data, so the available data are as varied as the collecting entities and their purposes (Guha-Sapir and Below 2002). The data for disaster research focused on particular places must have sufficient quantity and quality to be informative. However, informative disaster data are often lacking, particularly for areas highly exposed to natural hazards. A first impediment to disaster studies is thus the lack of available data. Second, as Hittelman et al. (2001) noted, most natural hazard catalogs are "collection[s] of conclusions based on imprecise and incomplete information" due to, for example, varied original data sources and formats, human errors, poor quality control, and diverse purposes and data-creation systems. Various specific problems have been identified in disaster databases. Amin et al. (2008) identified problems with the accuracy of the names and locations of disasters, lack of accurate maps, and deficient technical capacity of those responsible for data collection.

How to conduct research into disasters when data are limited is a key question addressed in certain recent studies. Ganapathy (2011) used Disaster Inventory System (DesInventar) data to analyze disasters in Vellore District, Tamil Nadu State, India. Although the availability of disaster data was relatively good for this study, as 30 years of data were available, the results obtained were district-scale frequencies of limited value for local application. Alfaro and Quesada (2010) analyzed the temporal trends of tropical cyclones and their impacts on populations in Costa Rica using 60 years of detailed national and global data from the Emergency Event Database (EM-DAT); they performed trend analyses at the local level, but the method could not be applied to other areas because of the lack of available data.

This work treats the Samala River basin in Guatemala as a case study to explore how disaster research can be performed with limited local-scale disaster data. The work uses the drainage basin as the study area because natural boundaries, not administrative ones, set the limits of the processes underlying the key natural hazards in this part of the world. Guatemala, like the rest of Central America, has been identified as highly vulnerable to natural hazards (Alcantara-Ayala 2002; Harmeling and Eckstein 2012). Despite the obvious risks facing the Guatemalan people, disaster studies at the regional or local scales are limited; existing risk and vulnerability studies of Guatemala are mostly at the national scale (Tesliuc and Lindert 2004; GFDRR 2009).

The Samala River basin is located in a tectonically and hydrometeorologically very active area that accommodates highly populated settlements—a combination that frequently leads to disaster. Very limited data are available for the country and consequently for the study area. Several international databases include Guatemala in their registers but, in general, the few existing records correspond mostly to those outstanding disasters in which international aid agencies have been involved (University of Richmond 2002–2006; ADRC 2009; CRED 2009). National data have been collected systematically only since 2008 and are used mainly for addressing emergencies when they occur and not for planning to reduce the risk. The selected study area might well be representative of many other areas in Central America and other disaster-prone areas of the world.

The aim of this study was to identify the usefulness of the currently available disaster datasets for spatial planning, and to find a way to use the available data to characterize disaster situations over time and space in the study area. The results should contribute to building a regional research base for DRR applications. Since many limitations were found in the available data, the study emphasized as follows: (1) evaluating the quality of data for the study area; (2) exploring the methodological possibilities for compiling data with availability limitations; and (3) assessing the potential for applying the data at the local scale, for example, in regional development. The study focused on the spatial and temporal dimensions of disasters, disregarding their resulting effects, mainly due to a lack of relevant input data in the identified datasets; however, the available data did not permit the distribution of hazards in the area to be described.

2 Study area

The Samala River catchment is part of the Guatemalan Pacific Ocean basin. Its 1,500 km^2 area comprises varied landscapes, from mountainous highlands of volcanic origin to coastal plains via steep volcanic flanks. The sudden change in elevation that creates this landscape variety is mainly the result of active tectonics at the convergence of the Cocos Plate with the North American and Caribbean plates (Kuenzi et al. 1979). An active volcano complex, the Santa Maria–Santiaguito complex, is located within the Samala River catchment. Santiaguito is an extrusive dacite dome complex erupting continuously since its 1922 formation on the flank of the Santa Maria volcano, which erupted explosively in 1902 (Kuenzi et al. 1979).

The influence of the surrounding Atlantic and Pacific oceans and the area's geographical location in the middle of the hurricane alley results in high rainfall concentrated in the six months from May to October each year, with high variability due to altitude differences in the catchment (from 0 to over 4,000 m above sea level). The constant eruptions build volcanic material deposits around the volcano and on surrounding mountains flanks. Precipitation waters carry these sediments to the streams flowing from the highlands to the Pacific Ocean. The sediments deposited along the river beds gradually decrease in size downstream, creating coastal plains built of alluvial fans of volcanic sediments.

The volcanic sediment influx and humid conditions make the study area fertile and agriculturally very productive. Guatemala is historically an agricultural country and people have settled and grown crops in this catchment since pre-Columbian times, especially on the coastal plain and in the highlands—a pattern that persists to the present. The catchment area is administratively organized into 29 municipalities. For 16 of these municipalities, the administrative boundaries do not correspond to the geomorphological limits of the Samala River catchment (Fig. 1).

The people living in the study area, due to its climate and volcanic setting, face potential hazards including volcanic eruptions, ash falls, landslides, flooding, and flash floods along the streams, especially during the rainy season and in areas where river beds become braided, such as the coastal plain.

3 Material, methods, and results

This work examines disasters using open databases. The databases comprise information on reported disasters that is created according to the context and purpose of each database. In this work, the information corresponding to a single disaster report is called a record. This decision was made due to the identified possibility of single reports referring to multiple disasters or imprecise descriptions of the referred-to disaster.

3.1 Data

Data on disasters due to natural hazards in Guatemala were searched online mainly through organizations concerned with disasters, emergency response, and humanitarian aid, and through the Guatemalan National Coordinator for the Reduction of Disasters (CONRED). The initial database selection was based on three criteria: Data must be available for Guatemala, the available data must provide location information for the records at least down to the department scale, and the period covered must exceed 10 years. Of the freely available international databases, EM-DAT and DesInventar were selected for this study. The national System of Information Management in case of Emergency or Disaster (SISMICEDE) database was also included in this study although it was established as late as 2008. SISMICEDE was included because it comprises detailed information that includes geographical coordinates for most of the entries giving higher spatial resolution than the

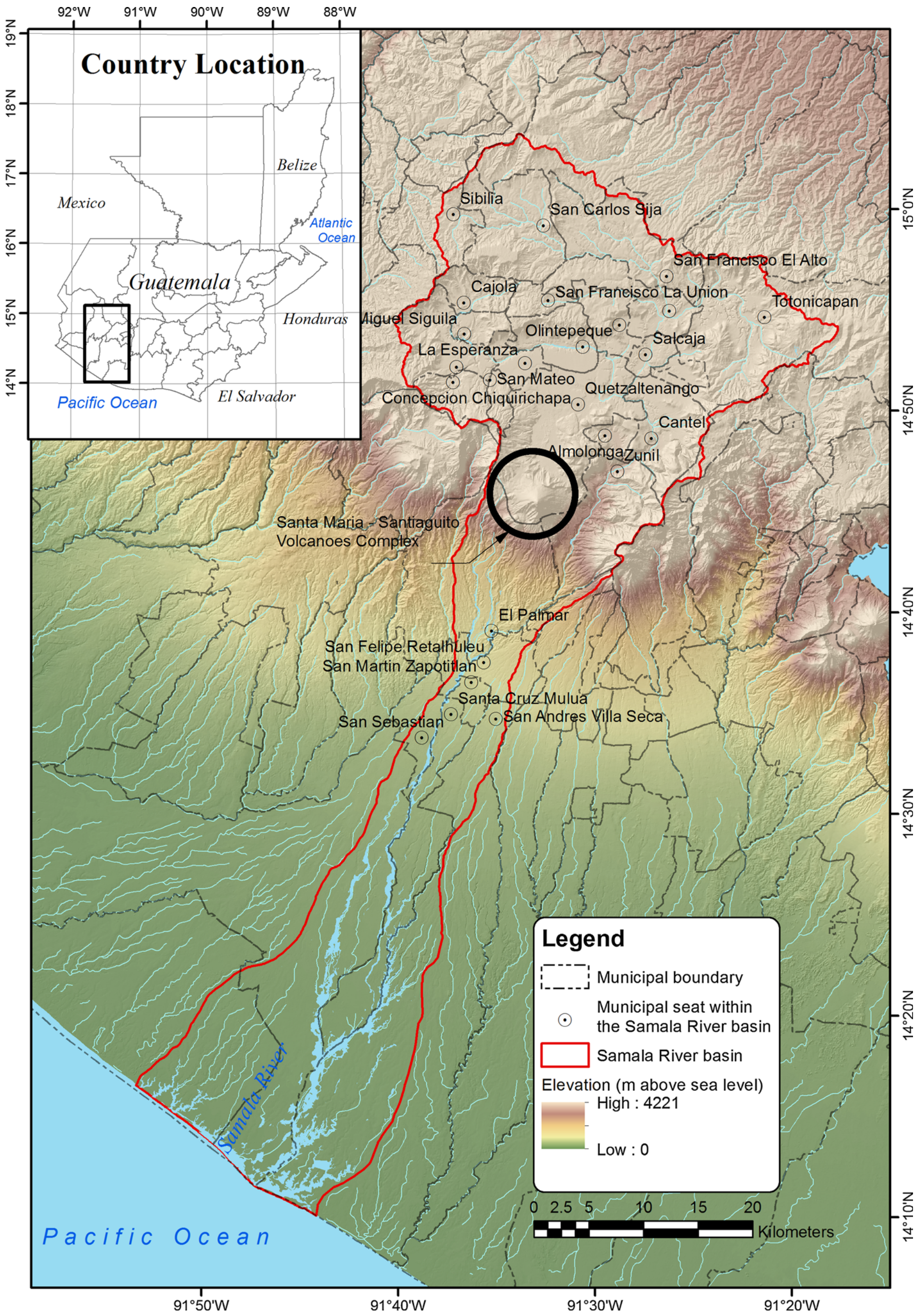

Fig. 1 Samala River basin. The inset map in (upper-left corner) shows the geographical location of the study area in Guatemala. The *black circle* indicates the Santa Maria–Santiaguito volcano complex. Geographical information source: Guatemalan Planning and Programming Secretariat (SEGEPLAN), 2012

other databases to the study. All datasets, as well as their data entry layout and potential limitations, were analyzed for the purposes of this research work.

EM-DAT has been created by the Centre for Research on the Epidemiology of Disasters (CRED) and is one of the best-known international sources of disaster data. Its goal is to "provide an objective basis for vulnerability assessment and rational decision making in disaster situations" (CRED 2009). EM-DAT data are collected globally at the country level and comprise information starting from 1900. Ideally, each record includes information on the disaster's underlying natural hazard, duration, location, description, losses, and damage in general terms (Table 1). The selection criteria of this dataset focus on mass disasters and require that a registered disaster have 10 or more people reported killed, 100 or more people affected, a declaration of a state of emergency, or a call for international assistance. The EM-DAT records thus correspond to those disasters that affected people at the national level. For Guatemala, EM-DAT has 85 reports of floods, storms, volcanic eruptions, earthquakes, and landslides from 1900 to 2011, 25 of which are located in the study area. EM-DAT's data limitations are determined by the selection criteria, which exclude disasters at the local scale, which although small could significantly affect local populations and their development. The country level of the database produces general-level information; for example, disaster location could be described in nonspecific terms, such as "Pacific coast," limiting the opportunity for specific geographical mapping. The complete dataset for Guatemala had to be requested from CRED due to incomplete location fields in the online-generated data.

DesInventar has been created by the Network of Social Studies in the Prevention of Disaster in Latin America (La Red), which aims to construct national databases of losses caused by disasters due to natural hazards in such a way that local-scale disasters are visible (DesInventar 2011). The system was implemented later in the Caribbean, Asia, and Africa, sponsored by the United Nations (UN), creating a bigger network using a more homogeneous methodology. DesInventar seeks to enable spatial and temporal analyses of the effects of disasters for risk-management purposes by recording those effects at a fine resolution (OSSO/La Red 2009). DesInventar obtains data from national institutional reports and newspapers. Ideally, each record includes information on the disaster's causal hazard (although not using a scientifically based classification system), duration, magnitude, location, description, source, and human and economic losses and damage (Table 1). Part of the information is categorized as "robust," meaning that it is quantitative. Other parts of the information are categorized as "fuzzy"; these describe effects that were not measured, so only the kind of effect is specified (Serje [date unknown]). The DesInventar dataset includes a total of 356 reports of incidents caused by natural hazards from 1988 to 2011 in the municipalities related to the catchment. The natural hazards related to the reports include volcanic activity, torrential floods, storms, sedimentation, rainfall, landslides, inundations, hail, gales, frosts, electrical storms, and earthquakes. The potential limitations of DesInventar data are determined by the sources used (mostly newspapers) and the unavailability of information for some data fields.

The SISMICEDE database, created by CONRED, has recorded, organized, verified, and updated the status of incidents affecting people in Guatemala as they have happened and at all scales since 2008. Its main purpose is to make data available to decision makers during emergencies (CONRED 2010). The database comprises information reported by CONRED's local delegates or by other institutions that are part of the CONRED system, so the information can be assumed to be reliable. Each record ideally includes information on the hazard's class, origin, duration, location, and probable cause, together with a description of the incident and of human and material damage (Table 1). Each incident is registered at the

Table 1 Comparison of the data entry templates of the EM-DAT, DesInventar, and SISMICEDE databases; information about effects is grouped by affected sectors/groups (i.e., people, houses, infrastructure, and the economy)

	EM-DAT	DesInventar	SISMICEDE
Identification	Code and name	Code	Code and name
Duration	Start and end dates	Start date and duration	Start and end dates
Magnitude	–	Magnitude	–
Location	Country and "place" (i.e., region or department level)	Municipal or village level; geographical coordinates	Address level (or village level at a minimum); geographical coordinates
Report description	Disaster type, sub-type, and sub-subtype (EM-DAT system)	Event, cause, and observations about both (uses popular terms)	Incident description, origin, and probable cause
Source	–	Documentary sources	Institutional sources
Effects			
People	Affected and killed (totals)	Affected, victims, evacuees, relocated, missing, wounded, and deaths (quantitative)	Affected, at risk, victims, assisted, evacuees, sheltered, missing, wounded, and deaths (quantitative)
Houses	–	Affected and destroyed (quantitative)	At risk; slightly, moderately, and severely damaged (quantitative)
Infrastructure	–	Roads; water and sewage systems; energy infrastructure; education, health, industry, and aid organization buildings (qualitative: whether or not there is damage)	Airports, roads, bridges, river and sea communications, water and sewage systems, communication infrastructure, energy infrastructure, schools, churches, and buildings (qualitative: whether or not there is damage)
Economy	Total damage	Crops and forests, livestock, total losses (qualitative: whether or not there is damage)	–

local level (i.e., village and town level) and the location information often includes geographical coordinates, although they are not always accurate. The SISMICEDE dataset of incidents caused by natural hazards from 2008 to 2011 has a total of 653 records in the municipalities partly or completely within the catchment.

The data entry templates of the three selected databases present the full set of information that the databases ideally contain (Table 1). Comparing the data entry templates reveals how the incidents are codified and identified in each database, indicating that the level of detail varies between databases. EM-DAT keeps records at a country scale, DesInventar at a municipal scale, and SISMICEDE at a local scale with geographical coordinates.

3.2 Data quality control

The quality of the selected data needed to be assessed to determine how to manage the available data for further spatial and temporal analyses.

3.2.1 Data quality control: method

Table 1 shows the full set of information that each database should contain. The completeness of impact and effect data ideally included in each dataset was checked by comparison with the identified templates.

The general attributes of the EM-DAT, DesInventar, and SISMICEDE data for the Samala River basin were evaluated to identify the advantages and disadvantages of each database for the purposes of this study. The results of this evaluation would help determine whether a single database would provide enough information for spatial and temporal analyses, or whether data compilation would instead be needed.

Finally, the completeness and accuracy of the location information in the datasets were reviewed by comparison with national geographical information. Entry locations found to be incomplete, deficient, or inaccurate were examined and rectified. National geographical data of departments, municipalities, villages, and schools were used as proxy data for the rectification, depending on the database scale. Unlocatable reports were removed from the datasets as they would not be informative for the study.

3.2.2 Data quality control results

The completeness evaluation of the effect data in the three selected databases (Table 2) indicated that the EM-DAT data are complete according to their specified fields. In contrast, much of the quantitative information required by the DesInventar template was not recorded or is instead given only as "fuzzy" qualitative data. The SISMICEDE database contains quantitative data on the effects of the recorded incidents on people and houses, but only qualitative information on infrastructure damage, which has not been consistently recorded over time. The SISMICEDE database lacks information about economic losses because such data are not included in the standard data entry template.

Tables 1 and 2 show that the available data were incomplete for some fields in the data entry templates. The attribute evaluation of EM-DAT, DesInventar, and SISMICEDE data for the study area (Table 3) indicates that all the datasets had advantages and disadvantages for local-scale spatial and temporal analyses, the analyses required for DRR applications and, more specifically, for spatial planning. The disadvantages of every database indicate

Table 2 Data completeness evaluation of EM-DAT, DesInventar, and SISMICEDE data for Samala River basin municipalities grouped by affected sectors; italicized cells indicate required data provided by the dataset, bold cells indicate data not provided, and bold italicized cells indicate data only partially provided

Affected sector	Database		
	EM-DAT	DesInventar	SISMICEDE
People	*Totals provided for most entries*	**Quantitative data required but many entries read only "yes"**	*Quantitative data provided*
Houses	Not included in the data entry template	**Quantitative data required but many entries read only "yes"**	*Quantitative data provided*
Infrastructure	Not included in the data entry template	*Qualitative data provided*	**Qualitative data required but not provided**
Economy	*Total damage provided for most entries*	***Qualitative data provided; quantitative data provided only for 3 % of entries***	Not included in the data entry template

that, as no single database would provide sufficient data quantity, detail, and time span to allow local-scale historical and trend analyses by itself, compilation was required.

The completeness of the location information was not evaluated for the EM-DAT data because the place descriptions correspond to departments or areas larger than a department; it was hence impossible to define particular locations for any data. All DesInventar location data were evaluated, but only to the municipal level because the sub-municipal location information lacks geographical coordinates and comprises only local names that could not be confidently located. Eighty-one percent of the SISMICEDE data had to be verified after evaluating the completeness of the location information (either the names or geographical coordinates of places); three percent of the data had to be removed because they were unlocatable. A total of 632 records were kept for this study.

3.3 Data compilation

3.3.1 Data compilation: setup method

According to the data quality control analyses, only two databases (i.e., DesInventar and SISMICEDE) are suitable for disaster analysis in the study area, although not individually. Both databases lack the qualities that would enable spatial and temporal analyses of disasters in the Samala River basin at a local scale by themselves. Compilation of data from both databases was therefore identified as a potential way to handle the available information. It was assumed that, to be able to compile data from different sources, both databases should meet certain basic criteria. First, the selected databases should capture the overall disaster situations in a similar way, despite their different sources, scales, and time spans. Second, the reporting systems should display consistent behavior, geographically and temporally. The method used to compile the various involved databases would depend on the behavior of each, which should be known over time. On the practical side, it was assumed that compiling data require compatible data that can be compared regardless of the particular limitations of the individual databases.

First, the compatibility of the databases was tested; the data entry systems and structures of the three databases were scrutinized and compared in order to identify differences and commonalities between their classification systems and the terms used for data entry. As

Table 3 Comparative analysis of the EM-DAT, DesInventar, and SISMICEDE data for the Samala River basin; the dataset attributes were evaluated to identify their advantages (+) and disadvantages (−) for local-scale DRR applications (i.e., for the aim of this study) (CRED 2009; CONRED 2010; DesInventar 2011)

Attribute	Database		
	EM-DAT	DesInventar	SISMICEDE
Database scale	Country level: (–) almost exclusively comprises only the biggest disasters in the country, which, in most cases, have captured international attention and occur infrequently	Town level: (+) because of the source, it includes incidents that attracted attention at the national level	Town/address level: (+) comprehensive detail, i.e., the reporting system allows for data corresponding to what has been important for national down to local-level managers
Number of entries for all the municipalities of the study area	25 (–) limited number of entries for the encompassed time span, too few to allow analyses by itself	356 (+) higher number of entries; local-scale incidents are included	653 (+) highest resolution of entries, corresponding specifically to local-scale incidents
Time span	1900–2011 (112 years) (+) long time span	1998–2011 (14 years) (+) moderate time span	2008–2011 (4 years) (–) limited; too short to allow temporal analyses by itself
Sources of information	Humanitarian, international aid, and governmental agencies: (–) only high-magnitude incidents are included	Newspapers mainly: (–) there might be a tendency to focus on incidents in the capital and biggest cities, or affecting the most important national infrastructure; incidents in less important areas could be underrepresented	Local and departmental delegates of CONRED, national agencies and organizations part of the CONRED system: (+) reliable sources as part of the national disaster reduction system
Spatial location descriptor	Not always available: described mostly in general terms, e.g., "southern coast"; (–) spatial mapping of individual records is impossible due to lack of specific location data	Name of municipality, sometimes local place names: (–) detailed spatial mapping is impossible due to lack of geographical coordinates	Name of municipality and address or geographical coordinates for most entries: (–) coordinates are sometimes wrong or correspond to the location of the nearest town, reducing data accuracy
Field entry characteristics	Predefined list of options for types; location names are typed in from the source information: (+) consistency in the classification system; (–) potential conflicts or inaccurate locations	Event and location (i.e., municipality) are selected from predefined option list; all other data fields are typed into the database: (–) potential conflicts or inaccuracy for most fields	All data fields are filled by typing in the information: (–) high potential for inaccuracy in all fields due to human error/typos
Entry language	English: (+) use and interpretation can be straightforward	Spanish: (–) potential interpretation problems due to local word use and translation	Spanish: (–) potential interpretation problems due to local word use and translation

the databases were created for different purposes, many dissimilarities were found and a new common set of classes was designed. The identifiers comprising the new common set of classes were chosen and classified based on the three databases in view of the natural hazards likely related to the origins of the various records. The new set of categories was used to reclassify all the data and allow comparative analyses among them.

The common set of classes was used to evaluate the quantitative compositions of the three datasets and to test whether they met the assumption that they depicted the same overall situations despite their differences. The three datasets were compared by (a) class, (b) origin of the underlying natural hazard, (c) month of occurrence, and (d) location of municipal territories for the available time span of each database (2008–2011). Percentages were used for these analyses to prevent the big differences in absolute numbers created by the different resolutions of the datasets from affecting the interpretation of the results. The analysis by origin of hazard was conducted to reduce the conflicts that could be created by the different ways of registering specific events, for example, the consequences of heavy rain.

The selection of the databases for the compiled spatial and temporal analyses was refined in view of the results of the quantitative composition analysis of the data comprising each database. Data displaying very dissimilar overall compositions and data with too many limitations for local applications due to their content, structure, or scale were disregarded.

3.3.2 Database compilation: setup results

The data entry templates and structures of the three databases were analyzed as part of setting up the appropriate data compilation indicated by the quality control results. The analysis of the data entry templates and structures revealed that the databases generally treat natural hazards related to the incidents reported, but that they manage the data in very different ways (Table 4). EM-DAT entries are classified using the terminology agreed on in 2007 by the Munich Reinsurance Company (Munich Re), CRED, Swiss Reinsurance Company (Swiss Re), the United Nations Development Programme (UNDP), the Asian Disaster Reduction Centre (ADRC), and the International Strategy for Disaster Reduction (ISDR) (CRED 2009; Munich Re 2011). This terminology establishes a structure that links the described disasters to the underlying natural hazards using four subsequent levels (Table 4a). Since natural hazards are causes of disasters but not the disasters themselves, the use of EM-DAT data could result in some uncertainties, as the terminology used might not be specific to the resulting disaster. A record of a storm, for example, might not clearly indicate what the consequences were.

DesInventar does not follow a scientifically based data entry structure, which is instead based on the most common terms used to describe disasters in the countries in the network from which the system was created (OSSO/La Red 2009). The use of classes according to a set of popularly used terms without consideration of the underlying natural processes may limit the usefulness of DesInventar data for scientific analyses.

SISMICEDE entries are classified at two levels, which are also related to the origin of the event. Entry details include the description of the incident, which is the effect produced at the local scale. All data fields are filled by typing in the information, which creates a high risk of errors such as misspelling or mistyping. In addition, the fields contain descriptions typed in by people who choose the wording themselves. Different entries describing similar incidents could be recorded using different words, so data verification

Table 4 Compatibility analysis of the data classification systems used by the EM-DAT, DesInventar, and SISMICEDE databases and the numbers of reports by class

Classes and number of reports included								Classes used in this study
(a) EM-DAT				(b) DesInventar	(c) SISMICEDE			
Origin	Type	Sub-type	Sub-subtype	Class	Origin	Event	Incident classes	
Geophysical	Earthquake	Ground shaking (3)		Earthquake (6)	Geological	Earthquake (10)		Earthquake
	Volcano	Volcanic eruption (5)		Volcanic activity (9)		Volcanic activity (1)		Volcanic activity
	Mass movement (dry)	Rockfall		Landslide: without rain as cause (22)		Mass movement	Landslide: no specific types and not associated with rain (1)	Landslide
		Landslide (1)	Mudslide					
			Lahar–debris flow	Sedimentation (2)				
		Subsidence	Sudden subsidence					
			Long-lasting subsidence					
Meteorological	Storm (7)	Tropical Storm		Electrical storm (1)	Hydrometeorological	Rainfall		Storm
		Local/convective storm	Thunderstorm					
			Generic storm	Hail (1)				
				Storm (8)				
				Rainfall (77)				
			Orographic storm (strong winds)	Gale (20)		Strong winds	Strong winds (8)	Strong winds

Table 4 continued

Classes and number of reports included								Classes used in this study
(a) EM-DAT				(b) DesInventar	(c) SISMICEDE			
Origin	Type	Sub-type	Sub-subtype	Class	Origin	Event	Incident classes	
Hydrological	Flood (7)	General river flood		Inundation (100)			Flood: no specific types (381)	Flood
		Flash flood		Torrential flow (37)				
		Storm surge/coastal flood						
	Mass movement (wet) (1)	Rockfall					Landslide: no specific type but associated with rainfall (101)	Debris flow
	Landslide	Debris flow		Landslide: with rain as cause (59)				
		Debris avalanche						
	Subsidence	Sudden subsidence					Sinking soil (11)	Subsidence
		Long-lasting subsidence						
							Structural collapse (95)	Water damage
Climatological	Extreme temperature	Cold wave (1)	Frost	Frost (14)		Cold front (24)		Extreme temperature
Total: 25 reports				Total: 356 reports	Total: 632 reports			

The classes of each dataset are organized in parallel and take account of the origin of the related natural hazards in order to establish how they relate to each other. The numbers in parentheses indicate the number of entries in each class of the corresponding database for the Samala River basin for the available time spans. The last column presents the set of classes used in this study to enable comparisons between the selected databases should be done, if possible. It should be noted that any required data verification is associated inherent uncertainty.

A new set of classes was developed to reclassify the data from the selected databases according to the identified differences between them (Table 4). The new set of classes follow the general data structure of EM-DAT, which classifies all the disasters according to the natural hazards that caused them (i.e., geological for disasters caused by geophysical processes and hydrometeorological for those caused by atmospheric and hydrological processes). It is important to note that further interpretation conflicts could arise from the noted differences in the classes used by the databases. For example, SISMICEDE considers rainfall a potential cause but not a class, whereas EM-DAT and DesInventar consider storm a class; consequently, a disaster caused by heavy rainfall would be recorded in the class "storm" in EM-DAT and DesInventar but could be recorded in one of several classes in SISMICEDE. In such cases, the comparison of similar reports between databases could be limited and the data interpretation could be affected.

The results of the quantitative composition analysis of the three datasets carried out to test how consistently they describe the disaster situations by (a) class, (b) origin of the natural hazard, (c) month of occurrence, and (d) location of municipal territories are shown in Fig. 2.

Figure 2a, b show that the percentage composition of DesInventar and SISMICEDE correspond fairly well to both natural hazard class and origin respectively. These composition analyses show that reports due to hydrometeorological causes, mainly of floods, are the most recurrent reports in both databases. EM-DAT data display similar maximums but, in contrast to the other databases, the predominance of records of hydrometeorological origin is less marked. These differences might be because EM-DAT's focus on large-scale disasters allows only mass events to be taken into account; therefore, the data for Guatemala are very few.

The composition analysis by month of occurrence (Fig. 2c) revealed similar behavior in all databases with a bimodal distribution over the year, the first peak occurring around May–June and the second around September–October. The percentages do not correspond exactly, but the denser months are clearly similar in all databases.

The composition analysis by municipality (Fig. 2d) could only be conducted for DesInventar and SISMICEDE. EM-DAT spatial resolution did not allow analysis of its data on a municipal basis. The analysis demonstrates that the higher frequencies of incidents recorded by both databases correspond to the municipalities of Quetzaltenango, Totonicapán, Retalhuleu, and Champerico of the 28 municipalities located partially or completely within the Samala River basin. The percentages differ between the two databases: The highest percentages in DesInventar occur in the most urban municipalities in the highlands, Quetzaltenango and Totonicapán, while the highest percentages in SISMICEDE occur in municipalities on the coastal plain. The differences could be created by the information sources: The newspapers might tend to report more on the cities, increasing the corresponding percentages for the DesInventar data. The actual spatial distribution of disasters in the Samala River basin needs to be further studied to identify causal processes that could guide DRR actions.

The composition analysis of the data demonstrates that reports of hydrometeorological origin are dominant in all three databases (Fig. 2a, b). This dominance over reports of geophysical and climatological origin is so great that further analyses presented here will deal only with reports with hydrometeorological causes. It was decided to exclude the quantitative impacts of the incidents caused by natural hazards from this study in view of the identified incompleteness of the relevant data in all the databases (Table 2). The

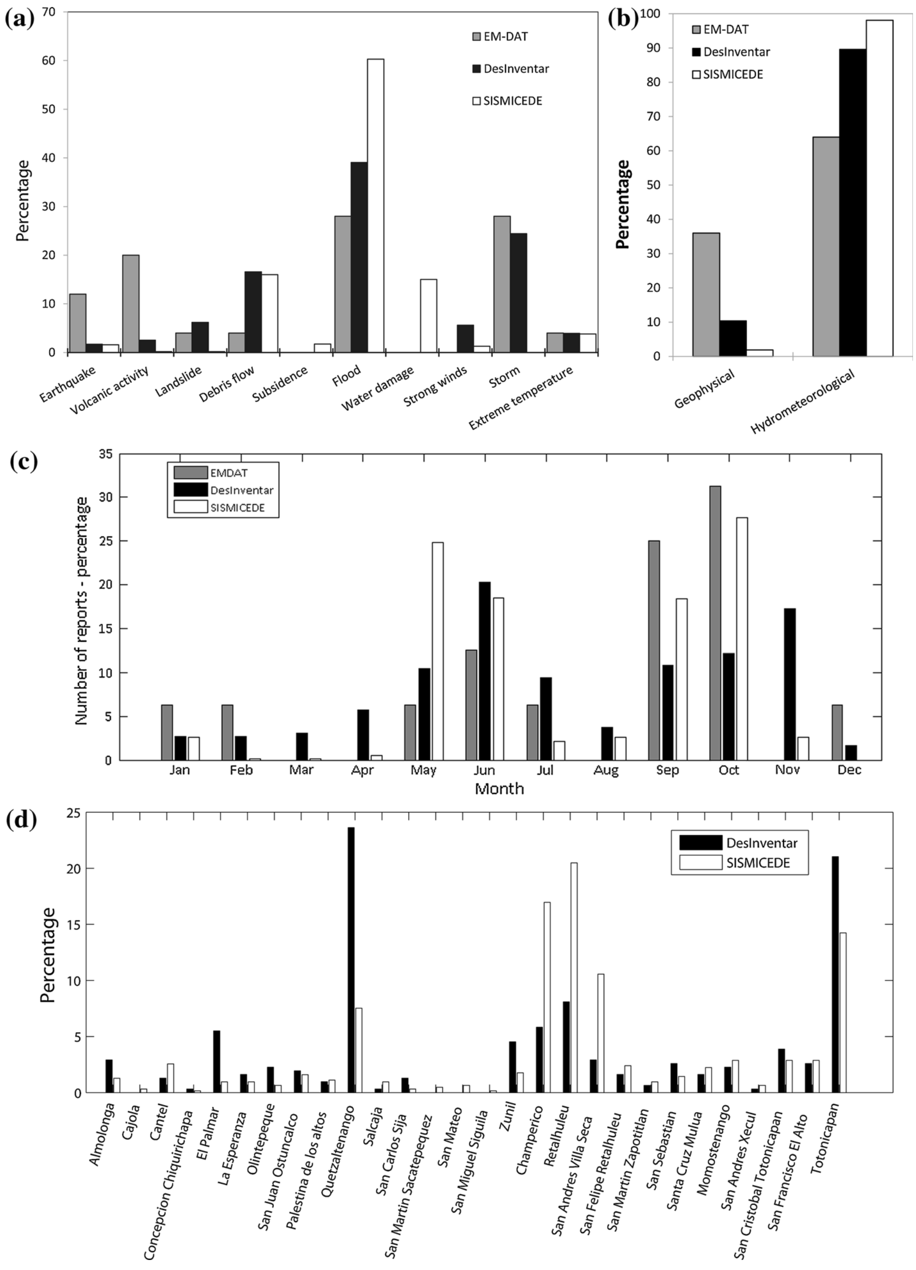

Fig. 2 Composition analysis of the database content. Bars indicate the percentage of the total number of records for the available time spans of the EM-DAT (1900–2010), DesInventar (1988–2011), and SISMICEDE (2008–2011) data for the Samala River basin municipalities by **a** class, **b** origin of the natural hazard, **c** month of occurrence, and **d** location of municipal territories

method designed in this work assumes that qualitatively analyzing the incidents in the study area provides enough information to understand, at least partially, the type of disasters in the study area in order to study their potential underlying causes. Because of the particular attributes of EM-DAT data and the resulting limitations of high-resolution studies identified in the previous analyses (Tables 1, 2, 3), the database was excluded from further analysis in this study.

3.4 Spatial and temporal analyses

3.4.1 Spatial and temporal analyses: method

A method was designed to compile the available data for the Samala River basin using the SISMICEDE spatial resolution and DesInventar time span (Fig. 3).

After the data assessment and quality control, the SISMICEDE database was selected for spatial referencing and the DesInventar database as the source of temporal data. SISMICEDE's fine resolution was used to spatially differentiate the reports referring to the Samala River basin, distinguishing among those recorded in all municipalities that lie partly within the basin. All the data were mapped using ArcGIS, and the basin boundary was used for the differentiation (Fig. 4); data referring to areas outside the basin were disregarded for this study. The retained SISMICEDE entries were aggregated by municipality as the smallest spatial unit in common with DesInventar. The number of reports per month was calculated for each municipality. The percentages that the monthly number of reports of the retained data represents in relation to the total number of reports by municipality were calculated.

The results of the quantitative analyses of SISMICEDE data were applied to DesInventar data. The percentages of the monthly number of reports of SISMICEDE data in relation to the municipal monthly total were extrapolated to select, proportionally, the number of reports in DesInventar in all the municipalities related to the study area. The resulting selected parts are assumed to correspond to the portion of data likely to refer to the study area. The method was tested by evaluating the relationship between the selected monthly municipal fractions of both databases obtained for the overlap period of both datasets (2008–2011). In view of the difference between the entries comprising both databases, the selected monthly municipal fractions were normalized for comparison.

Once the method was tested for the overlap period, a temporal analysis of reports in the Samala River basin was carried out for the entire DesInventar dataset time span (1988–2011). The resulting fractions were also analyzed seasonally, according to the hydrological year in Guatemala (i.e., November–October).

3.4.2 Spatial and temporal analyses: results

Spatial analyses of SISMICEDE data from 2008 to 2011 indicated that approximately 43 % of the total number of reports in the municipalities related to the Samala River basin occurred within the basin (Fig. 4). The numbers and percentages of reports referring to the basin from SISMICEDE and from DesInventar that resulted from the compilation process corresponded moderately when compared for the overlap period (Fig. 5).

Similarities have been identified between the temporal patterns of reports registered by DesInventar and SISMICEDE. This applies to both the yearly and seasonal analysis; the dry season is considered to run from May to October and the rainy season from November to April.

The temporal analysis of reports for the Samala River basin for the DesInventar data, as extrapolated from the SISMICEDE location data (Fig. 6), examined the data from 1988 to 2011 both yearly and seasonally. Only the temporal analysis of reports per rainy season

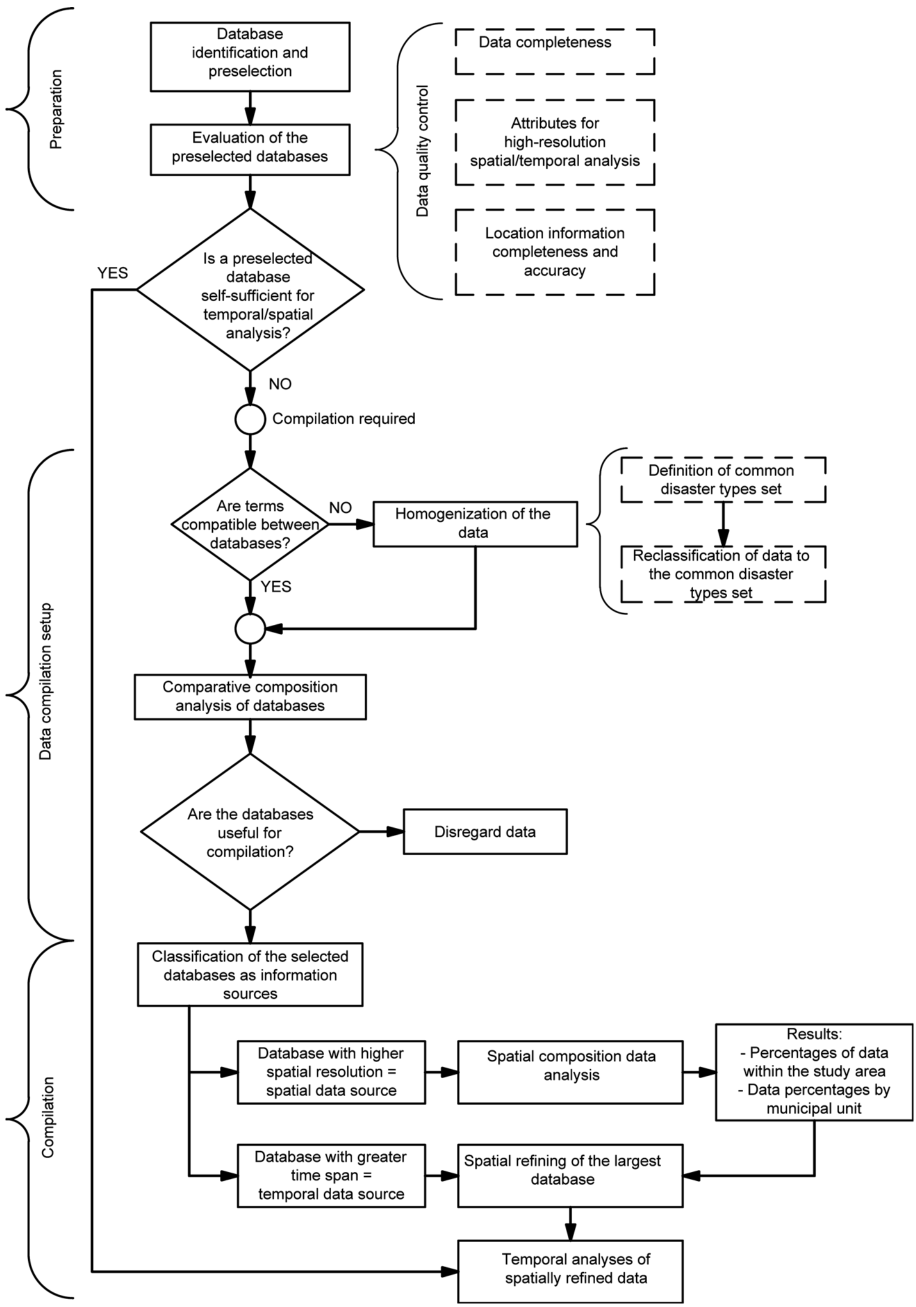

Fig. 3 Process flowchart of the compilation method used in this study

indicates a statistically significant increasing trend at a 95 % confidence level ($p = 0.004$), which could have been influenced by improved reporting over time. One significant outlier appears during the 1998–1999 dry season. This outlier was caused by hurricane Mitch, which struck Guatemala from 28 October to 2 November 1998; most of the reports in the

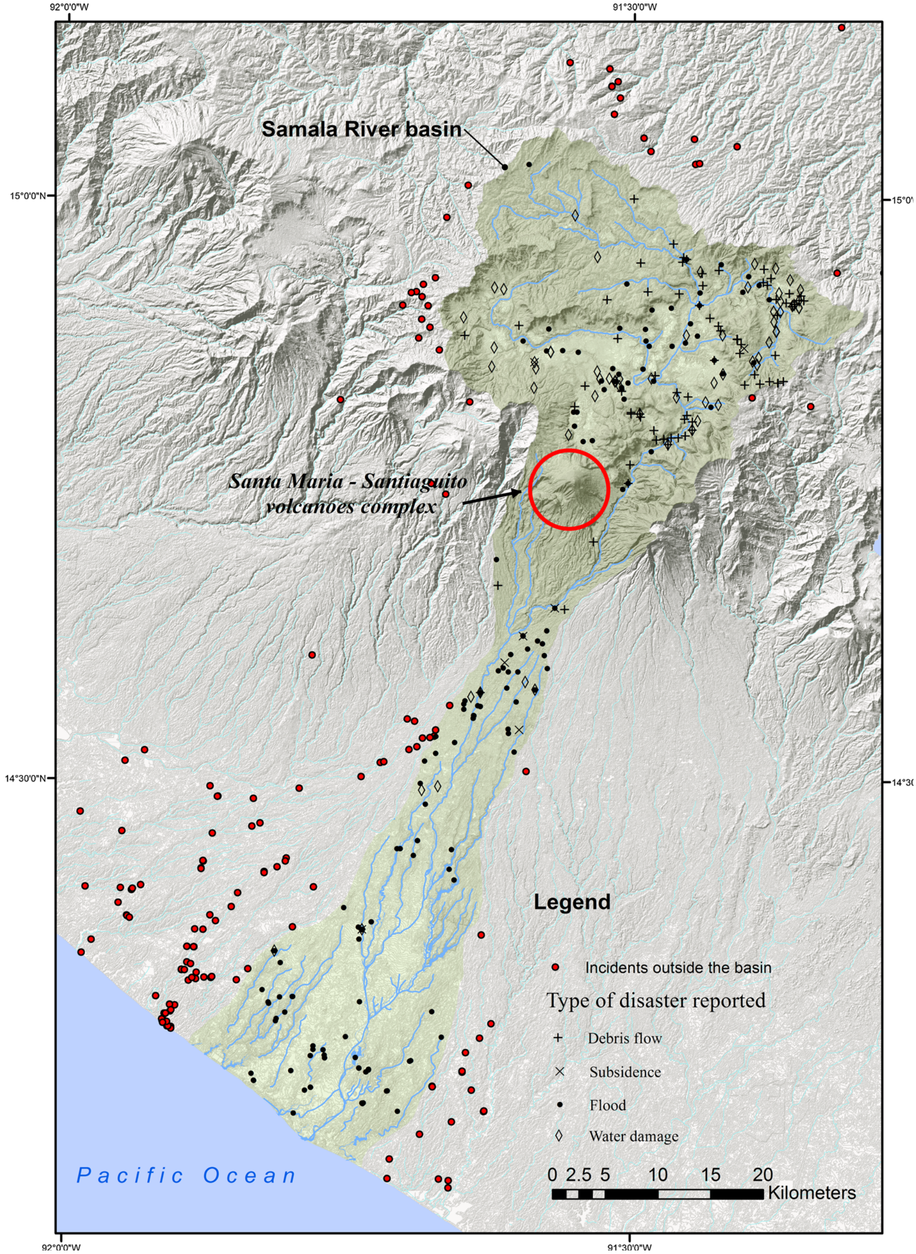

Fig. 4 SISMICEDE data, 2008–2011, for the municipalities completely or partially within the Samala River basin; geographical information source: SEGEPLAN, 2012

study area correspond to the first days of November that generally mark the beginning of the dry season but in 1998 marked the late ending of the rainy season.

Rainfall in the Samala River basin was analyzed over time on yearly basis to determine whether there are distinguishable patterns that could be related to disasters since rainfall is

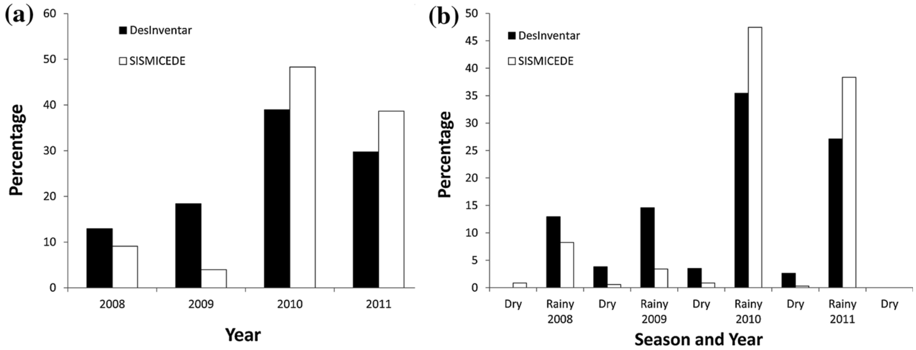

Fig. 5 Comparison of the normalized number of records of hydrometeorological origin in the DesInventar and SISMICEDE databases for the Samala River basin, 2008–2011. Resulting quantitative data for the Samala River basin from the municipal totals result from the compilation method developed here and are shown **a** by year and **b** by rainy (November–April) and dry (May–October) seasons

the main potential underlying natural hazard of disasters in the study area. The analysis was made using the total precipitation per hydrological year, the maximum monthly precipitation per hydrological year, and the maximum daily precipitation per hydrological year. The results showed no particular trend over time in any of the cases. The comparison of rainfall and frequencies of disasters per hydrological year from 1988 to 2011 showed no relationship between the two variables.

4 Discussion

Fragmented data are commonly encountered in disaster research in many places around the world. However, research at the local or regional scale focusing on the disaster risks associated with local conditions is greatly needed, especially in disaster-prone areas. Available databases for the study area were searched, identified, collected, and evaluated for quality and usefulness for spatial and temporal analyses. There were few applicable databases for Guatemala and the Samala River basin, and the identified information was found to be inadequate, mainly in terms of spatial resolution and time span. Since the ultimate purpose of the analyses presented here is to support spatial planning, the scale needed is eminently local and the spatial and temporal resolution should be high. It is important to consider the current state of the art of disaster data collection and that, as Downton and Pielke (2005) have concluded, the results of local-scale disaster studies tend to be less accurate than those of country-scale studies, probably because the available data are usually coarser than are needed for local studies. Global databases are greatly needed, but their use for spatial planning is limited because the data resolution is too general to provide useful information at the local scale.

Data verification for the available databases for the study area revealed that data mis-registration (e.g., misspelling of local place names) might be a systematic source of error in all the datasets. As a result, some records could be assigned to wrong places or be completely unlocatable. For spatial applications, lack of specific geographical location could hinder spatial analyses or make them highly uncertain or inaccurate. The typing process for data entry is a frequent source of error as well as a factor potentially hindering automated data processing.

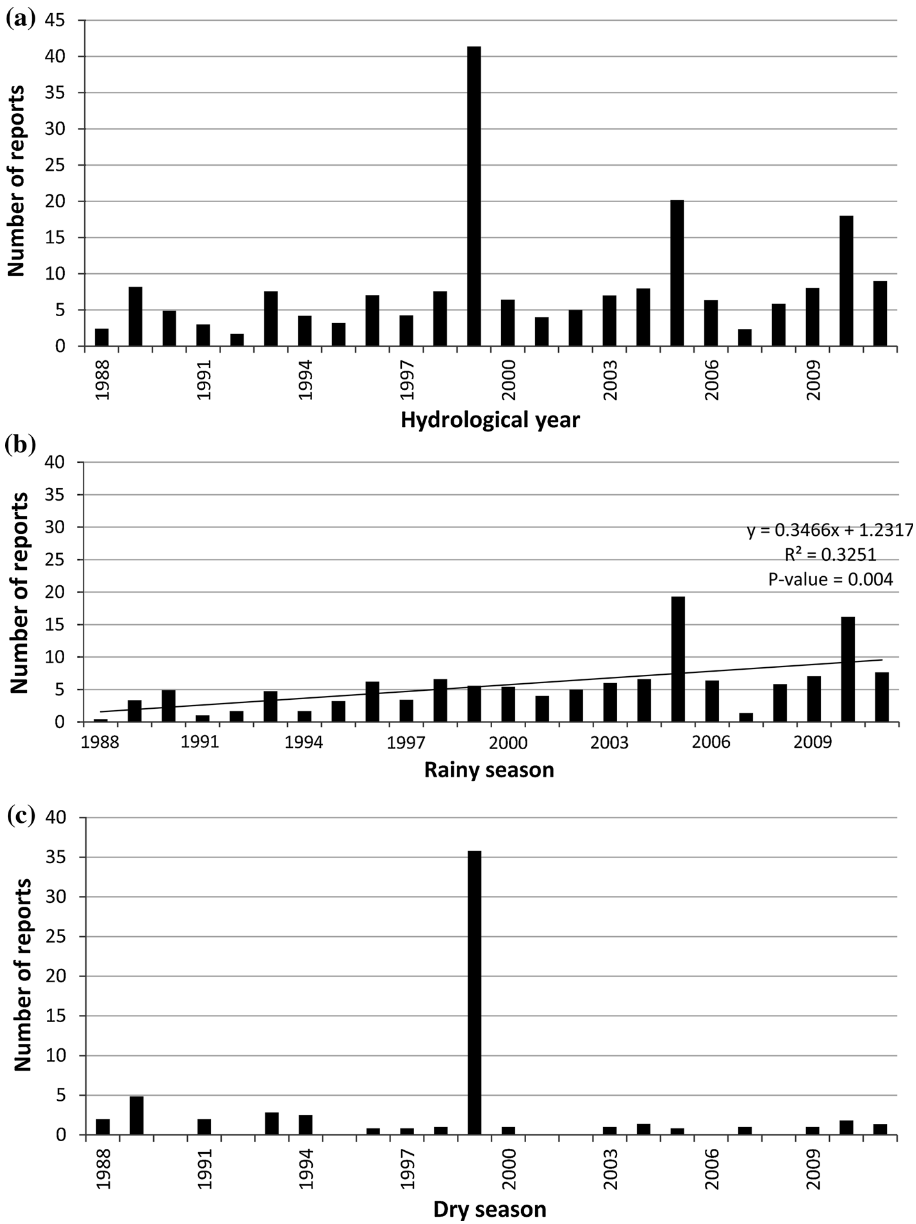

Fig. 6 Temporal analysis of DesInventar data of hydrometeorological origin, 1988–2011. The numbers of records by year are shown **a** by hydrological year (1 November–30 April), **b** by dry season, and **c** by rainy season. The dry season corresponds to November–April, the rainy season to May–October; a year corresponds to one dry season and one rainy season

Effect analysis was impossible in this study due to the lack of data on impacts; the collection of such data should be implemented and improved to enable better understanding of the disaster situation and further action to address it.

The results of this study regarding the limitations of available data are consistent with the conclusions of Amin et al. (2008) who found several frequent data problems, including inaccurate names and locations of communities (especially small remote ones) and disagreement in entry definitions and categories. Even given the potential uncertainties generated by these limitations and the lack of high-quality data for the studied area and country, the present kind of study is needed to improve our understanding of disasters, to support risk-mitigation actions, and in turn to improve the quality of data for future studies.

The results of the spatial analysis of SISMICEDE data from 2008 to 2011 indicated that only approximately 43 % of all the reports in the municipalities completely or partially in the Samala River basin occurred within the physical limits of the basin. This finding highlights the discrepancies between administrative and geomorphological boundaries, which indicate the importance of the spatial resolution of the input data. Studies at a subnational scale are needed to support spatial planning. The spatial resolution of the input data plays a crucial role in determining the kinds of studies that can be carried out and the accuracy of the results/interpretations of these studies. Data collected at a municipal scale provide information that is geographically too coarse and impede mapping the spatial distribution of disasters and relating this distribution to specific physical or social features of the territory. Local-scale studies that analyze data incorporating location indicated by geographical coordinates are needed.

The difference between the basin surface area and the surface area of the municipalities belonging partially or completely to the basin indicates that analyses focusing on a particular geographical area would provide significantly different results depending on the geographical unit of the data. Administrative boundaries frequently differ from geomorphological boundaries. Analyses using data with municipality as the spatial unit would likely link disasters occurring within the same administrative boundaries but that result from similar although independent natural processes.

The spatial analyses also demonstrated that approximately half of the floods reported in the municipalities related to the study area occurred outside the Samala River basin, while most debris flows, subsidence, and water damage occurred within its limits (Fig. 4). These differences could be the result of diverse geomorphological features or social circumstances over the territories of the Samala River basin municipalities; this diversity would create different disaster risks in different places, variations not captured when using municipal-scale data. However, as these differences seem important for the spatial distribution of disasters, local-scale research into the potential relationships between disasters and the natural and social conditions over the study area is required.

The composition analysis of the databases by month of occurrence indicated a bimodal distribution with peaks in May–June and September–October. This behavior is probably due to the predominance of reports of hydrological origin and related to the hydrological seasonality in the area, a relationship that should be further studied in future work. The analysis of the main natural hazard in the area (i.e., rainfall) over time showed a temporal pattern that corresponds with the pattern of disasters; however, clear numerical relationships were not identified between the two variables which indicate that other variables could be also causes for the disasters.

The most frequent class of reports in the study area comprises those of hydrometeorological origin; in contrast, those of geophysical origin represent 10 % or less of the total number of reports in the DesInventar and SISMICEDE databases. The most frequently reported classes in those databases are floods and debris flows; the DesInventar *storm* category, which could likely include floods and debris flows as well, constitutes 24 % of the database total. It is important to assess the situation regarding disasters in particular places, because risk-mitigation alternatives vary according to the kind of hazard: Some hazards require, for example, measures addressing the built environment, while others require non-structural measures, such as improving public awareness or training (Hyndman and Hyndman 2009). Disasters of hydrometeorological origin require measures that specifically involve spatial planning. The requirement for spatial planning measures as well as the predominance of hydrometeorological disasters in the Samala River basin leads this study to focus on hydrometeorological disasters. This focus is supported by international statistics indicating that, in Guatemala, most disaster-related deaths and economic damage are caused by storms, floods, and mudslides (PreventionWeb 2012).

The outcomes of this study indicate that limited data restrict the opportunities for mining spatial, temporal, and impact data on disasters that can be used in DRR planning in areas such as the Samala River basin. The temporal and spatial dimensions of the available data were limited and compilation proved to be the best identified alternative for investigating disasters in the study area. The conclusions that could be drawn from the analyses were limited due to the consequences described above; however, the results indicate that the available data can provide a valuable information pool that allows better-informed conclusions regarding the class, location, and temporal progress of disasters in the study area. At the same time, the results indicate that, despite their identified limitations, the available data are a useful source of information on the kind of disasters that take a toll on the studied area.

5 Recommendations

As Amin et al. (2008) have pointed out, emergencies tend to relegate the accuracy and completeness of data collection to a secondary place after the emergency response. The results of such priorities are decision-making systems that operate on an ad hoc basis instead based on scientific evidence. It is highly recommended that in Guatemala, CONRED, in its role of national coordinator, should promote a change on this practice in favor of a system that gathers and maintains high-quality data, with accurate and complete information describing the disasters, their locations, and their effects. Efforts are already being made in this regard, but the goal remains to be achieved. An improved disaster information system would provide the information needed during the emergency response, regarding what is needed, by whom and where, while at the same time providing data input for post-disaster analyses. Informative disaster data could enable better-informed DRR decision making and planning. In the long run, improving the disaster information system could promote recognition of the important role a disaster database can play in developing the entire country.

People in charge of reporting disasters need to have a clear understanding of what should be reported in the database and how this should be done in order to maintain database accuracy and quality. A basic training program should be created for the people working within the CONRED structure at the local to national levels, as well as guidelines for entering values in the database.

If national and especially local authorities and institutions viewed natural hazards as a shared concern and their hazard-related decisions as strongly interconnected, joint measures would be easier to visualize and implement. An important result of such a change of vision would be integrated management plans for territories lying within a common hydrological catchment, as they would face common hazardous processes due to the common underlying natural processes.

CONRED was created recently, after the previous national emergency agencies. The freely available data cover the period of the agency's existence; data from before that time are kept in archives but are cautiously released, if required. Although such caution is valid, longitudinal disaster data are greatly needed, and if any such data exist, they should be made known and studied. Research and mitigation purposes require a database as complete as possible that includes data from all existing sources. As such, old Guatemalan data should be digitized and made available but, as Hittelman et al. (2001) recommend, all errors and error sources, known and potential, should be carefully described to facilitate better interpretation.

It is true that changing and improving existing systems implies substantial investments but, as Amin et al. (2008) have found, advance investments in natural disaster data are highly effective because baseline data are crucial for disaster management. Bearing in mind the existing data limitations in Guatemala and the cost that overcoming such limitations would impose on a system that must simultaneously deal with many other development problems, alternative plans should be considered. Crowdsourcing, an alternative way of collecting data from a wider network of people and using more open platforms, could provide a feasible alternative to reduce data limitations. The resources needed to establish and maintain an online platform for local data collection and the later data processing are likely less than the resources required to operate a network of employees handling same amount of information. Great quality control would be required, but the potential for increased understanding of disasters would be worth the effort.

Acknowledgments The author thanks Veijo Pohjola, Allan Rodhe, and Jan Boelhouwers for the scientific guidance and support. Special thanks are given to CONRED in Guatemala for the opportunity to get to know their functioning and providing the data for this work. I thank EM-DAT and DesInventar for providing the requested data for this study. This research was funded by a stipend for PhD studies of the Swedish International Development Agency.

References

Adams W (1986) Whose lives count: TV coverage of natural disasters. J Communication 36(2):113–122

ADRC (2009) GLIDEnumber. In. Asian Disaster Reduction Center, Japan

Alcantara-Ayala I (2002) Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries. Geomorphol 47(2–4):107–124

Alexander, D. E. (1993) Nat. disasters. Springer

Alfaro MEJ, Quesada AR (2010) Ocurrencia de ciclones tropicales en el Mar Caribe y sus impactos sobre Centroamérica. InterSedes XI(22):136–153

Amin, S., & Cox, M. et al. (2008) Using data against disasters: overview and synthesis of lessons learned. In: Data against natural disasters: establishing effective systems for relief, recovery and reconstruction. The World Bank, Washington, DC. 1–22

Burby RJ, Dalton LC (1994) Plans can matter! the role of land use plans and state planning mandates in limiting the development of hazardous areas. Public Adm Rev 54(3):229–238

CONRED (2010) Coordinadora Nacional para la Reducción de Desastres. In, Guatemala

CRED (2009) EM-DAT: The OFDA/CRED International Disaster Database. In. Université Catholique de Louvain, Brussels, Belgium

DesInventar (2011) DesInventar Disaster Information Management System. In. DesInventar Open Source Development Team

Downton M, Pielke R (2005) How accurate are disaster loss data? The case of US flood damage. Nat Hazards 35(2):211–228

Ganapathy PG (2011) DesInventar database: disaster management plan for Vellore District, Tamil Nadu State, India. Int J Geomat Geosci 2(1):282–288

GFDRR (2009) Disaster risk management programs for priority countries. In. Global Facility for Disaster Reduction and Recovery, The World Bank, UN International Strategy for Disaster Reduction. 295

Guha-Sapir D, Below R (2002) The quality and accuracy of disaster data: a comparative analyses of three global data sets. World Bank, Disaster Management Facility, ProVention Consortium, Brussels
Harmeling S, Eckstein D (2012) Global climate risk index 2013. In. Germanwatch e.V., Berlin
Hittelman A, Lockridge P et al (2001) Interpretive pitfalls in historical hazards data. Nat Hazards 23(2–3):315–338
Hyndman D, Hyndman D (2009) Natural hazards and disasters. Brooks/Cole, Australia
ISDR (2009) Global assessment report on disaster risk reduction. United Nations, Geneva
Kuenzi D, Horst OH et al (1979) Effect of volcanic activity on fluvial–deltaic sedimentation in a modern arc-trench gap, southwestern Guatemala. Geol Soc Am Bull 90(9):827–838
Munich Re (2011) Hierarchy and terminology of natural hazards. In. Geo Risk Research, NatCatSERVICE, Munich
OAS (1991) Primer on Natural Hazard Management in Integrated Regional Development Planning. Dept. of Regional Development and Environment, Executive Secretariat for Economic and Social Affairs, Organization of American States, Washington, DC
OSSO/La Red (2009) DesInventar Disaster Inventory System Methodological Guide, version 8.1.9. In. P. DesInventar. Grupo de Investigación OSSO and La Red, Colombia
PreventionWeb (2012) Guatemala: Disaster Statistics. In. U.N. Int Strategy for Disaster Reduction ISDR
Serje, J. ([date unknown]) Desinventar: a methodology to build Disaster Inventories as part of the Risk Mitigation Process. In. UNDP/LA RED. 20
Sperling F, Szekely F (2005) Disaster risk management in a changing climate. In: Discussion paper
Tesliuc ED, Lindert K (2004) Risk and vulnerability in guatemala: a quantitative and qualitative assessment. In: Social protection discussion paper series. H. D. N. Social Protection Unit. The World Bank. 92
UN/ISDR (2004) Living with risk. United Nations, New York, Geneva
University of Richmond (2002–2006) The disaster database project. In. W. G. Green III