Nat. Hazards Earth Syst. Sci., 14, 189–208, 2014
www.nat-hazards-earth-syst-sci.net/14/189/2014/
doi:10.5194/nhess-14-189-2014

Natural Hazards and Earth System Sciences
Open Access

A quality assessment framework for natural hazard event documentation: application to trans-basin flood reports in Germany

S. Uhlemann[1,2,*], **A. H. Thieken**[2], **and B. Merz**[1]

[1]Helmholtz Centre Potsdam – GFZ German Research Centre for Geosciences, Section Hydrology, Potsdam, Germany
[2]University of Potsdam, Institute of Earth and Environmental Sciences, Potsdam, Germany
[*]now at: ASPEN Re, Aspen Insurance UK Limited, Research and Development, Zurich, Switzerland

Correspondence to: S. Uhlemann (steffi.uhlemann@aspen-re.com)

Received: 3 January 2013 – Published in Nat. Hazards Earth Syst. Sci. Discuss.: 7 February 2013
Revised: 16 December 2013 – Accepted: 18 December 2013 – Published: 4 February 2014

Abstract. Written sources that aim at documenting and analysing a particular natural hazard event in the recent past are published at vast majority as grey literature (e.g. as technical reports) and therefore outside of the scholarly publication routes. In consequence, the application of event-specific documentation in natural hazard research has been constrained by barriers in accessibility, concerns of credibility towards these sources and by limited awareness of their content and its usefulness for research questions. In this study we address the concerns of credibility for the first time and present a quality assessment framework for written sources from a user's perspective, i.e. we assess the documents' fitness for use to enhance the understanding of trans-basin floods in Germany in the period 1952–2002. The framework is designed to be generally applicable for any natural hazard event documentation and assesses the quality of a document, addressing accessibility as well as representational, contextual, and intrinsic dimensions of quality. We introduce an ordinal scaling scheme to grade the quality in the individual quality dimensions and the Pedigree score which serves as a measure for the overall document quality. We present results of an application of the framework to a set of 133 cases of event-specific documentation relevant for understanding trans-basin floods in Germany. Our results show that the majority of flood event-specific reports are of good quality, i.e. they are well enough drafted, largely accurate and objective, and contain a substantial amount of information on the sources, pathways and receptors/consequences of the floods. The validation of our results against assessments of two independent peers confirms the objectivity and transparency of the quality assessment framework. Using an example flood event that occurred in October/November 1998 we demonstrate how the information from multiple reports can be synthesised.

1 Introduction

The role of past and present natural hazard events as learning examples for hazard assessment, risk prevention and disaster mitigation has been stressed at many instances (Hübl et al., 2002; IRDR, 2011). The underlying rationale is that any predictive method (model) and the effectiveness of disaster response likewise depend on observations and monitoring as well as on experience from real-life situations (Hübl et al., 2002). Therefore, any systematic event analysis requires the evaluation of all available sources of data and information on past events as well as the standardized documentation of any future event.

Written sources that aim at documenting and analysing a particular natural hazard event in the recent past are published at vast majority as grey literature (i.e. literature not controlled by commercial publishers) and therefore outside of the scholarly publication routes (Uhlemann et al., 2013). Anecdotal evidence shows that many scientists are aware of the existence of grey literature (to a certain degree), i.e. they know about the publishing activities of state authorities or (inter-)governmental institutions in the course of natural disasters. However, the type and detail of information contained in the documents seem to be rather unclear, making it difficult to assess the potential applicability for one's own research question.

Studies investigating the use and influence of grey literature in science and research synthesis (e.g. MacDonald et al., 2010; Rothstein and Hopewell, 2009; Uhlemann et al., 2013) highlight two main reasons that hamper the effective use of the material in scientific discourse. First, practical aspects that can be largely described as barriers to accessibility (both in terms of the languages used and the options for finding these documents) impose strong constraints on the applicability of these items. Second, the credibility of information published in grey literature is commonly perceived as low. The assumption is that grey literature is never or rarely impartially refereed (MacDonald et al., 2010) and therefore scientific quality standards (peer reviewing and editorial control) are not ensured.

In contrast to the common perceptions of quality constraints, studies that have investigated the characteristics of grey literature (Ranger, 2004; Weintraub, 2000; Farace and Schöpfel, 2010) often highlight that these documents contain unique and significant scientific and technical information, and that the greatest advantage of grey documents is the considerably greater detail at which a topic can be treated. From their analysis of publications accessible on floods in Germany in the recent past, Uhlemann et al. (2013) conclude that ignoring grey sources in flood research also means ignoring the largest part of knowledge available on single flood events (in Germany).

However, in order to be used in scholarly argument, the issue of credibility of event documentation needs to be addressed, i.e. requires a quality assessment. Only then will or should it be included in the knowledge accumulation or in a decision-making process.

Defining the quality of publications or any other documentary evidence is in itself a challenging task, as no agreed standard measure has been proposed. Hjørland (2012) provides a concise list of methods for evaluating information sources as they are applied in the field of information sciences. In total they discuss 12 widely used concepts, most of which originate from critical research assessment (and are therefore primarily intended for "white" literature) and address quality through the classical peer review, author credentials, publisher reputation or journal impact factors. Inasmuch as these approaches are current gold standards in research assessment (Bailin and Grafstein, 2010), they are indirect methods for evaluating the quality of a document (or its academic worth), and as such have received substantial criticism (e.g. Seglen, 1997; Simons, 2008). Broader concepts that acknowledge the heterogeneity of information sources are checklist approaches (most common for evaluating websites, e.g. Hjørland, 2011), comparative studies (evaluating a study against "authoritative works" in the field, e.g. Bragues, 2009), or evidence-based evaluations.

Evidence-based evaluations aim at synthesizing the available evidence for a given question (e.g. how effective interventions in a river system are for habitat restoration of species x) to identify and assess consistent findings across diverse studies and to inform policy (Burton, 2010; Borenstein et al., 2009). They are most commonly applied in the course of systematic reviews and meta-analyses and have become standard in the health and medical sciences (Higgins and Green, 2011), and have also been transferred to environmental science and management (Centre for Evidence-Based Conservation, 2010; Norris et al., 2008; Osenberg et al., 1999). Beside the quantitative meta-analyses that provide a reproducible weighted average of the estimate of an effect, qualitative criteria-based methods of causal inference have been developed (see Weed, 2000 for a comparison of both methods).

Documentation and analysis of natural disasters primarily aim at describing the event in its course and therefore they do not address classical evidence-based objectives like cause–effect relationships under a particular experimental design. For this type of reporting evidence-based approaches are less suited to characterise the quality of the study. A special case of evidence-based evaluations are the techniques developed in the field of historic sciences for source criticism. When studying extreme events of the past, information about the climate or floods in the pre-instrumental period is commonly retrieved from documentary evidence that therefore constitutes the data basis of historic climatology or historic hydrology (see Brázdil et al. (2006) for a concise overview). In order to make use of historic sources such as chronicles, annals etc., methods have been developed to critically analyse the evidence of the documents. Rather than defining quality measures of a document, the sources are checked critically for the conditions under which the information of concern was produced. Glaser and Stangl (2003) as well as Glaser (2001) present a scheme for source criticism typically applied in historic climatology. In order to assess the reliability of the source intrinsic quality, characteristics like the bibliographical level of the author, his mental perceptions and the environmental level (depicting the general way of thinking and expression during an epoch) are used as indicators and the content of the document is cross-checked for compliance with known historical and scientific facts of that time. Only if some basic agreement is reached is the climate information extracted and used for further quantitative analysis (Glaser and Stangl, 2003).

In order to assess the credibility and applicability of recent sources of event documentation and analyses, the quality assessment needs to be accompanied by a contextual assessment, i.e. an assessment of the quality of information provided on the geophysical processes that caused the hazard and the resulting consequences. Furthermore, from a scientific perspective it is important that the information is accessible and of an intelligible nature. In summary, it is important to assess how much and how qualitatively good the information of the document is, and how likely it is that this information is actually used.

Therefore, a holistic assessment of the quality of written sources is needed in order to assess their applicability or,

in other words, their 'fitness for use'. This user's point of view is pursued in the field of information systems sciences and its linkages to organizations, management and consumer research, where the quality of a product (data, information) plays a critical role. It is well accepted in this field of research that the quality of data and information cannot be assessed independently of the purpose for which they are being used (Strong et al., 1997). For more than two decades, the development of methods and standards for data and information quality assessment has received substantial interest (see Madnick et al., 2009 for an overview). Using a factor analysis on a large set of data quality attributes, Wang and Strong (1996) identified four distinct data quality categories, all of which need to be addressed for a complete assessment of quality: accessibility, and contextual, representational, and intrinsic quality. Each of the categories is further defined by certain quality dimensions (QD). Synonymous with Wang and Strong's (1996) formulation of data quality, we can summarize the expectation towards a high-quality report on a natural hazard event as that it should be intrinsically good, contextually appropriate for the task at hand, clearly represented and accessible to the information consumer.

Considering the identified strengths and weaknesses of event-specific documentation (that is largely characterised by grey literature features) and the apparent barriers to using these documents as another source of data and information in flood risk research and management, the objective of this study is the development of a quality assessment framework for written contemporary sources. The aim is to provide a generic framework for information quality assessment and quality labelling of natural hazard event reports (including both grey and white literature) for an international research question. Using the example of trans-basin floods in Germany in the period 1952–2002 (Uhlemann et al., 2010), we want to assess the quality of event-specific reports that were produced for any of the top 40 trans-basin floods. Providing a holistic quality measure, we want to address the concerns of credibility of sources and of their contextual depth. In that we aim to improve the use and awareness of the information contained in flood event-specific reports.

The paper is structured as follows. In Sect. 2, first, we introduce the data used for this study and, second, present the quality assessment framework and the test for concordance to infer its validity for judging the quality of event reports. Results and discussion are presented in Sect. 3 starting with the concordance check and a discussion of the quality assessment framework. Following, we present the results of the overall quality of the entire material. We complete our results with an application to an example flood event. Finally, in Sect. 4 we conclude on the framework and the quality of the reports with respect to their applicability for flood research. The paper is accompanied by an open access data supplement (Uhlemann, 2013).

2 Data and methods

2.1 Data

We use a subset of the literature compiled in Uhlemann et al. (2013). The set comprises the results of a systematic search for publications that contain information on the sources, pathways, receptors and/or consequences (SPRC) for any of the top 40 trans-basin flood events in Germany in the period between 1952 and 2002 (Uhlemann et al., 2010). The main criteria for the search can be summarized as follows. Only reports are included that treat any of the selected flood events within Germany and that are of a regional scope or broader. Local studies and studies that document the flood largely outside the German territory are only considered if they also provide information relevant for the scale of trans-basin floods in Germany. The search included solely print material (both paper and e-prints) and website contents of either scientific or agency origin. This excludes reports from other internet sources or media. Furthermore, (personal) experience reports or narratives were not included. The main tools used for the search were (1) the Web of Knowledge, (2) the Karlsruhe Virtual Catalogue (KVK), which is the standard search tool for publications indexed in public open access library catalogues in Germany, and (3) open catalogues of federal or state agency libraries or their respective homepages. A detailed description of the search criteria applied for the systematic search is provided in Uhlemann et al. (2013).

The search resulted in the identification of 186 relevant publications (see Uhlemann et al., 2013 and the respective data supplement provided in Uhlemann, 2012). For the purpose of this study we use only reports that explicitly aim at documenting one or a number of flood events. In Uhlemann et al. (2013) these types of reports are classified as "Special Report 1" (reports on one, possibly two particular flood events aiming at documentation and analysis) or "Special Report 2" (reports on two to five, rarely more events, sometimes with the aim of comparative analysis but generally aiming at an event description). In total 105 reports of this characteristic are listed in the set of documents. As some reports contain documentation on more than one trans-basin flood event, the total number of event-specific documents sums to 133.

2.2 Quality Assessment Framework (QAF)

In their analysis of the importance of quality attributes for consumers of data, Wang and Strong (1996) identified four distinct quality categories (QC), all of which need to be addressed for a complete assessment of data quality: accessibility, and contextual, representational, and intrinsic quality. They stratify each of the categories by a number of quality dimensions (QD) and consequently develop their framework for data quality from the perspective of the needs of the data user. This fitness for use is considered through the

formulation of a task at hand. Synonymous to Wang and Strong's (1996) and the Strong et al. (1997) formulation of data quality, we can summarize the expectation towards a high-quality report on a natural hazard event as that it should be intrinsically good, contextually appropriate for the task at hand, clearly represented and accessible to the information consumer. For the development of our quality assessment framework for natural hazard event documentation we adapt the number and the definitions of the quality dimensions considered per quality category in the Wang and Strong (1996) concept to fit the purpose of assessing the quality of event documentation instead of data. In the case of our study the task at hand is formulated from a scientific user's perspective and phrased as: *What were the governing hydrometeorological and hydraulic processes that have caused the trans-basin flood event of year x in Germany and what were its consequences?*

Figure 1 provides an overview of the framework and the quality attributes considered. The four quality categories can be differentiated into those that capture quality aspects specific to the document (accessibility, and representational and intrinsic quality) and those that capture the quality of the document with respect to the information provided on the natural hazard event (contextual quality). The following conventions are made for the application of the quality assessment framework on flood event reports: (1) if a report covers more than one event it will be evaluated separately for each event in the contextual quality dimensions; (2) each report is assessed with respect to its own spatial focus (and not with respect to the entire flood-affected region); (3) within the document-specific quality aspects each report is assessed with respect to its own objectives (i.e. a report that focuses on the meteorological aspects or on damages will be assessed on these aspects and not towards an expectation of completeness in the event description). The rationale for the choice and definition of quality dimensions is given in Subsects. 2.2.1 to 2.2.4 of this section (as indicated in Fig. 1). Tables A1 to A3 in the Appendix then provide the basis for the quality assessments of any of the 133 event-specific documents on trans-basin floods

In order to quantify the level of information within each of the individual quality dimensions and to quantify the overall report quality we extend the concept of Wang and Strong (1996) by a pedigree scoring scheme. A pedigree scoring scheme was developed by Funtowicz and Ravetz (1990) and Van der Sluijs et al. (2005) in order to assess the uncertainty and quality of environmental research that is relevant for policymakers. Selected data quality dimensions of Wang and Strong (1996) and a pedigree scoring scheme based on the work of Van der Sluijs et al. (2005) were already combined to assess and illustrate the quality of flood damage data subsets that can be retrieved from the flood damage database HOWAS 21 (Thieken et al., 2009). In our study we apply a four level ordinal scaling scheme that grades the level of quality reached within each quality dimension in a range from 0 (no information/quality), 1 (low quality), 2 (medium quality) to 3 (high quality).

The ordinal scaling allows calculation of an overall quality for each document (and each flood event). This so-called pedigree score P (Funtowicz and Ravetz, 1990; Van der Sluijs et al., 2001) is the sum over the QD scores reached in all i dimensions divided by the maximum score possible, in this case ten dimensions by three points (Eq. 1). We choose equal weights for all QD in Eq. (1), meaning that P is 90 % equally influenced by representative, contextual and intrinsic QC and 10 % by the accessibility QC.

$$P = \frac{\sum_i \mathrm{QD}_i}{\sum_i \mathrm{QD}_{i,\max}} \quad i = 1, \ldots, 10 \tag{1}$$

P can reach a value between 0 and 1. A pedigree of 1 would mean that, with respect to the task at hand, the document is complete in its description of the event, that the information contained can be trusted and that the document is easily found and comprehended. The measure P can be interpreted in terms of quality labels, i.e. a document being of no, low, medium and high quality. The ranges of P are based on the consideration of an average score in all quality dimensions $\overline{\mathrm{QD}_i}$ of either 0, 1, 2 or 3. The breaks are defined by the upper and lower bounds of this average, e.g. the lower bound of low quality is defined by a minimum of half of the quality dimensions reaching a score of 1 at least. Table 1 provides an overview of the ranges applied for the interpretation of the overall quality of a document.

It is important to note that P is not meant to label a document as bad or good per se and any new task at hand will yield its own quality results. It provides a measure to assess the overall quality of a report and assists in creating an overview of the quality present in the material. At any rate, this overall score needs to be accompanied by an analysis of scores reached in the individual dimensions or combinations of dimensions in order to identify the contextual scope of the document and its strengths and limitations.

2.2.1 Accessibility quality

Accessibility assesses the bibliographic control of the document, which in turn determines the tools by which it can be found and retrieved, and its availability as full text. Furthermore, a document is only fit for use if it can be found using standard search terms for the task at hand. In the case of our study the assessment is based on the systematic search of Uhlemann et al. (2013) and the search terms and tools provided therein. We define the score classes of accessibility quality as ranging from inaccessible, e.g. documents not indexed in any public openly accessible catalogue like archival material, to full access, i.e. indexed documents with an additionally openly accessible full text.

We omit the "Security" quality dimension that is originally included in the concept for data quality of Wang and Strong

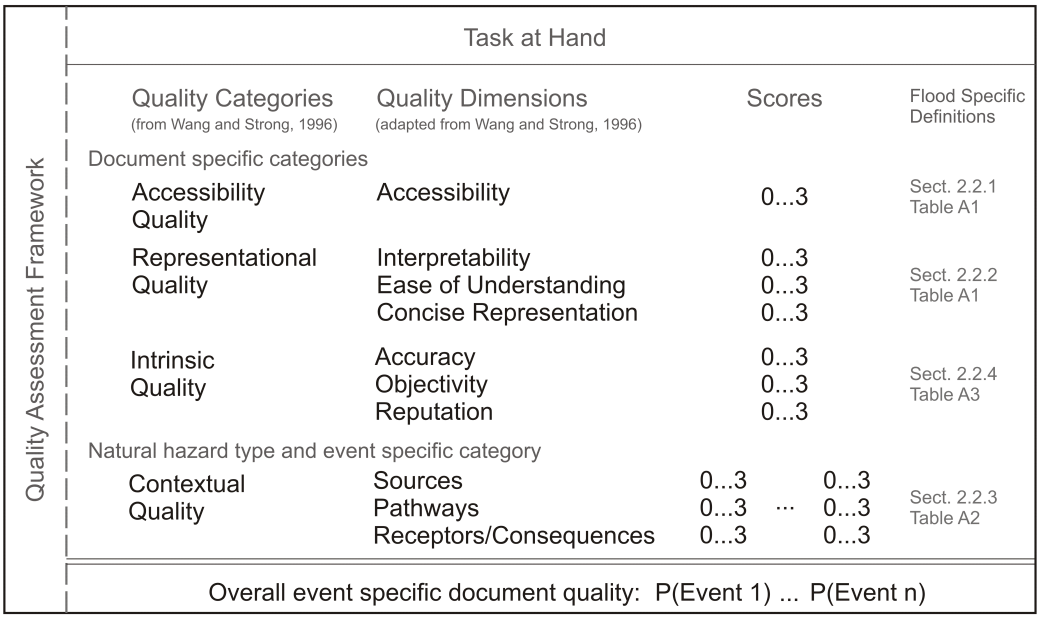

Fig. 1. Overview of the quality assessment framework.

Table 1. Quality label per document.

$\overline{\mathrm{QD}_i}$	$\sum_i \mathrm{QD}_i$	P	Quality label
0	0–4	0.00–0.13	No quality
1	5–14	0.17–0.47	Low quality
2	15–24	0.50–0.80	Medium quality
3	25–30	0.83–1.00	High quality

(1996), as the level of protection is not relevant for the quality of a document, or, if understood as copyright or access restriction, it is reversely relevant, as it limits access to the document and therefore is already included in the "Accessibility" quality dimension.

2.2.2 Representational quality

As natural hazards are confined to particular geographical regions, reporting on an event is largely endemic to these regions and consequently the language(s) used. This is useful for communication amongst authorities; however, it creates barriers in deploying these documents in the context of an international research question if the native language is different to English. The "Interpretability" dimension considers this aspect. It evaluates whether the document is drafted in appropriate languages and can therefore be comprehended. We define a document as comprehensible by any individual if it is fully drafted in English. A report that in turn is neither written in the language of the flood-affected region nor in English will be of little use for knowledge synthesis and therefore is rated as incomprehensible.

The "Ease of understanding" dimension relates to formal aspects like the clarity of the writing, the appearance and style of text and figures and the clarity of the structure of the document. In particular, the report should sensibly structure the content that it presents, i.e. should use headings and subheadings in a logical order. This will allow the reader to reference into the document easily without having to scan the entire text. The "Concise Representation" dimension evaluates whether the document is compactly presented without being overwhelming or too coarse and whether appropriate use of additional material (figures, maps etc.) is made. Generally this dimension assesses the amount of relevant and meaningful information provided by the standard of the document's objectives (either inferred from the title of the document or their direct definition in the text). For example, a report that is produced under the title of "The flood of 1983 in the river Rhine" but that consists largely of photo documentation of single damage occurrences without providing a broader contextual frame is not concise in this regard.

In addition to the three quality dimensions above, Wang and Strong (1996) consider "Consistent representation" (defined as the compatibility of the format of the data with that of previous data) to define the representational quality of data. As highlighted, we do not define an a priori standard of event documentation that could be used to infer compatibility of the evaluated document with this standard. We therefore exclude this dimension from the quality assessment.

2.2.3 Contextual quality

Contextual data quality as defined by Wang and Strong (1996) involves judgements on whether the data is value-added, relevant, complete, timely, and of an appropriate amount with respect to the task at hand. We adapt the definition of the contextual dimensions in such a way, that the variety of information needed to characterise any natural hazard event from its causes to the consequences can be best captured. Therefore we employ the commonly adopted Source-Pathway-Receptor-Consequence concept (SPRC) (Samuels and Gouldby, 2009). This concept is useful in that it can be universally applied to all natural hazards. It represents the systems and processes that lead to particular consequences of the hazardous event. We incorporate this concept into the definition of the contextual quality criteria and create three quality dimensions: sources, pathways, receptors/consequences. Within each of these dimensions the original contextual dimensions (added value, relevancy, completeness, appropriate amount of information) of Wang and Strong (1996) are inherently considered.

The grading for each of the three contextual quality dimensions is based on the amount of information presented with respect to a set of expected variables or attributes and the degree to which quantitative analyses are provided in favour of qualitative or descriptive information on the event. If no or an insignificant amount of information is provided, the score is zero. A quantitative analysis of the expected list of attributes leads to a score of 3, and largely descriptive and rather incomplete information leads to score 1. In that sense the framework also addresses the added value that a flood report may provide, as score 3 analyses contain attributes that are commonly difficult to obtain in the course of a flood risk assessment (or related flood research tasks).

Within the "Sources" QD we assess the degree of information presented for the atmospheric processes, the catchment state and the runoff processes. In order to be complete, the event documentation has to capture the processes on three temporal scales: preconditions, and initiating and maintaining conditions. It should be spatially resolved at a level that allows identification of the main regions of flood origin (geographical and/or stratified by elevation). Depending on the type of the flood (in particular the flood season), the report should include information on the following state variables or processes and highlight their role in the flood generation: circulation patterns, precipitation (snow/rainfall; intensity, duration, advection/convection/orographic enhancement), temperature, snow accumulation/melt, catchment state, soil saturation (e.g. through precipitation indices, monthly anomalies, or runoff coefficients), frost.

The "Pathways" QD grades the information provided on the flood propagation in the river, processes that influence the flood wave formation and the inundation effects encountered. Depending on the course of the flood, the following variables or processes need to be addressed: affected river stretches, timing and duration of flood peak, effects of superposition of flood crest at confluences, flood volume; effects and types of flood retention due to defence failures (breach location and type) or operations of defence structures (polders, dams, ad hoc defences by mobile flood protection); inundation extent (and duration).

We combine the evaluation of the amount of information provided for the affected elements of the flood and the damages encountered in the "Receptors/Consequences" QD. In a spatially explicit manner, the report needs to differentiate the documentation of the consequences and therein particularly of the (monetary) damages to the sectors that were affected (private households, business, infrastructure, etc.) and by the type of damage (direct, indirect; insured, uninsured; intangibles like fatalities).

2.2.4 Intrinsic quality

The intrinsic quality category reflects the overall "trustworthiness" of the report. There are general rules for creating trust, like correct and sufficient data, replicable methods and results, discussion of results and uncertainties, unbiased and impartial conclusions that are supported by evidence, procedures for quality checking etc.

"Accuracy" assesses to which degree the data used for the contextual aspects of the report are reliable, i.e. of sufficient amount and certified free of error, and whether adequate and documented methods were used. It therefore addresses the degree to which the results are reliable (error free) and reproducible (amount and sources of data are clear). It has to be noted that this dimension provides a summary assessment over the entire report. It is likely that the accuracy of the presented data and methods varies within a report and particularly across the different aspects of the SPRC. The quality grading then integrates over all aspects, providing a mean quality score, which is particularly relevant for reports that cover a number of the contextual aspects. Therefore the score should not be used independently of the quality framework, i.e. for labelling single information entities as in(accurate).

"Objectivity" assesses the validity of the results and conclusions reached, i.e. the degree to which they are supported by the evidence of the data and analysis. Any evidence of influence of interests (political, personal, institutional, etc.) leads to lower quality grades. As for accuracy, objectivity is assessed as a mean over the contextual aspects covered by the report.

The "Reputation" quality dimension is reflective of the common concepts of assessing research by indirect measures like author credentials, publisher reputation or level of peer reviewing (see e.g. Bailin and Grafstein, 2010 and Hjørland, 2012). We consider the technical experience of the producing body as important for the quality of the report; however, we acknowledge peer reviewing as the gold standard for quality control. Similarly we assume consortium work that included

numerous publishing bodies of high technical experience as having undergone quality checks by multiple resorts. The quality grades decrease with the degree of generality of the producing body. If the author of the report or the publishing body cannot be affiliated or seems dubious, the grade is set to zero.

2.3 Concordance check – Kappa test

To check for consistency in the interpretation of the definitions of the quality dimensions we test the framework using the kappa statistic κ, which is a widely used measure to express the degree of agreement between two or more independent judges (in the following referred to as peers) for categorical data. It is commonly used in systematic reviews, i.e. in meta-analysis, both in clinical studies as well as environmental studies (Higgins and Green, 2011; Centre for Evidence-Based Conservation, 2010). Within these studies often a limited number of experts (ranging between one and three) are used to test a particular selection scheme for its objectivity.

Kappa expresses the percentage agreement corrected for chance (Kraemer et al., 2002; Fleiss and Cohen, 1973), or in other words it evaluates the proportion of records for which there was agreement against the amount of agreement that would be expected by chance alone (Centre for Evidence-Based Conservation, 2010). For the quality assessment framework this means, that the scoring in the quality dimension by a number of peers is compared to a hypothetical random score result. Kappa then provides the respective test measure. For nominal categories Cohen's Kappa (Cohen, 1960) is given as

$$\kappa = \frac{P_\mathrm{o} - P_c}{1 - P_\mathrm{c}} \tag{2}$$

P_o – proportion of units in which peers agreed

P_c – proportion of units for which agreement is expected by chance

In an ideal case when there is perfect agreement or disagreement between the peers, Kappa is scaled to vary from -1 to +1, with zero indicating exactly chance agreement. However, in the likely case of encountering disagreements, the marginals of the concordance matrix are unequal and the upper and lower limits of κ are slightly less than 1 or, respectively, slightly larger than -1. The exact upper boundary of the maximum possible kappa κ_M is defined as (Fleiss and Cohen, 1973)

$$\kappa_\mathrm{M} = \frac{P_\mathrm{oM} - P_\mathrm{c}}{1 - P_\mathrm{c}}, \tag{3}$$

whereby P_oM is found by pairing the values of two peers such that the concordances are maximised. Negative values always suggest agreement poorer than chance and positive values indicate better than chance agreement (Cohen, 1960). A kappa of unity (or the maximum possible kappa) means that both peers are in total agreement and, in the context of this study, that the quality attribution to the documents is not randomly done. In that case, the definitions of the quality dimensions proposed would be perfect in the sense that their interpretation is "bijective".

The rating of the quality dimensions in this study is based on an ordinal scale. Cohen's Kappa is however restricted in its use to strictly nominal (independent, mutually exclusive and exhaustive) categories. For ordinal scaled data not only the absolute concordances need to be taken into consideration but also the relative concordances expressing the nearness of the assignments. Variants of Cohen's Kappa have been developed for ordinal scales (Fleiss and Cohen, 1973; Lowry, 2012). They consider the relative concordances by introducing weights to neighbouring classes of assignments, whereby strong neighbourhood behaviour is best captured by assigning squared weights. Weights are assigned to each combination of scores (score class i by peer A and score class j by peer B) according to

$$v_{ij} = 1 - \frac{(i-j)^2}{\max(i-j)^2}. \tag{4}$$

According to Fleiss and Cohen (1973) the proportions of Eq. (2) are then calculated as the weighted mean observed degree of agreement P_o

$$P_\mathrm{o} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{ij} v_{ij}, \tag{5}$$

and mean degree of agreement as expected by chance $P_\mathrm{c}.P_\mathrm{c}$ is given by the weight adjusted joint probabilities of the marginals,

$$P_\mathrm{c} = \frac{1}{N^2}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{i\bullet} n_{\bullet j} v_{ij}. \tag{6}$$

N – total number of reports;

m – number of score classes;

n_{ij} – number of reports assigned to score class i by peer A and score class j by peer B;

$n_{i\bullet}$ – total number of reports assigned to score class i by peer A;

$n_{\bullet j}$ – total number of reports assigned to score class j by peer B.

The sampling characteristics of κ can be captured by computing a confidence interval that expresses the chances that the population value of κ falls within the computed limits.

The standard error of κ can be approximated by (Cohen, 1960)

$$\sigma_\kappa = \sqrt{\frac{P_o(1-P_o)}{N(1-P_c)^2}}. \quad (7)$$

With large N it can be assumed that the sampling distribution of κ approximates normality and the respective confidence limits at 95 % are given by $\kappa = \pm 1.96\sigma_\kappa$.

3 Results and discussion

The paper is supplemented by the entire set of documents evaluated in the frame of this study including both the scores given in the individual quality dimensions as well as the overall pedigree. Additionally, the results of the kappa test are presented. The data supplement is available through an open access data server and can be permanently addressed and referenced to using the doi provided in Uhlemann (2013).

3.1 Kappa test results and method discussion

We test the quality assessment framework by drawing a random sample of 10 studies from the entire set of documents which are then evaluated by two independent and experienced scientists (5 studies each) from the field of flood risk assessment but outside of the collective of authors of this study and who were not involved in the design of the quality assessment framework. In the following we will refer to them as peers. In order to avoid positive biases towards report quality assessments we deliberately did not request for peers from the field of water authorities or other experts that largely publish their work as grey literature. Results from both peers are pooled and checked for concordance with the scores given by the authors of this study.

Table 2 comprises the results of the weighted kappa κ for each quality dimension and as a total by pooling the agreements of all QD. The number of cases N considered in the test is 100 for the pooled κ (10 quality dimensions, 10 studies), making the measure statistically more robust than for the single QD ($N = 10$). The κ measure is accompanied by its respective maximum value κ_M, its standard error σ and a confidence interval (CI at 95 % confidence).

For the entire framework we compute a weighted kappa of $\kappa = 0.89 \pm 0.14$ at 95 % confidence interval. The maximum possible kappa for the given marginals is $\kappa_M = 0.98$ and the κ therefore reaches 91 % of the maximum possible agreement. The result indicates a high degree of concordance between the peers and we can exclude pure chance operations in the application of the framework to an arbitrary document. Furthermore, investigating the concordance matrices shows that all disagreements deviate by a maximum of one score.

When computing κ for the individual QD the number of cases N is limited to 10 and the statistical significance of the kappa is limited entailing a high sensitivity to individual changes in just one pair of evaluations, as can be inferred by the large standard error and confidence intervals. Keeping the small N-size in mind, in the following we will discuss the agreements reached in the individual QD. In order to substantiate the interpretation we requested a feedback from the peers upon completion of their assessments.

Some distinct differences in the concordances can be observed between the quality dimensions. For the accessibility, interpretability and reputation dimensions, a nearly perfect agreement in the scoring between the peers can be observed. These QD can be assessed in a strictly objective manner, as they entail no need for textual interpretation by the peer.

A very high to high degree of concordance is given for the contextual quality dimensions, verifying that the assignment of a document to any of the grades is not ambiguous and will be similar for any peer. Differences in the assignments are the result of some degree of subjectivity of each peer in drawing the distinctions between the score classes. This subjectivity is largely related to the technical experience of a peer in any of the specific aspects like e.g. hydrometeorology. Furthermore, as the whole spectrum of aspects from sources to consequences is covered, the grading of a document then also depends on the peer's scientific background and experience (which will likely not be at the same technical level for all aspects). The same accounts for the concordance of the peers in the intrinsic "Accuracy" and "Objectivity" quality dimensions.

Inasmuch as there is perfect agreement in the assessments of accuracy, the author's and the peer's experience was that a common occurrence is the need for indirect inference of the data and methods used within technical reports from governmental authorities who draw on their own data and standard methods (as given by official guidelines, e.g. for the assessment of return periods) and presuppose clearness about this. Generally, assessing the accuracy of event documentation that aims at compiling information rather than presenting scientific results (data–method–result–conclusion type of analysis) poses some challenges to a peer. It is then up to the peer to decide whether the in-text (and sometimes in-figure) references allow for a sound judgement of the quality of the data and methods used. Similarly, assessing the objectivity of the conclusions reached is a largely investigative process that refers to circumstantial rather than hard evidence. It is then particularly a matter of the technical background and expertise of the peer that allows identification of flaws or misjudgements. For both "Accuracy" and "Objectivity" the assessment of the quality is based on a mean perception over the entire document and therefore differences are introduced by the need to weigh all aspects presented. Considering the above-mentioned degrees of freedom in the interpretation of the score definitions, the agreement reached amongst the peers can be considered substantial.

The least agreement is given for the representational "Ease of understanding" and "Concise representation" quality dimensions. The result indicates that the interpretation of the

Table 2. Kappa statistics for the individual quality dimensions ($N = 10$) and pooled over all quality dimensions ($N = 100$).

	Access-ibility	Inter-pret.	Ease of underst.	Concise repres.	Sour-ce	Path-way	Recept./ Cons.	Accu-racy	Objec-tivity	Repu-tation	All QD
κ	0.91	1.00	0.66	0.29	0.84	0.76	1.00	1.00	0.78	0.90	0.89
κ_M	0.95	1.00	1.00	0.91	0.97	0.97	1.00	1.00	1.00	0.95	0.98
σ	0.27	0.00	0.72	0.72	0.27	0.32	0.00	0.00	0.47	0.30	0.07
CI 95 %	0.53	0.00	1.42	1.42	0.54	0.62	0.00	0.00	0.91	0.60	0.14

definitions in both dimensions is influenced more by the peer's mental framework and expectations towards the form and content of a document, i.e. by the application of standards for scientific publications. The discussion showed that most of the uncertainty in the scoring occurs when the peer's initial expectation differs from the actual definitions of the scores in the quality dimensions. The representational quality criteria are defined to assess the overall quality of the document, prescribing a central role towards the objectives of a report. Inferring the objectives is in turn an investigative process. Frequently, they are not stated explicitly and have to be inferred by the reader mostly from the title of the document. Once inferred it is then still a matter of the peer's expectation towards the comprehensiveness and form at which content is presented. This opens a wide interpretational space for the peers and helps explain the discordances in the scoring. As the QAF is designed to allow assessment of the quality of a heterogeneous group of material (from long technical reports to short ministerial statements to full journal articles), the definition of an adequate size of a document in the conciseness dimension cannot be accompanied by a clear page limit. Therefore the discordances in this QD have to be accepted as a source of uncertainty.

In order to assess the effect of peer disagreement on the overall pedigree, we compare the resulting P values. For the 10 documents chosen, Fig. 2 compares the pedigree scores assigned by the author against the pedigree scores assigned by the peers. The maximum difference encountered is +0.13, equalling a score difference of four (a difference of one score leads to an alteration of P by 0.03 units). A slight bias towards lower pedigrees can be observed; however, given the small sample size, no statistically significant inference can be drawn. Our sample size is comparatively small, i.e. in total 10 studies that were drawn randomly and that constitute a good representation of the type of documents contained in the data set. Given the high kappa coefficients within the small sample, we expect that increasing the sample size will not significantly alter the overall kappa; however, it would improve the statistical robustness.

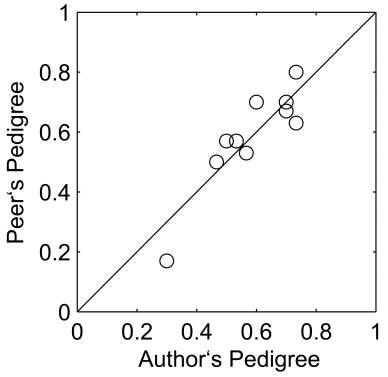

Fig. 2. Comparison of the peer's P with the authors' P ($N = 10$).

3.2 Quality distribution and document characteristics

For the 133 assessments undertaken in the course of this study the pedigree ranges between a minimum of $P = 0.23$ and maximum of $P = 0.97$, with a median of $P = 0.57$. Figure 3a provides an overview of the spread of the pedigree showing a distribution slightly skewed towards higher quality. The bulk of the documents (67 %) reach medium pedigrees in the range of $0.5 \leq P < 0.8$ highlighting that the overall quality of the documents is generally good. One quarter of all documents reach less than 50 % of the maximum quality (low quality) with only four reports being evaluated in the range of $0.2 \leq P < 0.3$. As a result of the stringent inclusion criteria applied during the systematic search (Uhlemann et al., 2013), reports that do not or hardly meet any of the quality criteria are not present in the evaluated material as a report that scores a pedigree below 0.2 (reaching a maximum of 6 out of 30 points) would likely not be accessible, be of low intrinsic quality and generally fail to provide information on the SPRC of the event and therefore not fulfil the inclusion criteria that relate to the task at hand. Eleven reports were evaluated to be of high quality ($0.8 \leq P < 1.0$) and one report can be considered as qualitatively near perfect (Gräfe, 2004).

In Fig. 3b we show the distribution of the pedigree stratified by the type of the document. Technical reports form the largest group of flood event documentation. Generally, technical reports (most of which are produced by governmental agencies) span the whole spectrum of quality; however, they exhibit a tendency towards higher quality, i.e. they form the largest part of documents that are assessed as being of medium to high quality. They provide a format that is obviously more suited for the documentation of flood events. In contrast, contributions in national technical journals that are

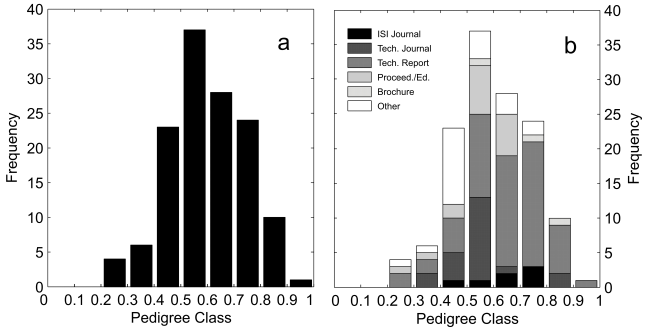

Fig. 3. Histogram of the pedigree results for all evaluated flood reports **(a)**, and further stratified by reference type **(b)**.

also used as a medium for publishing by governmental agencies tend towards lower quality. Reports produced in the academic/scientific environment that are published in ISI-listed journals are evaluated as medium quality with no deviations to poor or high overall quality. This means that they fulfil standards in presentation and the intrinsic quality is high; however, often they are constrained by the contextual scope, i.e. in most cases they focus on only one particular aspect, like the hydrometeorological conditions of the flood.

3.3 Components of quality

The overall quality of each report as expressed by the pedigree is reflective of the scores reached in the individual quality dimensions. In order to add meaning to the summary measure P and to gain more understanding of the features that determine report quality, we analyse the quality scores reached in the individual quality dimensions and their correlations.

A first survey of the composition of the pedigree by its ten dimensions (see data supplement) reveals that an overall low pedigree also means that the scoring in any of the individual quality dimensions is likewise low, i.e. a report that is of low overall quality is also comprised of evenly low scores in the individual quality dimensions. Effects of one quality dimension excelling the others by far are not encountered as they are not completely independent. This holds except for the contextual criteria, where it is likely that only one or two dimensions have been within the scope of the report. A report that is of low representational quality will likely also be of lower contextual and intrinsic quality. Similarly, as the overall pedigree increases, the scores in the quality dimensions also largely increase. Only in the range of 0.4 to 0.5 do the first quality dimensions reach the maximum score of 3; often this is either in the intrinsic or representational criteria. For reports of $P > 0.6$ the contextual quality dimensions additionally reach high scores.

In the following we will analyse the four quality criteria and their dimensions in more detail. Table 3 presents an overview of the performance of all documents in the individual quality dimensions and highlights the frequencies at which certain quality levels are reached. Additionally, Fig. 4 illustrates the scores that were reached by the documents per QD in relation to the overall document quality.

3.3.1 Accessibility

The majority of flood event-specific reports (62.4 %) can be found and retrieved using standard search terms on standard search engines within the national framework provided. However, only 17.3 % of all documents are both searchable and open accessible in their full text. 14 % of the material can only be found using non-standard searches, i.e. by searching the online portals of flood relevant institutions (where in fact the document may be available as a free download) (score 1). 6 % of the material used for this study is actually inaccessible and became available either by chance (donations etc.) or was found in reference lists. In Uhlemann et al. (2013) we already provided a detailed analysis of the accessibility of literature relevant for the task at hand, and the evaluation of the subset of material used within this study reflects the results presented for the entire set of literature.

Comparing the quality level reached in this dimension with the scores reached in the contextual dimension highlights that the degree to which the document is easily accessible has no implication for the contextual depth of the document. I.e., open access documents (one open access journal, two proceedings paper, three international river commission reports, four governmental press releases; the others: states or federal government reports by authorities with technical expertise; all published after 1996) likewise contain few to many aspects relevant for understanding a flood event and the overall pedigree exhibits some spread (see boxplot for score class 3 of accessibility in Fig. 4). However, both the representational as well as the intrinsic quality dimensions are graded as good to complete (scores 2–3) for open access documents, indicating that the degree of quality control increases (irrespective of peer review). This may be attributed to the author's or authority's additional effort towards meeting quality standards upon the document's enhanced public exposure.

3.3.2 Representational quality

Uhlemann et al. (2013) already highlighted that the majority of documents compiled in the systematic search is drafted in the language of the producing body, in this case in German (90.3 % for the subset used), a main characteristic of grey flood reporting. Less than 10 % are completely published in English and within the German documents 4 % are accompanied by English titles, captions and/or summaries. The language component however does not influence the overall quality of the documents. The spread of pedigrees for the German documents reflects the pedigree distribution with a median of $P = 0.57$. The five documents that are

Table 3. Distribution of the scores per quality dimension, given in percentage of documents per score class. Bold numbers indicate the scores reached by the majority of documents.

Quality Criteria	Access. Qual.	Representational Quality			Contextual Quality			Intrinsic Quality		
Quality Dim.	Access-ibility	Inter-pret.	Ease of Underst.	Concise Repres.	Sources	Path-ways	Recept./ Cons.	Accu-racy	Objec-tivity	Repu-tation
Score	Absolute frequency									
0	8	0	0	3	21	22	**92**	4	1	2
1	19	**115**	19	35	37	**61**	24	24	7	10
2	**83**	5	**72**	**53**	**44**	45	8	**62**	36	**89**
3	23	13	42	42	31	5	9	43	**89**	32
Score	Relative frequency (%)									
0	6.0	0	0	2.3	15.8	16.5	**69.2**	3.0	0.8	1.5
1	14.3	**86.5**	14.3	26.3	27.8	**45.9**	18.0	18.0	5.3	7.0
2	**62.4**	3.8	**54.1**	**39.8**	**33.1**	33.8	6.0	**46.6**	27.1	**66.9**
3	17.3	9.8	31.6	31.6	23.3	3.8	6.8	32.3	**66.9**	24.1

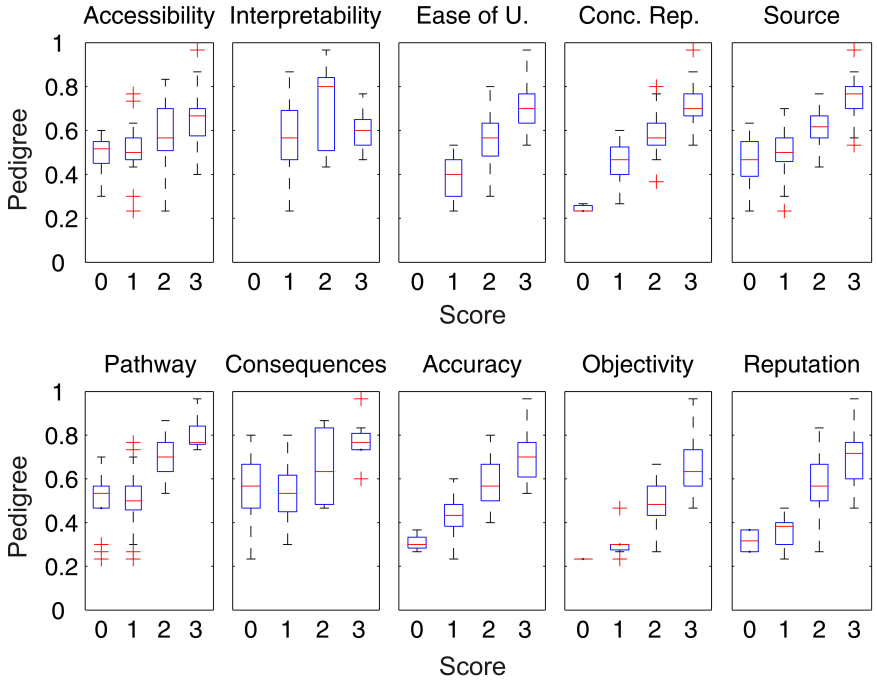

Fig. 4. Boxplots showing the distribution of the overall pedigrees reached by the documents in a particular score class of a quality dimension. The sample size per box can be inferred from Table 3.

accompanied by English annotations are of a slightly better quality (one governmental technical report, four contributions to national technical journals with respective publishing requirements). The 13 documents that are published fully in English (eleven articles in ISI Journals or proceedings, two European Commission (EC) contributions) are in turn characterised by lower scores in the contextual quality dimensions and none exceeds a P of 0.8. Two main reasons can be identified: the journal articles mostly present a detailed analysis of the meteorological aspects of the floods however they do not treat aspects of the catchment conditions, pathways and/or consequences. The proceedings papers in turn treat all aspects but, due to the condensed format, in a rather brief manner. The EC contributions concern official statements of the European Union towards political actions in the aftermath of the August flood in 2002. Here, the special effect occurs that the content provided scores as insignificant in all contextual quality dimensions.

The results obtained within the other two quality dimensions of the representational quality criteria ("Ease of understanding" and "Concise Representation") indicate that the flood-specific reports available for this paper are at the majority qualitatively good in their representational characteristics. This means that the quality of the use of language and terminology as well as the structure of the documents is generally good and also that the content is largely presented making adequate use of figures, tables etc. and providing a balanced report in terms of length of the document with respect to its objectives. However, it has to be noted that quite a number of reports also suffer from an overly condensed presentation (26.3 %), e.g. the presented content is not representative in its volume or presentation to fulfil the aims of the report. The most frequent occurrence in this context is the overall brevity of the report (less than one to a few pages of text). Some reports present results largely by figures, tables or photos that are insufficiently accompanied by explanatory text. Clearly, if a report is lacking in its conciseness and structure, the fitness for use is likewise limited. Figure 4 shows a clear correlation of both dimensions with the overall quality of the documents. Those reports that are of an overall good quality are exclusively well written and well structured.

3.3.3 Contextual quality

The summary statistics in Table 3 show that the contextual dimensions are assessed heterogeneously. 84.2 % of all reports include an analysis of the hydrometeorological causes of the flood, where one third contains detailed (score 2) and slightly less than one third either complete (score 3) to limited information (score 1). Documents that capture the hydrometeorological origin of a flood rather coarsely are of a generally broad and mostly condensed nature (conciseness scores 1 or 2). Most reports in this score category present a summary of the hydrometeorology based on secondary sources and provide a largely qualitative and descriptive characterisation of the flood initiating conditions (7 out of 37 present own data along the analysis). Reports that were assigned a quality score of 2 can be categorized into two types: first, reports with a focus on the atmospheric aspects providing detailed and mostly also quantitative analysis (however, omitting the other aspects of flood initiating conditions) and, second, reports that provide a more detailed description of the entire spectrum of the flood generating processes; however, these are not fully quantitative.

The pathways of the flood event are mostly treated partially and hardly any document fulfils the quality expectations of comprehensive documentation of the flood routing and particularly of the inundation processes. Most documents fail to provide systematic documentation of the entire flood-affected areas and the mechanisms that led to inundation (which is mandatory for quality score 3). Only five reports attempt to document the flood extent fully, mostly however in a descriptive manner. Two reports on the summer flood in 1954 actually provide a mapped overview of the flood-affected area. Routing charts are frequently presented together with analyses of the wave propagation in the river. Also, breach occurrences or volume and effects of retention on the flood crest are documented, providing considerable added value, as this information cannot be obtained in a data-based analysis that relies on series of discharges or water levels. The scores assessed for sources and pathways show a slight positive correlation. Most reports present a combined analysis of the hydrometeorology and the flood wave propagation, and often at comparable depth.

Almost 70 % of all documents do not include documentation or analysis of the consequences of the flood event. If considered, the information is mostly treated as an add-on in the report (18 %, score 1) presenting only the official overall damage (or an estimate of it) and/or giving examples of damages at (arbitrary) locations, not allowing for an overview of the amount and types of damages (or their spatial pattern). However, a number of documents, even though not delivering any descriptive or quantitative analysis of the damages (and therefore score 0), include photo documentation. In case of verifications of particular damage occurrences they may still provide useful information. The nine reports that contain detailed information on the consequences of the flood are all of a high quality. Except for one report on the August flood in 2002 that exclusively treats the damages of the event (therefore only reaching $P = 0.6$), in all other cases the sources and pathways are also treated in detail and $P \geq 0.73$.

3.3.4 Intrinsic quality

The intrinsic quality of a document represents the trustworthiness of the content presented. The accuracy of the data and methods used varies amongst the material; however, the majority of all documents (79 %) are of medium to good accuracy. As indicated in Sect. 3.2, a number of low to medium scores can be attributed to the fact that the document under validation failed to highlight the data or information sources that were used and/or did not document which methods were used to obtain the results and conclusions presented. Most documents present their results in an objective manner and only in a few cases could direct evidence of falsifications or unjustified interpretations be found. For three documents clear bias in the interpretation of the flood event could be found. These are (1) politically motivated, as for two reports from the former German Democratic Republic (GDR), where the actual inundation effects and damages of the described flood events were concealed to emphasize the successful management of the flood, and (2) they are the result of biased positions of the author, particularly in the discussion of flood engineering vs. restoration of flood plains (1980s), and since the 1990s also in the discussion of the effect of climate change on the frequency and magnitude of river floods. In both cases the conclusions reached are not supported by the presented data analysis. The other five reports of score

class 1 generally failed to meet their objectives. Generally, reports of low objectivity are also of an overall low quality. Once the quality in the other quality dimensions is given, the reports are mostly presented in an unbiased, objective manner. Also, if a report is accurate (score 3) it is always fully objective (score 3); however, not vice versa.

The reputation of a document is judged based on the authorship as well as on the level of independent reviewing that the document has been subjected to. As highlighted earlier, most reports were not published in a peer-reviewed journal. Therefore the majority of documents only receive the second highest score (67 %). For peer reviewed material (which includes a limited number of technical national journals too) and technical reports published by any international river commission or by a consortium of different agencies/stakeholders a thorough quality control was assumed. The box-plot for score class 3 in Fig. 4 highlights that the documents of high reputation are generally of higher quality. Assessing the type of documents underneath this distribution reveals that these higher quality documents of $P \geq 0.7$ are nearly exclusively non-ISI publications highlighting that the format of technical reports better suits the needs for event documentation.

3.4 Application: event analysis example

In the following we will highlight the potential of combining information from many reports in order to understand a particular flood event. The quality (overall and in the dimensions) of the reports is used to judge their applicability for the task. To improve the readability of the text we will use the record number of a report instead of the full citation to a document. The record number is a unique identifier which allows the identification of the document via the endnote database and evaluation tables provided in the data supplement. A summary for other flood events can be quickly obtained by querying the data supplement for the rank of the trans-basin floods (Uhlemann et al., 2010) which serves as a unique identifier, too. In the case of a document containing analyses of more than one flood it is listed several times; each time annotated with the respective trans-basin flood rank for which it has been evaluated using the QAF.

A full assessment of all flood events is beyond the scope of this paper and we use the example of the flood event in October-November 1998. According to the analysis of Uhlemann et al. (2010), the flood event of autumn 1998 ranks as the 5th strongest trans-basin flood in Germany in the period 1952–2002. It affected all river basins in Germany except for the Odra, lasted for 14 days from 29 October to 11 November 1998 and is the only autumn flood in the record. Compared with other flood events in the 90s (Rhine floods in 1993 and 1995, the Odra flood in 1997, the Danube flood in 1999), the event has received limited public and scientific attention. In total ten documents were identified in Uhlemann et al. (2013), eight of which belong to the group of special reports. These reports have been evaluated in the course of this study using the quality assessment framework. Table 4 provides an overview of all studies and the results of the quality assessment. We illustrate the spatial scope of the reports by aligning the vector of scores assigned in each quality dimension to four classes of spatial extent. Three of the reports cover the flood on a national scale, with the report of the German Meteorological Service (# 482) also making reference to other extreme weather and flood occurrences on an EU scale. Four reports have a largely regional focus, with # 148 extending in the analysis both to national and regional aspects and one report documenting the flood in the smaller catchment of the river Ruhr (# 30). The latter is the only report issued by a local water management company and all other reports have been issued by governmental authorities at federal or state level. The reports are drafted exclusively in German. The regional to supra-regional reports can be accessed easily, as they are indexed in public access catalogues and two of them are even fully open access. In turn, the three reports with a regional to local scope are exclusively available as downloads (# 30) or web text (# 37 and # 486) on the respective sites of the providers, but have not been indexed.

The result of the quality assessment for the documents of the 1998 event reveals a wide spectrum of document quality that resembles the distribution of P in the parent population, with $P_{\min} = 0.23$ and $P_{\max} = 0.8$ and a median of $P \sim 0.6$. Except for # 37, all reports present their analysis in a largely accurate and objective manner; however, most exhibit minor to substantial shortcomings in the way they are presented (in most cases due to a very condensed presentation and in the case of # 148 due to deficiencies in the structure of the otherwise very detailed and useful content). Considering the low quality of # 37, we exclude the study from further analysis. The content presented in the reports largely focuses on the analysis of the sources and pathways of the flood event. Only # 148 includes direct notions of the damages encountered, thereby constituting one of the most complete documents of damages that can be found in the entire set of documents analysed in the course of this paper.

In the following we will present a synthesis of the information contained in the documents on the SPRC of the 1998 flood event, highlighting where information is complementary, (non-)confirmative or missing. As indicated, the information contained in the seven reports we use for the synthesis can be considered as largely trustworthy (on average), with differences in P resulting from various levels of depth in the contextual dimensions. In order to keep the scope of this paper, the synthesis is necessarily kept short.

Sources: In the course of a persistent predominantly westerly circulation from mid to end October, connected with storm activities, a series of frontal systems transported moist Atlantic air masses over the central European continent (#486, #28, #148). #482 highlights that the scale of the frontal system extended on a W–E extension from the British

Table 4. Overview of the flood event reports available for the flood in October–November 1998 and their quality assessment. For each document the respective record number, the issuing body, the Pedigree and the vector of the scores assigned in each quality dimension (in the order given by the QAF, aligned to spatial scope of the report) are given. The contextual dimensions are highlighted in bold.

Record No.	Issuing Body*	*P*	QD-Vector aligned to spatial scope of report			
			EU	National	Regional	Local
# 482	DWD	0.57		[3|1-2-1|**2-0-0**|3-3-2]		
# 159	BfG	0.67		[2|1-3-2|**2-2-0**|3-3-2]		
# 94	BfG	0.47		[2|1-2-2|**0-1-0**|2-2-2]		
# 148	LA BW	0.80		[2 | 1 – 2 – 2 | **3 – 3 – 3** | 3 – 3 – 2]		
# 28	LA RP	0.67		[3|1-3-2|**2-2-0**|2-3-2]		
# 486	LA BY	0.50		[1|1-2-1|**2-1-0**|2-3-2]		
# 37	LA NI	0.23		[1|1-1-0|**1-1-0**|1-0-1]		
# 30	Union	0.63		[1|1-3-3|**2-2-0**|2-3-2]		

*Abbreviations: DWD – German Meteorological Service, BfG – Federal Institute for Hydrology, LA – State authorities of Baden-Wurttemberg (BW), Rhineland-Palatinate (RP) and Lower Saxony (NI), Union – Water management company for the Ruhr catchment.

Isles to Germany, also causing flooding e.g. in England and Wales. The flood in Germany was caused by a series of widespread extreme heavy rainfall events (locally very heterogeneous) (all reports) that met widespread saturated soil conditions due to an anomalous sum of precipitation in the month preceding the flood (all reports). The heavy rainfall of the flood was influenced by orographic enhancement and foehn (increasing the temperature and pressure gradient) (#486). Both the sums of precipitation in the month of October as well as the event precipitation sums (sub-hourly, hourly, daily) were unprecedented in several places (#148, #30) and were assessed with very high return periods or % anomalies for many regions (all reports).

Pathways: The runoff from the flood initiating rainfall events met already high flow conditions in smaller catchments of central north Germany and also Bavaria (# 30, # 486) and low to mean flow conditions in the main rivers – Rhine, Danube, Elbe, Weser, Ems (# 159). Only # 159 provides some few notations on Elbe, Danube, Ems and Weser (main rivers). The flooding exhibited some flash flood characteristics in smaller catchments of Baden-Wurttemberg (# 159, # 148). Two main flood waves were registered, with the first larger in western and central regions (# 148, # 28), i.e. mid-altitude mountains, and the second larger in alpine (# 159) and south-eastern catchments (# 486). The main rivers were affected at a increasing gradient south–north, with the upper and middle Rhine experiencing peak flow of small return periods $Q(T < 5a)$ (# 148, # 28) and higher peak flows with increasing contributions from tributaries Neckar, Main, Moselle. Generally all reports confirm that the most severely flood affected regions were predominantly small to medium sized (mountainous) catchments. Technical retention options were not operated on the river Rhine (# 28). However, a substantial amount of flood retention basins and dams were in flood operation in mountainous catchments (# 148, # 30). # 28 confirms that the upper Rhine was not and the middle Rhine only lightly affected by inundations the latter resulting in no consequences for adjacent communities. A detailed description (no map) of inundation mechanisms and areas is only provided in # 148 for Baden-Wurttemberg.

Receptors and consequences: An estimate of the overall national damage is given in # 148 with 250 Mio DM. Both # 148 and # 159 are concordant in their overview of the most affected geographical regions on the national scale. The overview of affected regions (# 94, # 148, # 159) indicates that substantially more areas were affected by flood damages than are included in the reporting. Two centres are identified: (1) the north German basins of Weser and Ems and (2) the northern part of the state of Baden-Wurttemberg and adjacent regions. For the latter, # 148 provides a very detailed documentation of the types of damages and sectors affected with special sections on particularly affected local regions (city reports). However, for the state of Bavaria # 486 provides no damage analyses. Similarly, no document provides a damage description for the basins of Weser and Ems.

4 Conclusions

The objectives of our study were two-fold: (1) we aimed at developing a framework for information quality assessment and quality labelling of natural hazard event reports; (2) we wanted to apply this framework for assessing the quality of flood event reports in order to investigate whether these

reports are a useful source of information for flood risk research.

For the two aspects we can draw the following main conclusions from the results of our study and can identify further needs for research.

In order to obtain a complete understanding of the quality of a document, its trustworthiness and applicability for a particular research question, more than the intrinsic quality measures common to approaches to critical research assessment need to be applied, i.e. the quality of a document is determined by accessibility, and representational, contextual and intrinsic characteristics. The quality assessment framework presented in our study considers all these aspects. The user's perspective is an essential strength of the framework, as the quality tag attached to a document is reflective of the actual task at hand (in this case the application to trans-basin flood events in Germany). Therefore it does not label a document as good or bad per se, rather its fitness for a particular use. Our validation of the framework by two independent peers highlights the objectivity and transparency of the approach. We can show that the overall agreement reached between the peers is substantial, that disagreement is in the range of deviations that can be expected from the degrees of freedom in the interpretation of the score definitions and from the influences of the peers' own mental framework and perceptions, and that the overall effect on the quality measure P is low. The framework provides a valuable new tool for assessing the quality of natural hazard event documentation and fills the methodological gap that existed for the quality assessment of grey literature.

The quality assessment framework (QAF) for event documentation is designed to be universally applicable to any natural hazard. The four quality categories are therefore defined hazard independently. Hazard-specific aspects are introduced in the definition of the quality dimension sources, pathways, and receptors/consequences in the contextual quality category. These definitions have to be adapted for hazards other than floods, which we believe is a straightforward task. It will be particularly interesting to apply the presented framework to other natural hazards and eventually to conduct a comparison between both the quality characteristics of event reports from different perils as well as to compare the amount and type of information available.

From the application of the quality assessment framework to trans-basin floods in Germany we can conclude that the majority of flood event-specific reports are of good quality, i.e. they are well enough drafted as well as largely accurate and objective, which allows application of their content for the interpretation and understanding of large-scale flood events and processes. Furthermore, our analysis highlights that the majority of reports present a substantial amount of information on the causes and consequences of flood events. Therefore, these reports can be considered as another important source of information for both scientific and practical questions in flood hydrology and flood risk management. We see the main benefit of our study in its contribution towards reducing the barriers to using information on flood event reports in the context of flood research.

This study is accompanied by the open accessible publication of the entire data set of flood event-specific reports and the results of the quality assessment (Uhlemann, 2013). Together with the data set of Uhlemann (2012), which provides the bibliographic metadata to the flood reports, this opens the opportunity for rapid inclusion of information from flood reports in any new research question. Using the pedigree quality measure conjointly with the quality dimensions allows a quick overview of the contextual aspects covered by a report for any particular trans-basin flood event and allows any user to access the literature in a more directed way. The next logical step towards a complete and concise catalogue of trans-basin floods in Germany would be the synthesis of the information from the many reports currently available per flood event into a standardized event description which takes account of the (un)certainty associated with the information available. We want to stress that the presented quality measure only provides a guideline for the overall quality and depth of information that can expected from an event report. It does not free any user from critically checking the accuracy of single information entities before using them in scholarly argument. Further research will be required to develop a framework to combine these sources of information with results from model- or data-based analysis. Possible frameworks can be the information expansion scheme provided by Merz and Blöschl (2008) or evidence-based methods like that of Van der Sluijs et al. (2005) or Norris et al. (2008).

The majority of flood reports analysed in our study can be classified as grey literature. In consequence, our study presents a quality assessment largely for reports produced by governmental agencies. Therefore, beside the scientific user's perspective, one main outcome of our study is the mapping of the report quality with respect to different types of reports and strategies of information dissemination pursued by the agencies. It will be important to communicate these results to these agencies to feed back the requirements and standards that need to be met to improve the impact of these valuable documents both for scientific as well as practical applications. Furthermore, the quality dimension considered in the quality assessment framework can be used to conduct an expert survey amongst different groups of users and producers of event documentation (from science to state agencies to re-insurance) to define relevant common attributes. This process can help to develop a standardized template of event attributes that need to be covered in an ad hoc and post-event analysis and largely facilitate any risk assessment which is based on past events. It can also help to reassess the relative importance of the individual quality dimensions in the quality assessment framework which are currently treated as equally important.

Appendix A

Table A1. Definitions of the quality dimensions and their respective scores for the "Accessibility" and "Representational Quality" quality categories.

Category Dimension	Accessibility Quality Accessibility	Representational Quality Interpretability	Ease of understanding	Concise Representation
Definition	The extent to which the report is available, and easily and quickly retrievable.	The extent to which report is in appropriate language(s) and can be comprehended.	The extent to which the report is presented in an intelligible and clear manner without ambiguity and can be easily comprehended. Relating to formal aspects of style, structure, writing.	The extent to which the report is compactly presented without being overwhelming or too coarse considering its own objectives. Balanced use of Fig./Tales/photos.
Score				
0	Citation not indexed in any OPAC/SCI. Document not retrievable online or via IL. Donation, archive material.	Language other than English and that of the region described.	Poorly presented, serious flaws in usage of language and terminology, almost not comprehensible, extremely poor appearance of text, figures.	Few lines or paragraphs only, not representative for objectives, or content cannot be related to objectives.
1	Citation not directly indexed in OPAC/SCI (articles in non-indexed journals) or with terms other than SS or indexed in non-public catalogues. However retrievable through IL.	Native language of the region of interest + English title (optional)	Largely comprehensible in use of language, few ambiguities. However, not well-structured document (e.g. title not meaningful, no headings, inconsistent order of topics, poor appearance/readability of parts of document, e.g. poor figure quality).	Strongly condensed content for objectives given. Or too much unnecessary detail (inadequate figure-text-ratio: Report consists mostly of photo documentation or figures in relation to explanatory text).
2	Citation indexed in OPAC/SCI, can be found through SS. IL, no OA. OR: Document not indexed but retrievable as OA from institutional homepage.	Native language of the region of interest + English title, abstract/ summary, table/figure captions	Succinct piece of writing with a fairly clear structure. Some shortcomings in appearance, terminology, and structure.	Largely balanced report (with respect to detail and figure/table use), yet inconsistent in parts (variance between the parts).
3	Citation indexed in OPAC/SCI and open access to full text. Found through SS.	All English	Succinct piece of writing with a clear terminology, appearance and structure (meaningful title, summary, clearly stated objectives and a logical order).	Well-balanced reporting in all parts (appropriate figure/text ratio). Appropriate size of document with respect to objectives.

Abbreviations: OPAC – Online public access catalogue, SCI – Science Citation Index, OA – Open access, IL – Interlibrary loan, SS – Systematic search.

Table A2. Definitions of the quality dimensions and their respective scores for the "Contextual Quality" quality category.

Category Dimension	Contextual Quality Source	Pathway	Receptors/Consequence
Definition Score	To which degree are the atmospheric processes, catchment state and runoff processes described that have lead to the flood?	To which degree are the processes of flood propagation in the river and the inundation effects described?	To which degree is the occurrence or absences of damages described (separated by affected sectors (private households, business) and damage/loss types (direct, indirect; fatalities))?
0	Not/insignificant part of the report	Not/insignificant part of the report	Not/insignificant part of the report
1	Rough typecasting of the hydrometeorological causes. The spatial and temporal dependencies/developments of the flood formation processes are only coarsely sketched. Largely condensed and descriptive with little quantitative evidence.	Rough typecasting of flood wave development. Little coherence in the description of the flood propagation and inundation (or no notion on inundation effects, i.e. only routing charts and/or return periods given). Reference to single aspects, like one dyke breach.	Damages/no damages only described for arbitrary locations, often qualitative only, not distinguishing types of damages and affected sectors. AND/OR: Overall damage estimate is given without further differentiation.
2	Either quantitative, detailed analysis of one aspect of the flood formation OR mostly qualitative analysis of all relevant processes (atmospheric, catchment state and runoff) that have lead to the flood. Evidence of spatial and temporal courses is given.	The flood wave propagation is described and a more or less coherent picture of the course of the flood over time and space is drawn. Notions of defence failures or ad hoc measures, mostly qualitative. Some indications of flood extent or inundation mechanisms.	Estimation of the overall (expected) damage is given. Some damage numbers are provided, distinguishing the affected sectors and types of damages allowing for limited conclusions on the flood impact in the area. OR: spatial reference to unaffected regions
3	The proportions of and the spatial and temporal correlations amongst the flood generating processes are given (quantitative analysis). The dominating CP(s) before and during the event are described. At best indications of advective, convective components, orographic enhancement.	The flood propagation, the influence of dyke breaches, peak amplifications, retention etc. are clearly documented in their causes and in their effect on the flood crest. The course of the flood wave can be clearly understood in its spatial and temporal development. Flood inundation extent is fully documented.	The affected sectors and types of damages are documented in a spatially explicit manner and allowing identifying hotspots of damage and unaffected regions.

Abbreviations: CP – Circulation pattern.

Table A3. Definitions of the quality dimensions and their respective scores for the "Intrinsic Quality" quality category.

Category	Intrinsic Quality		
Dimension	Accuracy	Objectivity	Reputation
Definition	To which degree are the analyses and results reliable (i.e. quality and amount of data and/or sources used, application of appropriate methods)?	To which degree is the report unbiased (unprejudiced) and impartial (independent of political, religious, personal, institutional, business interests) and therefore objective?	Is the author/institution reputable, i.e. does it have scientific or professional experience in the field? To which degree has the document undergone quality control (reviewing)?
Score			
0	No notation on the origin of data/sources and methods on which the results are based	Strong bias, the report clearly falsifies results (faked statistics; strong exaggerations or understatements)	No verifiable source or author. Source or author cannot be identified or affiliated. Quality control of the document cannot be assumed.
1	Quality and amount of data and/or information sources unclear or mainly raw data (uncertified/ unofficial) used. Limited documentation of methods. Results reproducible to a limited degree.	Some aspects are objective, some not; biases can be either inferred directly or assumed from the choice of results that is presented, i.e. results/interpretation not fully supported by evidence.	Reports from institutions or authors with technical reputation but mostly not directly flood related, i.e. core experience in related disciplines. Degree of quality control is either unclear or is assumed to be low.
2	Largely replicable analysis, minor inconsistencies (i.e. documentation of methods); most data used is official/certified, largely sufficient amount of data or information from secondary (reputable) sources.	Most aspects are objective, no apparent falsifications, exaggerations etc. However influence of interests on (the choice of) results cannot be fully excluded. Results/ conclusions largely supported by evidence.	Reports from institutions/authors with a technical reputation in hydrology. Some quality control of the document can be assumed (internal quality check; however, no peer reviewing).
3	Official/certified data/sources of adequate amount. Standard and documented methods applied; discussion on accuracy of data and/or uncertainties provided.	Impartial, objective work. Results presented supported by evidence. No evidence of political, religious, personal, institutional, business interests on results of the study.	Report from authors or institutions with clear flood/hydrologic expertise; (Peer) reviewed article or consortium work that has been cross-checked by several resorts.

Data set description

The data used for this publication are freely available as a data supplement under the creative commons license and can be permanently addressed following the doi given in Uhlemann (2013).

Acknowledgements. We thank the GFZ German Research Centre for Geosciences and the University of Potsdam for their financial support. The main author is grateful for partial funding through the excellence in teaching programme (Junior Teaching Professionals) of the Potsdam Graduate School and a scholarship by the equal opportunities office, both University of Potsdam. We want to thank the authorities and respective individuals who provided information and access to relevant publications for this study, assisted in their retrieval, and for donations. We dedicate our special thanks to Florian Elmer and Philip Bubeck for taking over the peer roles in the kappa test and for their valuable comments. The work of students Florian Betz and Sylvia Wesser who assisted in the application of an early version of the quality assessment is gratefully acknowledged.

The service charges for this open access publication have been covered by a Research Centre of the Helmholtz Association.

Edited by: M.-C. Llasat
Reviewed by: two anonymous referees

References

Bailin, A. and Grafstein, A.: The critical assessment of research : traditional and new methods of evaluation, Chandos, Oxford, 121 pp., 2010.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R.: Introduction to Meta-Analysis, John Wiley & Sons, Ltd., Chichester, UK, 421 pp., 2009

Bragues, G.: Wiki-philosophizing in a marketplace of ideas: Evaluating Wikipedia's entries on seven great minds, MediaTropes eJournal, 2, 117–158, 2009.

Brázdil, R., Kundzewicz, Z. W., and Benito, G.: Historical hydrology for studying flood risk in Europe/L'hydrologie historique pour une meilleure connaissance du risque inondation en Europe, Hydrol. Sci. J., 51, 739–764, 2006.

Burton, I.: Forensic Disaster Investigations in Depth: A New Case Study Model, Environ. Magazine, 52, 36–41, 2010.

Centre for Evidence-Based Conservation: Guidelines for Systematic Review in Environmental Management, Environmental Evidence: www.environmentalevidence.org (last access: 27 September 2012), 2010.

Cohen, J.: A coefficient of agreement for nominal scales, Educ. Psychol. Meas., 20, 37-46, doi:10.1177/001316446002000104, 1960.

Farace, D. and Schöpfel, J.: Introduction Grey Literature, in: Grey Literature in Library and Information Studies, edited by: Farace, D. and Schöpfel, J., De Gruyter/Saur, Berlin/New York, 1–7, 2010.

Fleiss, J. L., and Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educ. Psychol. Meas., 33, 613–619, doi:10.1177/001316447303300309, 1973.

Funtowicz, S. O. and Ravetz, J. R.: Uncertainty and Quality in Science for Policy, Kluwer, Dordrecht, 1990.

Glaser, R.: Klimageschichte Mitteleuropas, Primus Verlag, Darmstadt, 2001.

Glaser, R. and Stangl, H.: Historical floods in the Dutch Rhine Delta, Nat. Hazards Earth Syst. Sci., 3, 605–613, doi:10.5194/nhess-3-605-2003, 2003.

Gräfe, H.: Ereignisanalyse – Hochwasser August 2002 in den Osterzgebirgsflüssen, Sächsisches Landesamt für Umwelt und Geologie, Dresden, 2004.

Higgins, J. P. T. and Green, S.: Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0 [updated March 2011], The Cochrane Collaboration availabel at: http://www.cochrane-handbook.org/ (last access: 21 June 2012), 2011.

Hjørland, B.: Evaluation of an information source illustrated by a case study: Effect of screening for breast cancer, J. Am. Soc. Inf. Sci. Technol., 62, 1892–1898, doi:10.1002/asi.21606, 2011.

Hjørland, B.: Methods for evaluating information sources: An annotated catalogue, J. Informat. Sci., 38, 258–268, doi:10.1177/0165551512439178, 2012.

Hübl, J., Kienholz, H., and Loipersberger, A. (Eds.): DOMODIS – Documentation of mountain disasters. State of discussion in the european mountain areas, Interpraevent, Schriftenreihe 1, Handbuch 1, Klagenfurt, 40, 2002.

IRDR: Forensic Investigations of Disasters. The FORIN Project, Integrated Research on Disaster Risk (IRDR), Beijing, available at: http://www.irdrinternational.org/2013/01/25/forin-report-1/ (last access: 11 October 2012), 2011.

Kraemer, H. C., Periyakoil, V. S., and Noda, A.: Kappa coefficients in medical research, Statistics in Medicine, 21, 2109–2129, doi:10.1002/sim.1180, 2002.

Lowry, R.: Kappa as a measure of concordance in categorical sorting, available at: http://www.vassarstats.net/kappa.html (last access: 5 November 2012), 2012.

MacDonald, B. H., Wells, P. G., Cordes, R., Hutton, G. R. G., Cossarini, D. M., and Soomai, S.: The use and influence of information produced as Grey Literature by international, intergovernmental marine organizations: Overview and current research, in: Grey Literature in Library and Information Studies edited by: Farace, D. and Schöpfel, J., De Gruyter/Saur, Berlin/New York, 167–180, 2010.

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H.: Overview and framework for data and information quality research, ACM J. Data Inf. Quality, 1, 1–22, doi:10.1145/1515693.1516680, 2009.

Merz, R. and Blöschl, G.: Flood frequency hydrology: 2. Combining data evidence, Water Resour. Res., 44, W08433, doi:10.1029/2007WR006745, 2008.

Norris, R., Nichols, S. J., Ransom, G., Webb, A., Stewardson, M., Liston, P., and Mugodo, J.: Causal criteria analysis. Methods manual: a systematic approach to evaluate causality in environmental science, eWater Cooperative Research Centre, Canberra, 2008.

Osenberg, C. W., Sarnelle, O., and Goldberg, D. E.: Meta-analysis in ecology: Concepts, statistics, and applications, Ecology, 80, 1103–1104, 1999.

Ranger, S. L.: Grey Literature in Special Libraries: Access and Use in: GL-Conference series, Sixth International Conference on Grey Literature : Work on Grey in Progress, New York Academy of Medicine (USA), 2004.

Rothstein, H. R. and Hopewell, S.: Grey Literature, in: The handbook of research synthesis and meta-analysis, edited by: Cooper, H. M., Hedges, L. V., and Valentine, J. C., Russell Sage Foundation, New York, 103–125, 2009.

Samuels, P., and Gouldby, B.: Language of Risk – Project Definitions (Second Edition), FloodSite Project, Document Number: T32-04-01, 2009.

Seglen, P. O.: Why the impact factor of journals should not be used for evaluating research, Brit, Med. J., 314, 498–502, 1997.

Simons, K.: The Misused Impact Factor, Science, 322, 165, doi:10.1126/science.1165316, 2008.

Strong, D. M., Lee, Y. W., and Wang, R. Y.: Data quality in context, Commun. ACM, 40, 103–110, doi:10.1145/253769.253804, 1997.

Thieken, A. H., Müller, M., Kreibich, H., and Merz, B.: Flood damage and influencing factors: New insights from the August 2002 flood in Germany, Water Resour. Res., 41, W12430, doi:10.1029/2005wr004177, 2005.

Thieken, A. H., Seifert, I., Elmer, F., Maiwald, H., Haubrock, S., Schwarz, J., Müller, M., and Seifert, J.: Standardisierte Erfassung und Bewertung von Hochwasserschäden, Hydrol. Wasserbewirts., 53, 198–207, 2009.

Uhlemann, S.: Supplement to: Data Expansion: The potential of grey literature for understanding floods, Deutsches GeoForschungsZentrum GFZ, doi:10.5880/GFZ.5.4.2012.001, 2012.

Uhlemann, S.: Supplement to: A quality assessment framework for natural hazard event documentations: Application to trans-basin flood reports in Germany, Deutsches GeoForschungsZentrum GFZ, doi:10.5880/GFZ.5.4.2013.001, 2013.

Uhlemann, S., Thieken, A., and Merz, B.: A consistent set of trans-basin floods in Germany between 1952–2002, Hydrol. Earth Syst. Sci., 14, 1277–1295, doi:10.5194/hess-14-1277-2010, 2010.

Uhlemann, S., Bertelmann, R., and Merz, B.: Data expansion: the potential of grey literature for understanding floods, Hydrol. Earth Syst. Sci., 17, 895–911, doi:10.5194/hess-17-895-2013, 2013.

Van der Sluijs, J. P., Risbey, J. S., and Ravetz, J.: Uncertainty Assessment of Voc Emissions from Paint in the Netherlands Using the Nusap System, Environ. Monitor. Assess., 105, 229–259, 2001.

Van der Sluijs, J. P., Craye, M., Funtowicz, S., Kloprogge, P., and Ravetz, J.: Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: The NUSAP system, Risk Anal., 25, 481–492, 2005.

Wang, R. Y. and Strong, D. M.: Beyond accuracy: what data quality means to data consumers, J. Manag. Inf. Syst., 12, 5–33, 1996.

Weed, D. L.: Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related, Int. J. Epidem., 29, 387–390, doi:10.1093/ije/29.3.387, 2000.

Weintraub, I.: The role of grey literature in the sciences, available at: http://59.67.71.236/download/%7B41613BFD-DC32-45C2-8A25-02135485FDB3%7D.GreyLit.pdf (last access: 9 November 2010), 2000.