Journal: NHESS
Title: **A quality assessment framework for natural hazard event documentations: application to trans-basin flood reports in Germany**
Author(s): S. Uhlemann et al.
MS No.: nhess-1-143-2013
MS Type: Research Article
Iteration: First review

General Comments

The paper addresses the problem of the credibility of grey literature on floods (which presently hampers its implementation by researchers) and proposes a quality assessment framework (QAF) for its evaluation. The point under investigation is a relevant technical question (within the scope of NHESS) which has received little attention in the past, and whose proper treatment could improve present capacity of analysts of understanding flood risk.

The paper adapts existing tools, from other disciplines, to the problem under investigation proposing, this way, an innovative method up to international standards. In this regards, proper credit is given to previous work and authors' contribution is clear; reference is appropriate and fully accessible by fellow scientists. In general, implemented data and methods are clearly described but for statistical tools as better discussed in the following section. The title and the abstract are appropriate; the presentation is well structured but the paper is still too long. Figures and tables are adequate. The technical English language is fluent and precise. Results and conclusion are significant but too much specifically related to the case study. According to this, I suggest some revisions before paper's publication. In the following general and specific comments are supplied.

N.B. text references (i.e. page and lines) are made taking into account the discussion paper format

Major criticisms

1) The "squared weighting scheme" implemented for the kappa test is not clear (section 2.3). This makes difficult also the understanding of kappa test results (section 3.1). Both sections should be re-written and made clearer.

2) Section 3.4: this section does not aim at demonstrating the applicability of the QAF, as stated at the beginning of the section (this was already done in previous sections); rather, the objective is to highlight how available reports and related information are actually useful to understand/answer a specific technical question, if they are jointly used. The section should then be re-written according to this perspective. Moreover, it should be put into light which is the "weight" of information coming from reports with different quality in shaping the overall information (credibility).

3) Conclusions are too much related on German reports and their quality: this was already (extensively) discussed in previous sections. Conclusions should be more generic, discussing how the QAF can be implemented in research, with which improvements and consequences.

4) At present, all quality dimensions have the same weight. However, it could be argued that some dimensions are more relevant than others. This point should be better investigated or at least identified as a priority for future research.

Specific comments

Abstract

Comment 1: page 144 lines 23-25
"Using an example flood event that occurred in October/ November 1998 we demonstrate how the information from multiple reports can be synthesised under consideration of their quality". This is not done in the paper. In section 3.4 there is not any consideration of the quality of reports and their role in the overall information credibility. It's just one main criticism highlighted in previous section.

Introduction

Comment 1: page 145 lines 7
What do you mean with *"any systemic approach"*? Any systemic approach to what?

Comment 2: page 145 lines 15
"Contextual depth" is extensively defined in the following but, at this point of the paper, its meaning is not clear to a wide audience. Please specify.

Comment 3: *page 147 lines 17-7*
"They are commonly applied in the course of systematic reviews and meta-analyses are used to synthesize the available evidence for a given question to identify and assess consistent findings across diverse studies (i.e. statistical analysis of causal linkages, effectiveness of interventions) and to inform policy (Burton, 2010; Borenstein et al., 2009)". Not clear, please rephrase.

Comment 4: *page 147 lines 23-24*
What is *"the environmental level (depicting the general Zeitgeist)"*? Not clear

Comment 5: *page 147 lines 29*
As for contextual depth, *"intrinsic quality assessment"* is explained later in the paper and its meaning is not so clear at this point. Please, specify

Section 2.1

Comment 6: page 149 lines 22-23
"The spatial, temporal and contextual frame for the search is given by the task above". Not clear, specify.

Section 2.2

Comment 7: page 151 lines 26-28
"Within each of these dimensions the original contextual dimensions of Wang and Strong (1996) are inherently considered". Not clear, please specify.

Comment 8: page 153 lines 4-6
"Assuming an average score QD_i of 0, 1, 2, or 3 over all dimensions (example: an average score of 2 would 5 result in a score sum of 10×2 = 20 and P = 20/30 = 0.67), P can be interpreted according to the quality labels of no, low, medium and high quality". How ranges for quality labels have been defined is not clear from this explanation. Please clarify

Comment 9: page 153 lines 22-23
"In defining the quality dimensions we consider the spatial scope at which the report documents an event as reference for the quality expectation and assessment". How the spatial scale plays on reports quality is not clear, even in the following application (section 3.4). Please clarify

Comment 10: page 153 lines 23-25
"In the document specific categories we introduce the convention that each report is assessed with respect to its objectives". Not clear, please specify

Section 3.1

Comment 11: page 162 lines 1-2
"In the following we will discuss the agreements reached in the single QD with respect to the strengths and weaknesses of the definitions of the dimensions and their respective grades". Not clear, please specify

Comment 12: page 162 lines 12-14
"Differences in the assignments can be attributed to the large amount of variables that are covered in any of the contextual QD which introduces a minor degree of subjectivity of a peer in drawing the distinctions". Why should more variables reduce subjectivity? Not clear, please specify

Comment 13: page 164 lines 3-6
"In order to assess the effect of peer disagreement on the overall pedigree we compare the resulting P values (Fig. 2). The maximum difference encountered is +0.13 equaling a score difference of four (a difference of one score leads to an alteration of P by 0.03 units)." The meaning of Fig. 2 is not clear. Relation between P difference and score difference is not clear. Please, specify.

Section 3.2

Comment 14:
ISI journals cannot be considered grey literature. The proposed QAF can be used both to evaluate grey and official literature. That is fine but must be clarify earlier in the paper.

Section 3.3

Comment 15:
Most of discussed results are not evident form Table 3 or Figure 4 but supplementary material is required. This should be highlighted.

Comment 16 page 168 lines 15-16:
"Figure 4 shows (…) a pair wise correlation with the score class 3 of the contextual dimensions and accuracy". Not clear, please clarify.

Comment 17 page 168 line 20:
"83.5%". Is it correct? According to the table the right value is 84.2%

Comment 18 page 170 line 13:
"GDR". What does it mean? Not defined before

Section 3.4

Comment 19 page 172 line 5:
"See section 3.2". Reference is not correct.

Comment 20 page 174 line 3:
"Q(T < 5a)". Is it an error?

Section 4

Comment 21 page 176 lines 14-16:
"A natural extension of the example application presented is the combination of data based and model-based results with the quality-labelled information of the reports resulting essentially in an uncertainty assessment of the available knowledge". This seems a very important point but is not clear. Please, rephrase and clarify.

Comment 22 page 176 lines 20-22:
"Evidence-based or related methods are a natural successor of the results of this study that can assist in combining quantitative and qualitative measures of uncertainty". This seems a very important point but is not clear. Please, rephrase and clarify.

Technical corrections

Page 159 line 7: "bijective"

Page 167 line 2: Fig. 4 is the right one

Table A1: "efinitions"