Reply to Anonymous Referee #1

Note: We include our replies to the referee's comments directly at the respective points in the text. Referee comments are kept in italics and our replies are kept in normal font style.

1.1 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.

1.2 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.

The squared weighing is a commonly adopted weighting scheme for ordinal scaled data. In these data the neighbourhood of classes plays an important role as items that are allocated in neighbouring classes adhere to more similarity than when they are allocated to opposite ends of the scale. The weighting scheme considers these near-by characteristics and puts more weight on neighbouring classes.

We agree that the current section on the kappa test is overly technical in its presentation and will rewrite this section to clarify the relation to the data at hand. Accordingly we will revisit the presentation of the results.

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).

The referee is perfectly right in this observation. The application highlights 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. However, it does not yet provide a framework for information expansion that includes defined weights. This is subject for further research. We will consider this and rewrite this section accordingly, i.e. not allude to the section as a demonstration of the QAFs applicability rather than an illustrative example to highlight the potential of flood event documentation for understanding a specific flood event.

In the concluding section (pg. 176, lines 14-23) we discuss the next steps needed to develop a framework that formalizes the combination of information from many sources (i.e. combining quality labelled information from event reports with model or data based analysis) and under consideration of the uncertainties attached to each information. We will rewrite this paragraph to make this clearer.

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.

We will consider the referees comment and rework the concluding chapter. I.e. we will shorten the chapter substantially limiting the conclusions to only the most important findings with a German specific notation and rather add more generic aspects. These are:

- the use of QAF for providing the basis for a better ad-hoc and post event analysis. What are the critical factors that need to be analysed in the course of an event and what are critical considerations that need to be taken in the design of a report

- the contribution/addition of this study to event databases/catalogues and the improvement to an event set of floods by providing additional structured and quality labelled information.

- recommendations to report producers (better reporting can help improving capacities and organizational structures, as well as credibility)

- provide an outlook on future options for including event reports in research which are given by rapid technical and publishing developments (linked data, open access, semantic search options)

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.

The referee raises a very important point. In fact, the weighting scheme has been the most discussed point amongst the authors too. There are many possible points of view on a document's quality. Based on the framework by (Wang and Strong, 1996) the 4 quality criteria (QC) are the main pillars that define the overall quality of information (in their case data) from a users point of view. So, one option would be to give even weight to each of the QC. This however also means that the scores reached in each of the dimensions per QC will be averaged. The main argument for using the same weight for each of the dimensions was that it is most reflective of the chosen task at hand and therefore the user's perspective of our particular study. We accompany this choice by the notification that "It is important to note that P is not meant to label a document as per se bad or good 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 instance, this overall score needs to be accompanied by an analysis of scores reached in the single dimensions or combinations of dimension in order to identify the contextual scope of the document and its strengths and limitations. (pg. 153, lines 9-14)".

The way to proceed and therefore a field for further research will be a user survey in order to define those quality dimensions/categories that are most relevant and in order to derive any weights.

We will stress this important aspect more clearly and add it to the concluding chapter as a field of further research.

1.3 Specific comments

1.3.1 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.

We adapt this sentence according to the answer provided for major criticism no.2. I.e. we rephrase to: "Using an example flood event that occurred in October/ November 1998 we demonstrate the information from multiple reports can be synthesised."

1.3.2 Introduction

Comment 1: page 145 lines 7

What do you mean with "any systemic approach"? Any systemic approach to what?

We rephrase: "any systematic event analysis"

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.

We rephrase this sentence avoiding the usage of the term "contextual depth" in order to avoid confusion with the term that is later on used in a defined way. Instead of "However, the expectation towards the contextual depth of the documents seems to be rather unclear" we

rephrase: “However, the type and detail of information contained in the documents seems to be rather unclear”

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.

We will rephrase the entire paragraph from page 146, line 27 to page 147, line 7

“Evidence-based evaluations aim at synthesizing the available evidence for a given question (e.g., how effective are interventions in a river system 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).”

Comment 4: page 147 lines 23-24

What is “the environmental level (depicting the general Zeitgeist)”? Not clear

This is a term used in historic hydrology/climatology, however, its use is probably not so wide spread. We paraphrase this term rewriting the bracket to: “(depicting the general way of thinking and expression during an epoch)”

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

We remove the word ‘intrinsic’ from this sentence as we refer to the general quality and therefore need no specification here.

1.3.3 Methodology

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.

The sentence is repetitive to what is explained afterwards. We therefore is delete it.

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.

We include the dimensions in brackets so that the sentence reads: "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"

Comment 8: page 153 lines 4-6

"Assuming an average score QDi 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

We rephrase: 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 QD_i and the breaks are defined by the sum of scores reached by the QD.

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

In most cases the extent of a flood event will be larger than the one that is described in an event report (national and federal states or local restrictions). At this point in the paper we want to highlight, that we will assess the quality of each report with respect to its own spatial scope and not with respect to whether the reports is reflective of the entire flood extent.

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

Similar to the previous comment. Each report has its own objectives, i.e. one report maybe only focussing on the meteorological aspects, another only on damages, and another on the entire event. We want to evaluate the quality of each report in the quality that it reaches with respect to its own specific objectives.

1.3.4 Results

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

We reduce the sentence to "In the following we will discuss the agreements reached in the single QD", as this is the simple task we are about to perform.

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:

We mean the large amount of technical aspects that are covered within each dimension of the contextual quality category. We rephrase this sentence to: "Differences in the assignments are the result of some degree of subjectivity in drawing the distinctions between the score classes which will depend on the technical experience of a peer in any of the specific aspects like e.g. hydro-meteorology."

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 equalling 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.

Per document that was included in the peer review process, Fig. 2 plots the Pedigree scores that were given by the author against the pedigree scores that were given by the Peers. P is the ratio of the total sum of scores given per quality dimension divided by the maximum sum of scores possible (in our case: 10 dimension * max 3 = 30). The plot and example highlight that the peers and authors result for the quality assessment of any document are very close.

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.

We will include this as a notion in the methodology section.

Comment 15:

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

We add a note on that at the beginning of section 3.2.

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.

We rephrase to: "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."

Comment 17 page 168 line 20:

"83.5%". Is it correct? According to the table the right value is 84.2%

You are right. The percentage should be 84.2%.

Comment 18 page 170 line 13:

"GDR". What does it mean? Not defined before

We define the abbreviation. GDR – German Democratic Republic.

Comment 19 page 172 line 5:

"See section 3.2". Reference is not correct.

Reference to any section not needed here. Will be removed.

Comment 20 page 174 line 3:

"Q(T < 5a)". Is it an error?

It is correct. However we rephrase to: 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.

1.3.5 Discussion

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.

In the paragraph related to by comments 21 and 22 we want to highlight that our study provides a starting point for an improved understanding of flood events. In our case we use reports and provide a quality assessment scheme. 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).

1.4 Technical corrections

Page 159 line 7: "bijective"

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

Table A1: "efinitions"

The errors will be corrected.

References

Merz, R., and Blöschl, G.: Flood frequency hydrology: 2. Combining data evidence, Water Resources Research, 44, W08433, doi:08410.01029/02007WR006745, 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.

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 analysis, 25, 481-492, 2005.

Wang, R. Y., and Strong, D. M.: Beyond accuracy: what data quality means to data consumers, Journal of Management Information Systems, 12, 5-33, 1996.