Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-383-RC2, 2021

Comment on hess-2021-383

Anonymous Referee #2

Referee comment on "Building a methodological framework and toolkit for news media dataset tracking of conflict and cooperation dynamics on transboundary rivers" by Liying Guo et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-383-RC2, 2021

This study proposed a framework to extract news articles related to transboundary rivers from a large news paper database, and demonstrated the application of the framework. It is a timely topic and relevant to the topics concerned by HESS. However, I have several concerns regarding the quality of this study.

My major conern is on the significance of this study. The study constructed a framework to retrieve news texts related to transboundary rivers from a large database. The framework primarily relies on the term generator proposed by the authors to filter the news articles in the database, and the term generator is essentially a dictionary mapping of related terms. From my perspective, this study neither developped new cutting-edge techniques nor applied any advanced text mining techniques to resolve water resources issues, it looks more like a data preparation section of another paper rather than a standalone paper. In particular, the method is primarily based on term-based filtering, which is simple and straight-forward, and does not contribute much to current methodological studies; and the final product is a news database related to transboundary rivers, which only provides unstructured text data and does not directly provide any additional information, insights or solutions to any existing water resources issues. I would suggest the authors to work on two directions: (1) futher develop more structured database through extracting more information from the original news texts with advanced text mining tools; (2) develop a few relative simple cases to demonstrate the use of the data (i.e. implement some of the potential analysis.)

In addition to the major concern above, I also have a few minor concerns as listed below.

(1) The authors put too much potential impacts of this study as the significance of this study. I think the authors should clearly state what are the contributions of their work and what are the potential impacts.

(2) There are many subject judgements in the workflow of the framework proposed by the authors, for example, the determination of the terms and 5 blocks, the determination of "satisfactory keyworks" in line 88, the manual relevance checking. what is "the balance between relevance and coverage" in line 137, etc. How can the authors ensure the subject judgement are not biased?

(3) In section 2, the potential analysis is listed as one part of the workflow, but it is only talked on the conceptual level. It is to some extent misleading to be listed as one step of the framework.

(4) In line 111, the author mentioned only English newspapers are collected. I wonder that whether it will cause some bias. For example, if there are two countries along a transboundary river, but English is the official language of only one of the two countries; then the collected news will be unbalanced, and the contents may only reflect the comments from one country.

(5) Part of the study is developed based on TFDD, e.g., line 128, line 191 and line 186. Please discuss the difference between your work and TFDD, and what is improved.

(6) Line 206, how "5 time" is determined? Line 211, how to revise term frequency? Only based on subjective judgement?

(7) Line 221, please state clearly how the database sort the news articles by relevance, since you used the sorting function several times.

(8) Line 225-227. Only meta data are structured. It will be more helpful if the unstructured contents of the news can be somehow structured.

(9) Line 252, where LDA is used?

(10) As shown in your case study (figure 6), for some river basins the results are not acceptable (e.g., Columbia.) Have you evaluated how many occurances of such regions in all your retrieved data? Do you have any measures to control the quanlity of the data?

(11) In addition to transboundary rivers, is there any broader impacts of your study to the field of water resources and hydrology?

(12) Too many unnecessary details are provided in the major content of the paper. I would suggest the authors to write the main texts concisely.