Response to reviewer #1

We thank the reviewer for reviewing our study. Below we list the reviewer's comments in bold and discuss how we incorporated them in our paper.

This paper presents TAGGS, an innovative method to group natural hazards related Twitter tweets, which is very useful for the response and rescue after the natural hazards happen, mitigating the loss. Overall, this paper fits the interest of NHESS Journal; given the high-quality of its scientific innovation and writing, the paper deserves an acceptance, though some minor revisions are needed.

We thank the reviewer for his/her encouraging words and endorsement of our innovative approach.

The paper puts emphasis on its innovative geotagging algorithm, namely TAGGS, which basically is a method dealing with toponym recognition and resolution, especially for tweets. Although it does a great job reviewing related works, it overlooks some toponym recognition and resolution work on short texts, which could be useful for the case of tweets as well. Moreover, only those fields in the meta-data are considered as spatial indicators. What about the context in the tweet itself? For example, if a tweet mentions "Washington" and "president", it is very likely the "Washington" is referring to Washington D.C.. This could be the next step if the authors are going to further their approach. Here are two related literature that the authors may refer:

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., & McKenzie, G. (2016). Things and Strings: Improving Place Name Disambiguation from Short Texts by Combining Entity Co-Occurrence with Topic Modeling. In Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20 (pp. 353-367). Springer International Publishing.

Y Hu, K Janowicz, S Prasad (2014): Improving Wikipedia-based place name disambiguation in short texts using structured data from DBpedia, In Proceedings of 8thA CM SIGSPATIAL Workshop on Geographic Information Retrieval, Nov. 4,2014, Dallas, TX, USA.

We agree with the reviewer that these studies are indeed both interesting and valuable and would help to improve our approach in further research. Unfortunately, this is currently beyond the scope of our work and, therefore, we included both references in the section on future work and possible improvements (Sect. 4), which now reads like this:

> *In future work, we aim to continue improving our algorithm. Currently, using the approach described in this paper, we only parse each tweet using the spatial information from that tweet itself and from other tweets mentioning the same toponym. In future research, we plan to expand on this approach by detecting sudden changes in the number of mentioned locations in an area. This technique would allow us to improve the geoparsing algorithm by considering sudden increases in mentions of nearby locations, using such a peak as an additional spatial indicator. Other improvements could be made by taking into account additional context, such as entity co-occurrence (Hu et al., 2014; Ju et al., 2016) or the geography of Twitter networks (Takhteyev et al., 2012).*

Some parts of the writing could be clarified or improved: In line 24 of Section 2.2, the expression "tweets older than 24 hours" is confusing. Also, what is the reason to choose "24 hours" as the scanning window? What's the difference if I choose "6 hours" or "72 hours"?

In line with the reviewer's comments we have revised several sentences:

> *Once locations had been assigned to the tweets, the same procedure was applied to a later scanning window (Sect. 2.2.5 / Fig. 4), which included new incoming tweets. At that stage, tweets that are outside the scanning window were no longer considered. Meanwhile, new incoming tweets were immediately geoparsed using the toponym resolution table.*

and

> *All new tweets were retrieved from the tweet database and separately analyzed for toponyms and respective spatial indicators (Sect. 2.2.1 and 2.2.2), while tweets that fall outside of the scanning window were discarded.*

In addition, we have varied the size of the scanning window and included a detailed analysis in the paper. The text and figures now read as follows:

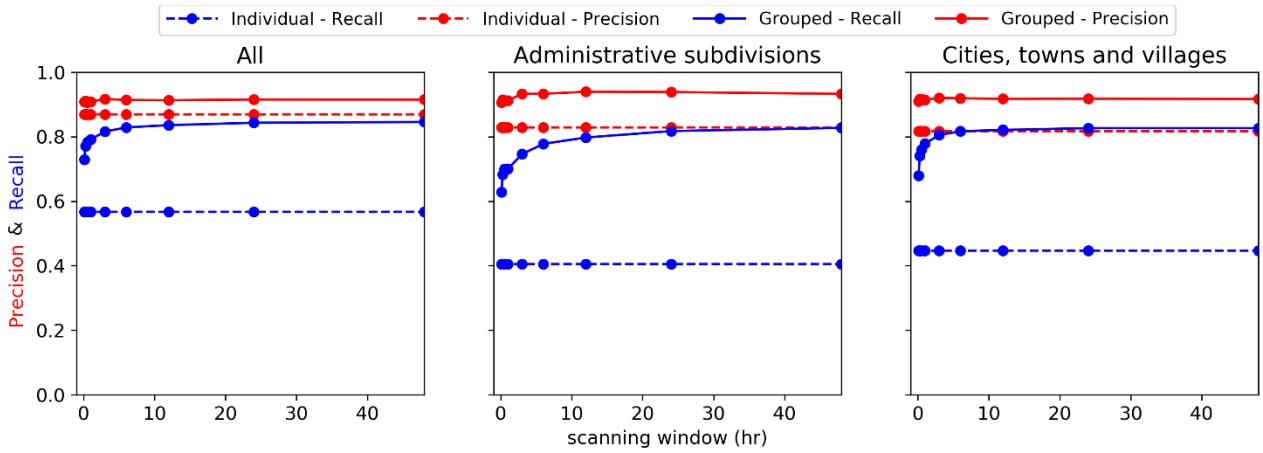

Figure 1: Recall and precision scores for individual and grouped geoparsing with a varying size of the scanning window.

> *Figure 1 shows the recall and precision measures for a varying scanning window size, ranging between 6 minutes and 48 hours. In theory, when using an infinitesimally small scanning window for grouped geoparsing, the results would be identical to the individual geoparsing. It is clearly visible that, in general, both precision and recall increase when the size of the scanning window is larger. This is expected, because a larger number of tweets are grouped and therefore, the likelihood that spatial information is available increases. Although an increase of recall and precision is still visible for a larger scanning window, the increase is not substantial, which indicates that spatial information is available for most toponyms. When new floods occur, it is not feasible to take location mentions of previous floods into account. Therefore, we hypothesize that when the scanning window becomes too large, the performance of the model will be lower. Unfortunately, because of memory (RAM) constraints in our current setup, we cannot test this.*

Ideally, the size of the scanning window depends on the volatility of the event type, where events with a longer average duration (people will likely refer to the same event over a longer timespan), such as droughts, could benefit from a larger scanning window and vice versa for shorter events.

It is nice to see "thresholds" are used to balance between precision and recall, but it seems like the authors only use "0" and "0.2". It would be better to see a precisionrecall curve, which is typical for the task of information retrieval.

We thank the reviewer for his/her valuable comment and have updated the analysis of the validation study. We now show a precision-recall curve for each administrative level for both singular and grouped geoparsing. The text and figures now read as follows:

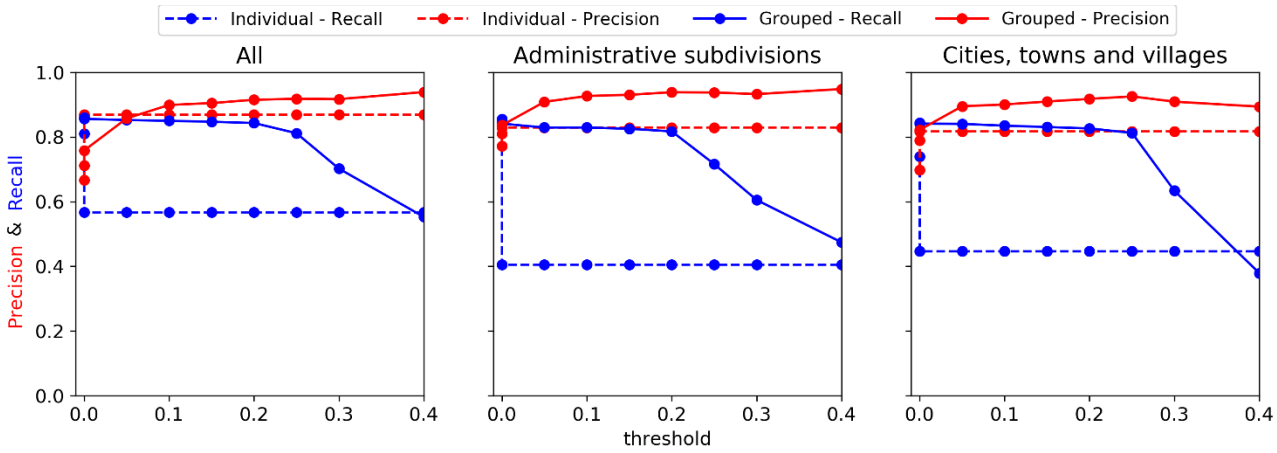

Figure 2: Recall and precision scores for individual and grouped geoparsing with a varying threshold.

Figure 2 shows the recall and precision scores for individual and grouped geoparsing with a varying threshold. The trade-off between precision and recall is visible in the first window: When a higher threshold is chosen, more location matches are discarded, while the likelihood of a correct match is higher for the residual locations. For individual geoparsing, as only the spatial indicators of the post itself are considered, the scores behave discreet. In contrast, for grouped geoparsing, the scores are averaged between tweets within the same group, and therefore the decrease is more gradual. At very high thresholds, the precision for grouped geoparsing starts to drop (for administrative subdivisions and cities/town/villages). This is likely because the scores assigned to tweets in small groups fluctuate more than for large groups (Sect. 2.2.4) and hence there is more uncertainty in the location being assigned correctly. Therefore, when the threshold increases, small groups have a larger share in the response set (as large groups will always have averaged medium scores) which causes the precision to drop. Approximately between a threshold of 0.1 and 0.25, precision and recall measures for grouped geoparsing are optimal and higher than using any other threshold for individual geoparsing.

In figure 3, for Toponym recognition, it should be 2.2.1, instead of 2.2.1.

We have updated the section reference accordingly.