Interactive comment on "Improving the understanding of flood risk in the Alsatian region by knowledge capitalization: the ORRION participative observatory" by Florie Giacona et al.

February 18, 2019

1 General comments

The paper presents a data set on regional river floods which is certainly a valuable contribution to natural hazards research. It is furthermore well structured. However, I have comments on the definition of variables and the statistical analyses. I think this can be a lot more informative if the associated sections are revised. In many of the analyses given, binary variables are related and in this case Pearson's correlation coefficient is not informative, nor can its significance be tested with a t-test. All these analyses need to be revised. The resulting paper will be a lot more valuable than the current version. See my detailed comments below.

2 Detailed comments

Comments are sorted according to the different sections and subsection.

2.1 Introduction

- p.2, l.8 "Statistical analyses" is the first concept mentioned in your introduction. Make it a valuable concept in your paper!
- p.2, l.8 "chronologies" is a synonym for time series?

2.2 Further filtering and additional event typologies

- Table 2. Is the "Multi-criteria severity scale for flood events" really as informative as it could be? There are initially 2 dimensions: "Damage level" (3 categories) and "Spatial extension" (3 categories). There are thus 9 possible combinations from which only 7 are attained. Using the "Severity" this is mapped onto 1 dimension with 4 categories. In the text, category "IV" is mapped onto "III". I cannot follow the argument why this reduction of information is made. 7 combinations are not so much. Please explain what is the reason behind this reduction to 3/4 categories.

2.3 3.4 Statistical analyses

This section apparently give the methodological background needed to understand and reproduce the statistical analyses made in this paper. Unfortunately, I see both goals not met. I have a rough idea what is going to happen with these methods and I fear that the results are prone to misinterpretation. The description is not sufficiently detailed to enable the reader to reproduce the research. I suggest to rewrite this subsection completely with more mathematical rigour, without Pearson's correlation but including conditional means, conditional probabilities, and, if you like, logistic regressin. The reason is given in the following section.

- p.10 l.16 "... was analyzed through a comprehensive set of basic statistical methods". I noted only *one* method, which is calculating Pearson's correlation coefficients. I agree with "basic" but not with "comprehensive".
- p.10 l.17 "each categorical variable". A categorical variable is a variable X which can take *one* value out of a set of categories, e.g. $X \in \{A, B, C\}$, compare `https://en.wikipedia.org/wiki/Categorical_variable`. This holds for your variables "hydrological configuration" and "severity class". The latter can take values from $\{I, II, III, IV\}$ and can be even ordered. It is thus a special case of a categorical variable. The other "variables" can take on more than one value,
  - p.7 l.25 "Causes. For each event, one or several causes (if any)" and is thus not a categorical variable,
  - p.7 l.26 "Damage types: human ..., material, functional, environmental, unknown. For each event, zero to four damage types can be documented" and is thus not a categorical variable,
  - "hydrological configuration", it is not completely clear to me, if more than one value can be attained.
- p.10, l.18 "each categorical variable ... was coded in terms of presence/absence". I don't think that is what you meant. A categorical variable can take values out of a set of categories. A binary variable can take on 0 (absence) or 1 presence. What are these binary variables now? My understanding is that you have groups of binary variables as follows
  - Group *causes*
    - Ice breakup (binary)
    - Ice jam (binary)
    - Snow melting (binary)
    - Heavy rainfall (binary)
    - ...
    - Unknown (binary)
  - Group *consequences*
    - Environmental (binary)
    - Functional (binary)
    - Human (binary)
    - Material (binary)
    - other (binary)

  This section needs to be more precise. Use mathematical definitions and equations.

- p.10, l.19ff. "Presence/absence codes were also given for each event in all municipalities". What does that mean? A municipality A gets assigned $A = 1$ if it was affected by the event? Please be more precise.

- p.10, l.20. "For the rivers, distinct codes were given, depending on whether water overflow was documented or not". What codes are given to the rivers? Is a river denoted as, e.g., "Rhine" and gets assigned a $Rhine = 3$ if there was an overflow? Please be more precise.

- p.10, l.22. "Also, visualization and analysis of the spatial characteristics of each event under a GIS environment was possible". I agree with "visualization".

- p.10, l.24. "All correlations between variables, rivers, municipalities, etc. were evaluated". I can see how a correlation between variables is evaluated but not how to obtain a correlation coefficient for rivers. What is meant? The variable "river length", "river flow", "water level"? Same question for "municipalities", is it the number of people living there? This needs to be more precise.

- p.10, l.24.ff The Pearson's correlation coefficient is easy to interpret for pairs from a bivariate normal distributed variable. If we want or not, this is what most of us have in mind when seeing this coefficient. I expect that you obtain Pearson's correlation coefficient not just for pairs from a bivariate normal distributed variable but also for other pairs, e.g. from ordered categorical variables as your "Severity class" or for binary variables as, e.g. *Cause*: Heavy rainfal (0/1), etc. I don't think it is clear, how to interpret Pearson's correlation coefficient in this case, neither is Student's t-test valid. This makes your analyses questionable.

- p.11, l.1 "percentage" $\rightarrow$ fraction

- p.11, l.1 "having occured each *calendar* month"

- p.22, l.2 "time trends" $\rightarrow$ large time-scale variations

- p. 22, l.4 "... for each series, the moving averages values" what are these values? How are they standardized? How can the standardized value be interpreted? This should be made more precise (and shorter) using equations.

2.4 Results

- p.11, l.15 "... all damage types are positively correlated with the 'source number' variable". What is the source number variables? This has not been explained in the previous sections. It appears from this section now that it is the number of sources reporting a given events, i.e. an positive integer variable including 0. Please explain this variable in the appropriate section before.

- p.11, l.16 "... Pearson correlation coefficient is 0.26 with human damage presence/absence, non-zero at the 0.05% significance level..." What does this number imply if we do not have pairs from a bivariate normal variable? What is "0.05%"? Do you mean a level of the test of 0.05? Or equivalently a level of 5%? Saying "non-zero", I expect you mean a two-sided test. However, as mentioned earlier, results of a t-test are questionable for non-normal input variables as the presence/absence variables. I suggest deleting this correlation analysis and

obtain conditional mean values instead. For source number S_i and human damage H_i for event i that implies

$$\bar{S}_{|\text{human damage}} = \frac{1}{\sum_{i=1}^{N} H_i} \sum_{i=1}^{N} S_i \, H_i, \tag{1}$$

with i counting the events from 1 to N, h_i being the binary variable for human damage associated with event i. This gives the conditional mean source number for events with human damage. This can be compared to the conditional mean for no human damage, e.g.

$$\bar{S}_{|\text{no human damage}} = \frac{1}{\sum_{i=1}^{N} (1 - H_i)} \sum_{i=1}^{N} S_i \, (1 - H_i), \tag{2}$$

If you like, you can obtain confidence intervals for these means using the central limit theorem or even t-test them for significant difference. Here, the t-test is very likely to hold as means become quickly normal distributed due to the central limit theorem, see e.g., Wilks [2011]. Alternatively, you can also show conditional distributions as histograms.

- p.11, l.18ff "There is also a very strong correlation between the source number variable and the extreme-sized events (class IV) and large-sized events (class III)." It appears here that "Severity" is not used as an ordered categorical variable with values from 1 to 4 but as a collection of binary variables. I suggest to use the same conditional mean as explained before for all four categories.

- p. 11, l.25ff Here, you obtain conditional probabilities for the different causes. That is very easy to understand and meaningful! You should use this kind of conditional analyses instead of correlation. As up to 4 causes can be selected for an event, you can also obtain probability estimates conditioned on two or more causes at the same time. Confidence intervals can be obtained with the help of the binomial distribution.

- Table 3. Total of right column is 100 not 1. Furthermore, I would call it "Fraction (%)" instead of "Percentage".

- p.13, l.1ff Again you relate two binary variables (e.g. classII and snow melting) with Pearson's correlation. Please change this in conditional probabilities as before. Estimate

  - probability of a class III event given snow melting
  - probability of a class III event given heavy rainfall
  - etc

  Vice versa, you can even estimate the probabilities of the various causes given a class III event. Both would be much more helpful than a correlation coefficient.

- p.13, l.30 Instead of correlations, I's suggest to estimate probabilities of common flooding. Replace Table 5 with joint probabilities instead of Pearson's correlation coefficient. As the tabel is symmetric, only the upper or lower triangle needs to be given. The space in the other triangle can be used e.g. for uncertainty information. There is a typo in the Table caption: *Bod*

- p.14, l.7ff Again, replace the correlation analysis with probabilities. This can be probably also nicely shown on a map.

- p.15, l.11ff Again, correlation should be replaced with conditional probabilities.

- I stop citing every case where Pearson's correlation coefficient should not be used. It should be replace in all cases using binary variables by either the conditional mean or conditional probabilities.
- p.18, l.2 and l.14ff. Here we have Pearson's correlation for class III events (binary) and time. If you want to show an increase/decrase in the probability for class III events in time, logistic regression [e.g., Wilks, 2011] would be a good choice. Same holds for the change in probability of causes and damage types.
- p.19, l.3 Floods of the Ill and total number of floods can be studied with conditional means of total number of floods conditioned on Ill floods, see above.

All through Section 4, the authors should replace the correlation analyses with either conditional means, conditional probabilities or logistic regression as indicated above. I did not explicitly mark every occurrence which needs replacement. This should have became clear from the examples discussed.

References

D. S. Wilks. *Statistical methods in the atmospheric sciences*. Academic Press, San Diego, CA, 3rd edition, 2011.