Referee 3

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.

Authors' response: Thank you very much for your generally positive appreciation of our work, your positive and constructive feedback and your detailed review. We agree that some points in the variable definition and statistical analysis sections need to be precised and this will be done in the revised version of the paper. Below are our detailed responses to all the comments and questions raised.

Let us however make first a general answer regarding the statistical methodology. We do not base our answer on Gilks et al. (2011) that is sadly not available to us but on the references at our hand.

Most important thing is that for binary variables, computation of standard metrics such as spearman rank correlation, Kendall tau and Pearson cross product-moment coefficient lead exactly to the same result (this can be checked by developing the formulas). The result gives a measure of the link between the two variables that one indeed should refer in a generic way as "association" (how the variables behave together or not) instead of "correlation" which may have the more restrictive sense of the strength of the potential linear link. For a couple of variables with one binary and one continuous variable or one binary and one discrete unbounded variable, the point-biserial coefficient is the standard choice, which can also be rewritten at the Pearson cross product-moment coefficient. Hence, there is nothing wrong with using Pearson cross product-moment coefficient in such cases as a measure of association/correlation, and other standard choices arguably provide exactly the same result.

For categorical variables with more than two values, however, this is no longer true (which may be what the reviewer has in mind), and one may need to look at other metrics such as the Tetrachronic correlation which does not exactly correspond to the Pearson cross product-moment coefficient (even if it remains close). This is not required for us because our variable definition is made to avoid this more difficult case.

The second point is how to test the significance of this association/correlation, and this is a trickier question. Arguably (and surprisingly), this may even be a kind of grey area in statistical science because despite a deep search it was not completely possible for us to get a complete definite answer. Yet, what is clear is that having one or two binary variables within the Pearson cross product-moment coefficient does not blur the convergence of its estimate to asymptotic normality on the basis of standard theorems (Central limit theorem, Slutsky theorem, etc.). As a consequence there is nothing wrong in principle in using the standard t-test as we do in such cases. However, what is true is that in such cases convergence to asymptotic normality of the estimate with sample size may be slow (and clearly much slower than in the optimal case of an underlying bivariate Gaussian population) which may be why the t-test is often (not always, there is no consensus) referred as inappropriate in such case. On the other hand, when applying tests, nobody generally cares about the convergence to the asymptotic distribution of the estimate and simply apply it to get the yes/no answer… More practically, we have rather large samples at hand (366 events, more than 500 years of data) so this may not be really an issue…

Yet, the problem comes when considering that, for binary variables, the same association measure may be seen, as said before, either as spearman rank correlation, Kendall tau and Pearson cross product (and similarly, with one binary and one continuous variable or one binary and one discrete unbounded variable, as the point-biserial coefficient or as the Pearson cross product coefficient). And these different quantities have different asymptotic distributions for their estimates providing different p-values and hence different answers to the same question of stating if the variables are associated or not. In full rigor, one should therefore try different tests in each case (as one should try different tests for the identity of the mean of two Gaussian distributions…). However, due to the very high amount of couples of variables we consider and to the scope of the paper (see below) we prefer not doing that. This would indeed increase the complexity and length of the paper without much added value with regards to the scope of the paper.

Finally, we indeed want to stress that, even if the statistical analysis is an important point of the paper, it is only one among different aspects of our work which also includes the constitution of the database with a mixed historical-participative approach. The latter leads a very large data set and the aim of the statistical analyze if only to demonstrating that the data overall makes sense and to highlight its main characteristics in space and time. The size of the analyzed data set makes that only a very simple strategy like ours can be systematically applied to all event characteristics and time series. For instance, we fully agree than more advanced approaches base on statistical modelling using e.g. logistic regression would be much more appropriate to answer specific questions such as, e.g. linkages between binary and continuous variable, but doing this systematically and interpret all related results would clearly expand the paper length beyond reasonable standards.
It was already stated in the conclusion and outlook section of the paper: "However, only a rapid analysis of this abundant information was performed. Hence, further research could clearly exploit the data in many directions, for example, explicit statistical modeling with advanced nonseparable spatiotemporal covariance models instead of the rather simple descriptive techniques that were used in this work. Coupled with the diachronic analysis of land use, settlements, and practices, this may facilitate further understanding of the main pattern of change in flood risk", but we will make sure in the revised version of the paper that this message is even more explicit.

All in all it seems to us that the statistical debate goes far beyond what we want to show in the paper and even beyond the discussions that the standard readership of NHESS may expect to find in the journal. We also believe that our methodological approach, even simple, is sufficient to reach want we want to show. To avoid ambiguities, we will in the reworked version of the paper shortly justify our methodological choice with the arguments detailed above, replace "correlation" by "association" as much as possible and say even more clearly that further work should ground on the data but chose (much) more advanced statistical modelling strategies to answer specific questions.

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!
Authors' response: According to our main comment, statistical analysis is not at the hearth of this paper, and it is also a bit restrictive at this stage. This will be changed into "analyses of long data series is a prerequisite for any effective management of natural hazards".

p.2, l.8 chronologies is a synonym for time series?
Authors' response: In a strict sense, yes, it is the temporal distribution of past events, but for us it goes a bit beyond since from the point of view of the historical analysis the rough time series comes with various elements allowing to put past events within their bio-physical and societal contexts. See, e.g., Giacona et al. NHESS 2017). This will be precised in the revised version of the paper.

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.
Authors' response: Thank you for this point. First, we really wanted to have a measure that combines spatial extent and damage level to asses flood severity, which led us define 7 possible combinations. This is not that much, but these different combinations are somewhat difficult to rank (is a regional flood with slight damage worse or less severe than a local flood having caused strong damages?). Also, scales usually used in the risk / natural hazard worlds have 5 levels, not seven, which is a bit too much. We therefore mapped the 7 combinations into 3 or 4 categories. This is not a huge reduction of the complexity, but it already makes things easier for the analysis. Especially, the obtained severity class is arguably really an increasing function of flood severity (class II more severe than class I, etc.). By the way, the damage and spatial extension scores for each event have obviously been kept, so that they remain usable for further work including specific analyses within more detailed categories

(e.g. only regional extent damageable events, localized damageable events, etc.). This will be further precised in the reworked version of the paper.

2.3 3.4 Statistical analyses
This section apparently gives 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.
Authors' response: see main comment at the beginning of our answer.

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 coeffcients. I agree with basic but not with comprehensive.
Authors' response: We agree that this sentence was a bit cumbersome. "Comprehensive" was for us not related to the set of methods used (arguably mostly an association analysis, even if the computation of smooth underlying trends can be considered as another method, why not?) but to the whole set of variables it was applied: the quantitative and presence/absence variables (including rivers and municipality) both in the event table (correlation between events) and in the chronology table (time series table). This leads to a huge amount of information arguably sufficient to highlight the main pattern of the very large amount of data at our hand. This point will be further precised in the reworked version of the paper.

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={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,
Authors' response: We agree that our use of "categorical" may be seen as insufficiently precise from a strict mathematical point of view. But in fact we only consider binary variables in our analyses (which are indeed specific categorical variables) since we transform all information such as cause, damage type, etc. in a vector of presence (1)/absence(0) variables whose length depend on the considered group (e.g. 5 for the severity class, the four class and unknown, etc.). This will be precised in the reworked version of the paper.

p.7 l.25 Causes. For each event, one or several causes (if any) and is thus not a categorical variable,
Authors' response: OK, this will be changed (see previous comment, we have as much binary variables as possible causes, plus the unknown case).

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,
Authors' response: ibid.

hydrological configuration, it is not completely clear to me, if more than one value can be attained.
Authors' response: No one single configuration per event, either Rhine, Tributaries or both. This will be slightly reworked for more clarity, which leads four binary variables (the three configurations and the unknown case, with only one of these taking the value 1 and the other zero).

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.
Authors' response: Again, this simply means that we transform all information such as causes, damage types, etc. in a vector of presence (1)/absence(0) variables whose length depend on the considered group (e.g. 5 for the severity class, etc.). This will be reworked for more clarity but we think that there is no need to introduce mathematical notations / equations for such simple operations.

p.10, l.19. 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.
Authors' response: Yes, exactly, this will be precised in the reworked version of the paper.

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., Rhin and gets assigned a Rhine = 3 if there was an overflow? Please be more precise.
Authors' response: Again, we transformed the information into as many variables as necessary. For each event we have a list of rivers which have been affected and for each of them we know if water overflow was documented or not. As a consequence, for each event we have the variable "overflow for Rhine" which takes the value 1 if this was documented, 0 if not, "flood without overflow for Rhine" which takes the value 1 if this was documented, 0 if not, and ""flood with or without overflow for Rhine" which takes the value 1 if Rhine was flooded during the event and 0 in not. And this is done for the 13 rivers we specifically consider plus for a binary variable corresponding to all other rivers specifically known in the target area as well as for unknown rivers (if any). Also, for each event, we count among the 13 rivers we focus on, the number of rivers with overflow, of rivers without overflow and of flooded rivers in total. At the end of the day, for each event, we have therefore (13+2*)*3=45 binary variables plus 3 positive discrete variables. This will be precised in the reworked version of the paper.

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.
Authors' response: Analysis here refers to the fact that it was then possible to do the correlation analysis and the computation of all time-series. We agree that it does not refer here to specific GIS treatments, so that the word "analysis" will be removed in the reworked version of the paper

_ 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 coeffcient 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.
Authors' response: This refers to the correlation between the presence / absence between couples of rivers or couple of municipalities in the event table. For rivers, the analysis could be done for specific binary variables referring to overflow, flood without overflow or both. For simplicity, only the latter case was considered in the paper. This analysis highlights to which amount couples or rivers or municipalities tend to be affected by the same flood events which is exactly the aim of an association analysis.

_ p.10, l.24._ The Pearson's correlation coeffcient 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 coeffcient. 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.
Authors' response: see main comment at the beginning of our answer.

p.11, l.1 percentage: fraction

Authors' response: Thank you, this will be changed in the revised version of the paper.

p.11, l.1 having occured each calendar month
Authors' response: Thank you, this will be changed in the revised version of the paper.

p.22, l.2 time trends: large time-scale variations
Authors' response: Thank you, this will be changed in the revised version of the paper.

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.
Authors' response: Simply, for each time series (e.g. number of events per year in the river Rhine as function of time, number of type III events per year as function of time, etc.) we computed 51- and 31-year moving averages. In the plots, rather than superposing the raw counts and these moving averages, we rescaled the latter as follows: for each series, the moving average values were multiplied by the ratio between the raw annual count standard deviation and the moving average standard deviation. This gives both the row series and rescaled moving average series the same variance, i.e. the same range of variability on the plots, allowing easier visualization. Only the shape of the rescaled moving average series is analyzed which is exactly the same as the one of the non-scaled moving averages series (increase, decrease, etc.). On the other hand, we do not analyze the quantitative rescaled values in terms of trends (+0.xxx event per year, etc.) as this should indeed be done with the non-scaled series. This will be clarified in the reworked version of the paper. Again, we do not think it is really necessary to introduce equations within the text core for such simple operations.

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.
Authors' response: Thank you, this will be included in the 3.2 Section of the revised version of the paper.

_ p.11, l.16 _... Pearson correlation coe_cient is 0.26 with human damage presence/absence, non-zero at the 0.05% signi_cance 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 Si and human damage Hi for event i that implies … (see latex doc) with i counting the events from 1 to N, hi 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. 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.
Authors' response: See main comment at the beginning of our answer. And yes, 0.05% refers to a level of 5% of the t-test, this will be precised in the reworked version of the paper.

p.11, l.18 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.
Authors' response: see main comment at the beginning of our answer.

p. 11, l.25 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.

Authors' response: See our main comment and our response about categorical/binary variables.

Table 3. Total of right column is 100 not 1. Furthermore, I would call it Fraction (%) instead of Percentage
Authors' response: Thank you, this will be changed in the revised version of the paper.

p.13, l.1 Again you relate two binary variables (e.g. class II and snow melting) with Pearson's correlation. Please change this in conditional probabilities as before.
Authors' response: see main comment at the beginning of our answer.

probability of a class III event given snow melting
Authors' response: see main comment at the beginning of our answer.

probability of a class III event given heavy rainfall
Authors' response: see main comment at the beginning of our answer.

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 coeffcient.
Authors' response: see main comment at the beginning of our answer.

p.13, l.30 Instead of correlations, I's suggest to estimate probabilities of common flooding.
Authors' response: see main comment at the beginning of our answer.

Replace Table 5 with joint probabilities instead of Pearson's correlation coefficient. As the table 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
Authors' response: Regarding the choice of a correlation analysis, see our main comment at the beginning of our answer. We agree that only half of the table is required due to symmetry and we will do that in the revised version of the paper, but this is not going to save space because of the diagonal structure of the resulting half table. Thanks for the typo, it will be corrected.

p.14, l.7 Again, replace the correlation analysis with probabilities. This can be probably also nicely shown on a map.
Authors' response: see main comment at the beginning of our answer.

p.15, l.11 Again, correlation should be replaced with conditional probabilities.
Authors' response: see main comment at the beginning of our answer.

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.
Authors' response: see main comment at the beginning of our answer.

p.18, l.2 and l.14. Here we have Pearson's correlation for class III events (binary) and time. If you want to show an increase/decrease 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.
Authors' response: see main comment at the beginning of our answer.

_ p.19, l.3 Floods of the III and total number of floods can be studied with conditional means of total number of floods conditioned on III floods, see above.
Authors' response: see main comment at the beginning of our answer.

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.
Authors' response: see main comment at the beginning of our answer.