Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-250-AC1, 2016

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Estimating extreme river discharges in Europe through a Bayesian Network" *by* Dominik Paprotny and Oswaldo Morales Nápoles

Dominik Paprotny and Oswaldo Morales Nápoles

d.paprotny@tudelft.nl

Received and published: 25 August 2016

We would like thank the referees for the time spent in reviewing our article and their valuable comments. Below, we list all the comments and our response.

Section 2.4: The choice or the Gaussian copula could be better justified, e.g. are there any physical explanations?

-> There are no physical explanations; the choice of the type of copula is based purely on the dependency structure.

Mention in the main text the two types of copulas you are testing as well.

-> Names of the other copulas (Gumbel, Clayton) will be added to the main text.

Briefly explain in the supplement, why you choose the selected types for comparison and not other/further representatives of no/lower/upper tail dependence. Are there any physical explanations? The test only shows, that Gaussian performs better than the gumbel and clayton, but it might still be a bad choice. Supplement 3 even indicates that the Gaussian copula is not a very good choice.

-> The three copulas used in the study have, primarily, the advantage of having one parameter which is obtained directly from the observational data. In this way, the method is more flexible. Naturally, the choice is not perfect, though this is not different to numerous possibilities in regression or extreme value analyses. These copulas cover basic asymmetries from data: upper and lower dependence. Since this is a first approximation of the joint distribution pointing to possible asymmetries observed in the data was considered sufficient. Future research could be directed to exploring other asymmetries in bivariate distributions in which case the copula families to be tested should be larger.

I do not understand how you get and use the conditional rank correlations for continuous distributions. The references you provide only give a short explanation and refer to further literature again. Is there some standard literature on the definition of conditional rank correlations for continuous distributions? To my understanding the conditional rank correlation depends on the state of additional parents, yet in fig. 3 you only give single numbers (no conditioning is visible). Yet, if the rank correlation is independent of the states of the other parents, you miss to model the joint effects and could use a naive Bayes instead (which would be far more simple than going all the way over BNs and copulas). On the other hand, if the conditional rank correlation depends on other parents states, there must be a way to calculate it for each possible conditioning state (to be able to perform inference) or it must be determined for a discrete approximation of the conditioning variable (which increases the number of required parameters significantly; similar as for using discrete BNs from the beginning).

Maybe you could comment on this in the discussion forum.

-> Conditional rank correlations are calculated as shown in eq. 4, except that the conditional distributions are used inside the arguments to the right of the equal sign. For the Gaussian copula conditional correlations are equal to partial correlations and these are constant. Hence the importance of validating the choice of the Gaussian copula. We thank the referee for noticing this fact, we will make it more evident in the text. The full explanation could be found in the publication we refer to in that paragraph – Hanea et al. 2015; it contains the theorems (3.1, 3.2) as well as the proof (Appendix A). That proof shows that in a non-parametric Bayesian Network the joint distribution is uniquely determined. Therefore, the correlations shown in Fig. 3 are all conditional on their parents. As we note further in the methodology in the main text, conditioning is done by sampling the BN. This procedure is shown by Hanea et al. 2006; unfortunately, the reference got lost in the original submission: Hanea, A. M., Kurowicka, D., Cooke, R. M.: Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets, Qual. Reliab. Engng. Int., 22, 709–729, doi:10.1002/qre.808, 2006.

Section 2.6: I find this subsection difficult to read and suggest to revise the sentence structure of this section.

-> The section will be revised and reorganized for clarity.

Why do you use 30-year time periods? What would be the effect of using shorter/longer time periods?

-> 30-year periods were used, because 1) it maximises the number of stations for validation and 2) it is the most commonly used time period in climate research, and was also used in the research project for which this method was developed. If we use a 50-year period (1951–2000), we can only validate the model on 378 stations (one-third of the current validation data set). Yet, the results for Q_100 calculated from 50 years of data are: Rˆ2 = 0.92 and NSE = 0.90, which is slightly better than using 30 years of data.

Please mention the distributions you are testing in addition to GEV. GEV performs best compared to what?

-> There were 15 other distributions analysed; we will mention a few most relevant in the text (generalized Pareto, gamma, lognormal or Weibull)

Section 3.1, p.13, l. 5: "2-year discharge has the same performance as Q_MAMX" <- where does this information come from?

-> To align better the text to the graphs, we will correct this information to “10-year discharge has almost the same performance as Q_MAMX”

Why do you suddenly have 4 different time periods? Section 2.6 mentions only 3.

-> The 1981–2010 was indeed not mentioned earlier. It was only used to analyse the results, and not to produce the final database. In the revision, this will be mentioned in section 2.6.

The regional performance seems to depend strongly on the number of stations used per region.

-> We do not believe there is a link between regional performance and number of stations. For instance, we can break down further the regions into countries, and this shows that, when comparing modelled and observed Q_MAMX, e. g. Sweden (139 stations) has Rˆ2 of 0.71, while Norway (101 station) has 0.90, which in case of the former is worse than the value for Scandinavia, and the latter is better. Finland (36 stations) has Rˆ2=0.88, and included in “other regions” (75 stations), where Rˆ2 equals 0.79.

You mention several times ’the model performance remains acceptable’. What is your understanding of acceptable?

-> We indeed use ‘acceptable’ two times in the text. We used Moriasi et al. (2007) as a reference. Information about this will be added to the text.

How do you explain the better performance of the BN quantified from the smaller dataset of 917 records?

-> The difference is at most 0.01 in Rˆ2 or NSE, so it is really negligible.

Page 18: You might extend your comment on using different types of copulas: How suitable is the Gaussian to model all interactions? How well does it fit the data (are there objective measures)? Would it be possible to use different types of copulas in the same BN and thus find a better description of each interaction? Which other types of copulas could be useful to check? What do you expect, to which extend could the model be improved, by using different types of copulas?

-> We will add to the mentioned paragraph a summary of answers to the reviewer's questions, which are addressed in other places in the manuscript and especially in the supplement. Briefly:

1) We test the joint normal copula assumption using the determinant of the rank correlation matrix (sec. 3 of the supplement);

2) We show the statistical analysis of the results in section 3.1 and 4.1, concluding that the fit is good in contrast to other studies and methods;

3-5) Other copulas could potentially be used, as for some distributions tail dependence and other asymmetries may be present. Even though the normal copula works well most of the time. Skewness for example may be modelled by copulas based on mixture distributions. This would correspond to copulas with more than two parameters [Joe 2014]. There is surely potential for improvements by more detailed analysis of the dependency structures in future research we will highlight this fact in the discussion section.

Figure 3 and supplement 4: Why do you use a discrete BN in the shown examples for inference? This does not correspond with your objective to find/use a continuous BN with a low number of parameters. The discrete conditional distributions you show in the

supplement, are not smooth. I guess, this should not happen, if you stick to continuous representations.

-> Empirical distributions are not smooth. A discrete BN uses conditional probability tables, while the class of continuous BN we use require empirical one-dimensional margins and rank correlations. The visual impression that the marginal distributions are discrete is merely caused by the representation of the marginal empirical distributions as histograms, as noted below Figure 3.

Page 9, l. 3: to be precise, the actual number of required conditional probabilities/parameters is a bit smaller, since some parameters result from probability theory (the parameters that describe a distribution for a specific condition have to sum up to one)

-> We thank the reviewer for pointing it out; the following remark will be added to the text:

"If a node has 7 parents, as it happens in the BN described in the next section, and it is discretized into 5 states, a probability table with 5ˆ8 = 390,625 conditional probabilities would be required. 5ˆ7 = 78,125 may be estimated by difference, as probabilities must add to 1. Thus, 312,500 probabilities would need to be specified. Similarly, if we were to discretize into 10 states each node 90,000,000 probabilities would need to be specified. Even a discretization into 5 states for each node in our model would make the quantification prohibitive given the data available. Considering other nodes, which also have parents, would make it even more restrictive for the use of discrete BNs."

Typos and minor issues: p. 6, l.30: the influence of DIFFERENCES models; p. 8, l.14: need to explained; p. 8, l.22: a set of noTes and arcs; p. 9, l. 1: which is the actually case; p. 9, l.28: Hanea 2006 is missing in the list of references; p.10, l.27: values these climate variables; p.11, l. 9: This variable is influence; p.11, l.11-12: check complete sentence; p.11, l.21: allowed to performed; p.15, l. 8: median return periods are show; p.17, l.12: I would not consider this fact as "evidenced", but rather as indicated; p.18, l.

3: Potential incorporation different time spans; p.18, l.21: non-Gaussian copula would a better model; p.20, l. 6.

-> We thank the reviewer for the detailed listing of smaller mistakes. All listed typos will be fixed in the manuscript.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-250, 2016.