Diagnosis of underlying assumptions regarding the Non-Parametric Bayesian Networks

1 Introduction

In this document we provide additional details related to the methodology of Non-Parametric Bayesian Networks (NPBN) described in section 2.4 of the paper. Firstly, we present a diagnosis of the copula models for our BN in order to justify the use of the Gaussian copula in the NPBN. Secondly, we describe the procedure and results of the validation of the graphical structure of the BN. Additionally, we give more examples of conditionalization of the model mentioned in section 2.5.

2 Copulas

Representing probabilistic dependence through a bivariate copulas requires selecting one of the many copula types. A detailed review is presented by Joe (2014). Here, we investigate three of the most popular copula types in order to determine which one is best in representing the joint distribution of variables included in the BN model of discharges. First, the Gaussian copula, which has the following cumulative distribution function:

$$C_\rho(u,v) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right), (u,v) \in [0,1]^2 \quad (1)$$

where Φ is the bivariate Gaussian cumulative distribution and ρ is the (conditional) product moment correlation between the two marginal probability distributions u and v in the interval [0,1]. Second, the Gumbel copula, parameterized by δ:

$$C_\delta(u,v) = \exp\left\{-\left([-\log(u)]^\delta + [-\log(v)]^\delta\right)^{1/\delta}\right\}, \delta \geq 1 \quad (2)$$

Third, the Clayton copula, parameterized by α:

$$C_\alpha(u,v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-\alpha}, \alpha \in [-1,\infty) \quad (3)$$

In contrast to many other types of copulas, these copulas require one parameter. These copulas model an important aspect of joint distributions known as tail dependence. The upper tail dependence coefficient λ_U for two random variables X and Y is:

$$\lambda_U = \lim_{u\to 1} P\left(X > F_X^{-1}(u) | Y > F_Y^{-1}(u)\right) = \lim_{u\to 1} P(U > u | V > u) \quad (4)$$

Roughly, a value of $\lambda_U > 0$ indicates that it is likely (more than normal) to observe values of U greater than u given that V is greater than u for u arbitrarily close to 1. Lower tail dependence would be defined similarly as eq. 4, but then for the lower quadrant of the joint distribution. The Gaussian copula presents no tail dependence $\lambda_U = 0$, while Clayton

presents lower tail dependence $\lambda_U = 2^{-\frac{1}{\alpha}}$ and the Gumbel copula presents upper tail dependence $\lambda_U = 2 - 2^{\frac{1}{\delta}}$. The investigation of these copulas covers a range of dependence structures that are usually observed in data.

Apart from a visual inspection, we employ two measures in order to advise on the copula best representing a particular bivariate distribution. Firstly, we compute semi-correlations, an approach suggested by Joe (2014). The semi-correlations are the Pearson's product moment correlation coefficients computed in the upper and lower quadrants of the normal transforms of the original variables. For positive correlation, semi-correlations in the upper right (*NE*) and lower left (*SW*) quadrants are:

$$\rho_{ne} = \rho(Z_1, Z_2 | Z_1 > 0, Z_2 > 0) \quad (5)$$

$$\rho_{sw} = \rho(Z_1, Z_2 | Z_1 < 0, Z_2 < 0) \quad (6)$$

where (Z_1, Z_2) are the standard normal transforms of (X, Y). For negative correlation, semi-correlations in the upper left (*NW*) and lower right (*SE*) quadrants are ρ_{nw} and ρ_{se} are defined similarly (Joe 2014). In general, larger absolute values of the semi-correlations than the "overall" correlation indicate tail dependence.

As the second diagnostic tool we utilize one of the test statistics in the "Blanket Test" discussed by Genest et al. (2009), which is the Cramèr–von Mises statistic (*M*). The test statistic of interest for a sample of length *n* is computed as follows:

$$M_n(\boldsymbol{u},) = \sum_{|\boldsymbol{u}|} \left\{ C_{\hat{\theta}_n}(\boldsymbol{u}) - B(\boldsymbol{u}) \right\}^2, \boldsymbol{u} \in [0,1]^2 \quad (7)$$

where $B(\boldsymbol{u}) = \sum 1(U_i \leq \boldsymbol{u})$ is the empirical copula and $C_{\hat{\theta}_n}(\boldsymbol{u})$ is a parametric copula with parameter $\hat{\theta}_n$ estimated from the sample. Notice that the statistic is the sum of squared differences between the empirical copula and the parametric estimate. If the correlation is negative, we compute the *M* statistic for Gumbel and Clayton with the rotated copula. The results of the two measures applied to the variables of our Bayesian Network are presented in Table S1, while graphs for a few selected cases are shown in Fig. S1.

Table S1. Semi-correlations and "Blanket Test" statistic for all pairs of variables used in the Bayesian Network for extreme river discharges. Lowest *M* values are bolded.

X	Y	ρ	ρ_{ne}	ρ_{sw}	ρ_{nw}	ρ_{se}	*M* gumbel	*M* gaussian	*M* clayton
Area	MaxDischarge	0.82	**0.78**	**0.52**	0.30	0.07	**0.011**	0.026	0.535
Area	Steepness	-0.59	-0.28	0.01	-0.46	-0.22	0.213	**0.089**	0.190
Buildup	Lakes	-0.14	-0.25	0.25	-0.16	0.05	1.746	**1.543**	1.614
Buildup	Marshes	-0.16	-0.28	0.17	-0.19	0.06	3.688	**3.164**	3.704
Buildup	MaxDischarge	0.16	0.24	0.17	0.26	-0.19	0.283	**0.280**	0.293
Buildup	RunoffCoef	-0.33	-0.17	-0.04	-0.28	-0.03	0.528	**0.211**	0.158
Buildup	Steepness	-0.38	-0.18	-0.11	-0.26	0.11	0.372	**0.061**	0.241
Lakes	Marshes	0.57	0.33	0.41	-0.30	-0.15	4.852	4.617	**3.550**
Lakes	MaxDischarge	0.29	-0.16	0.14	0.38	-0.05	1.865	1.527	**1.282**
Lakes	RunoffCoef	0.27	0.18	0.00	-0.18	0.14	0.691	**0.727**	0.947
Lakes	Steepness	-0.37	-0.05	-0.09	-0.33	-0.01	1.586	**0.397**	1.029
Marshes	MaxDischarge	0.31	-0.21	0.10	0.43	-0.05	3.644	3.127	**2.734**
Marshes	RunoffCoef	0.34	0.25	0.08	-0.16	0.04	1.791	**1.724**	1.872
MaxDischarge	MaxEvent	0.14	-0.20	0.16	0.03	0.21	0.159	0.083	**0.045**
MaxDischarge	RunoffCoef	0.15	-0.22	0.19	0.00	0.19	0.190	0.072	**0.027**
MaxDischarge	Steepness	-0.28	-0.33	-0.06	-0.09	-0.10	0.158	**0.135**	0.203
MaxEvent	Steepness	0.61	0.30	0.15	0.07	0.25	0.136	**0.115**	0.777

RunoffCoef	Steepness	0.30	0.13	-0.03	0.16	0.18	0.032	**0.017**	0.156

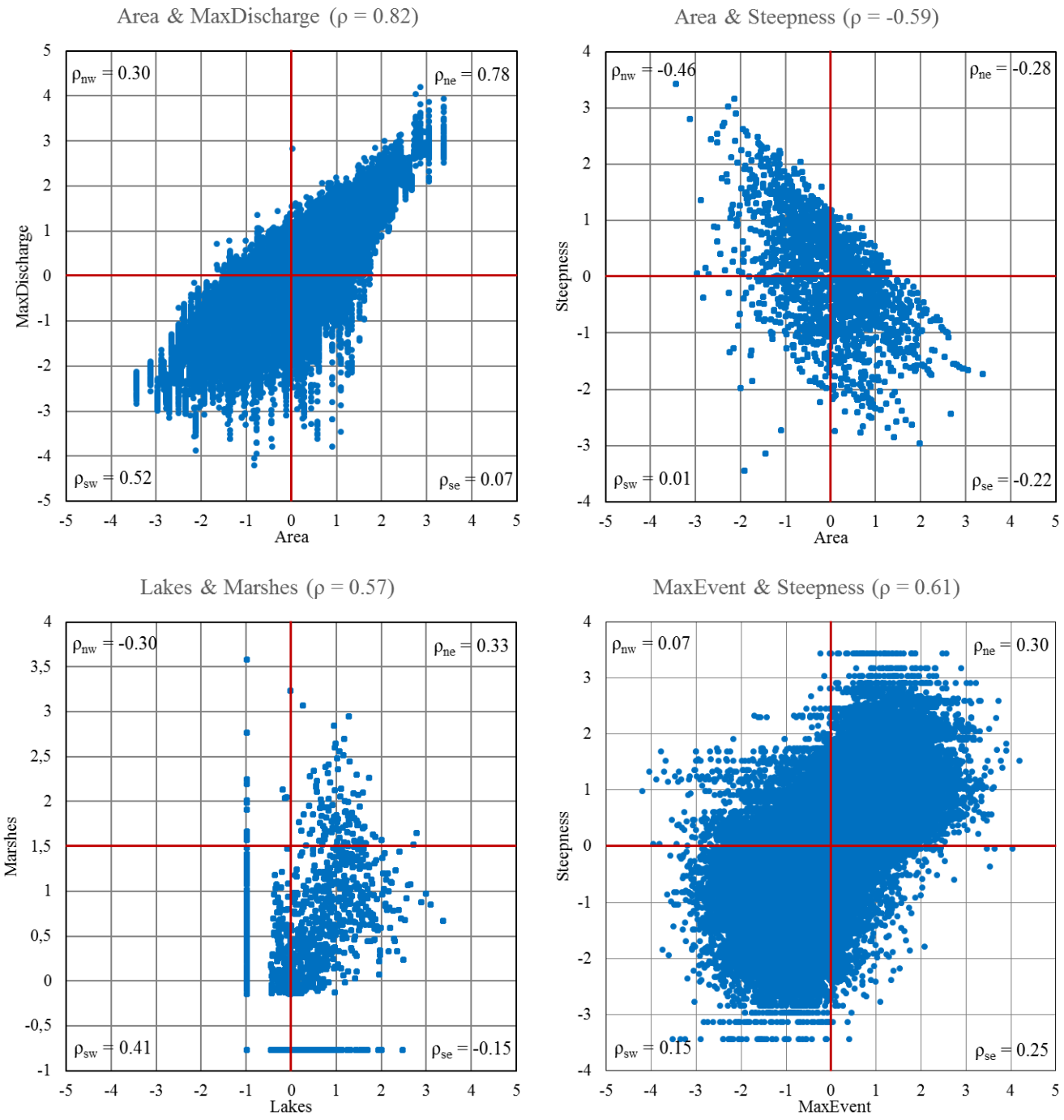

Figure S1. Graphs of selected pairs of variables of the Bayesian Network. The values of variables were transformed to standard normal, with correlation indicated for the whole sample and for each quadrant.

Analysis of the results indicates that the Gaussian copula is a good representation for most bivariate pairs of variables. This is indicated by relatively small differences in semi-correlations and low values of *M* statistic for the Gaussian copula. The difference between the empirical and parametric copulas is the smallest if Gaussian copula is used for 12 out of 18 pairs of variables included in the BN (examples: Fig. S1b and S1d). The *M* statistic indicates the Clayton copula as the best one for 5 pairs (e. g. Fig. S1c), and only for one pair – Area & MaxDischarge – the Gumbel copula gave the best result (Fig. S1a). In case of 3 of 5 pairs for which the *M* statistic indicated the Clayton copula as best-fitting (MaxDischarge & RunoffCoef, MaxDischarge & Steepness, Lakes & Marshes), the difference is small with respect to the same value for the Gaussian copula. Also, the difference in semi-correlations indicates only slight tail dependence, hence Gaussian copula is still a valid assumption. In summation, the results point towards the Gaussian copula as a suitable assumption for most of the bivariate distributions in the Bayesian Network for extreme river discharges. In our data, the variables most clearly displaying tail dependence (upper) is the Area & MaxDischarge pair.

3 Validation of the Bayesian Network

The Bayesian Network is constructed and validated in terms of accuracy of the results it produces in the main text of the paper. We however also verify to what extent the assumption of joint normal copula is valid. For that purpose, the determinant of the rank correlation matrix can be used. A rank correlation matrix is created by calculating the rank correlation between every possible pair of variables. The determinant is equal to 1 if all variables are independent, and equal to 0 if there is linear dependence between variables that have been transformed to standard normals.

The determinants can be utilized in two ways. Firstly, we calculate the determinants of the empirical rank correlation matrix (DER) and empirical normal rank correlation matrix (DNR). The former is obtained by transforming the marginals to uniforms and then calculating the product moment correlation of the transformed variables, while the latter is obtained by transforming the marginals to standard normal and then transforming the product moment correlations to rank correlations according to the following formula:

$$r(X,Y) = \frac{6}{\pi}\arcsin\left(\frac{\rho(X,Y)}{2}\right) \tag{8}$$

Those determinants will, in general, be different because the empirical copula will typically be different from the normal copula. Therefore, we can analyse whether the determinant of DER is within the 90% confidence bound of the determinant of DNR. If it does, that shows that a joint normal copula is a reasonable assumption. Secondly, the same comparison could be done between DNR and the determinant of a rank correlation matrix for a non-parametric Bayesian Network using normal copula (DBN). For details on this methodology we refer to Hanea et al. (2015).

In case of our BN, DER remained within the 90% bound of DNR if no more than ca. 310 samples were drawn in the procedure. DNR was within the 90% bound of DBN for up to ca. 400 samples. Those are a relatively small values, indicating

that the joint normal copula may not be the best assumption. However, Hanea et al. (2015) notice that the test is severe for large datasets, and the BN for extreme discharges contains 75,000 samples of each variable.

The rank correlation matrices can also be analysed not only in terms of determinants, but also through calculation of the so called d-calibration score (Morales Nápoles et al. 2013):

$$d(\Sigma_1, \Sigma_2) = 1 - \sqrt{1 - \eta(\Sigma_1, \Sigma_2)} \tag{9}$$

$$\eta(\Sigma_1, \Sigma_2) = \frac{|\Sigma_1|^{1/4} |\Sigma_2|^{1/4}}{\left|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2\right|^{1/2}} \tag{10}$$

where Σ_1, Σ_2 are the correlation matrices of interest. This score is a measure of "closeness" between two correlation matrices. The score is 1 if the matrices are equal and 0 if one matrix contains a pair of variables perfectly correlated, and the other one does not, and the score will be "small" as the matrices differ from each other elementwise. The distance between the empirical and empirical normal rank correlation matrices is within the uncertainty bounds if more than 800 samples are drawn. In case of the distance between the empirical normal and normal rank correlation matrices, it is within the uncertainty bounds if less than 250 samples are drawn. This confirms the results of the test based on the determinants.

4 Conditionalizing the Bayesian Network

To provide additional visualisation of the BN's conditionalization described in section 2.5, in Fig. S2–S4 we present three different states of the BN. The example is the same as in Fig. 4 of the paper, i.e. river Rhine at Basel station in Switzerland in year 2005. In Fig. S2 the Bayesian Network is unconditional. In Fig. S3 it is conditionalized on two variables (area and steepness); it can be seen how the distributions of all variables have changed. In Fig. S4 seven nodes were used to conditionalize the network, providing a better estimate of discharges.

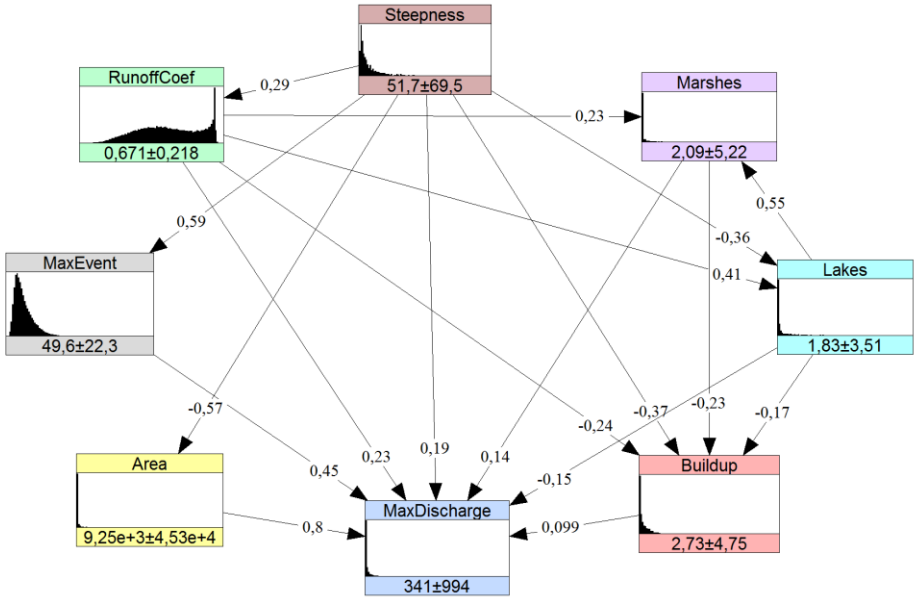

Figure S2. Unconditional Bayesian Network

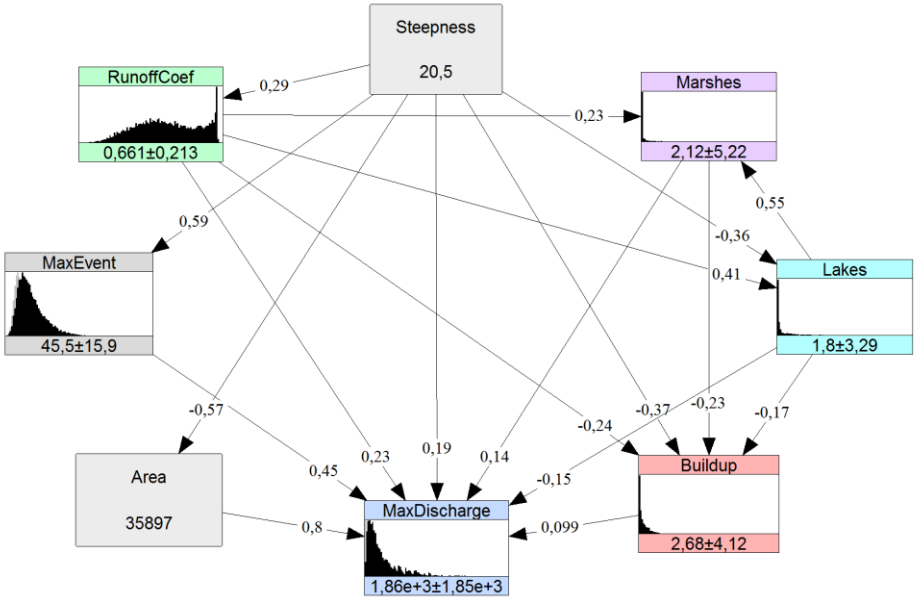

Figure S3. Bayesian Network conditionalized on two variables.

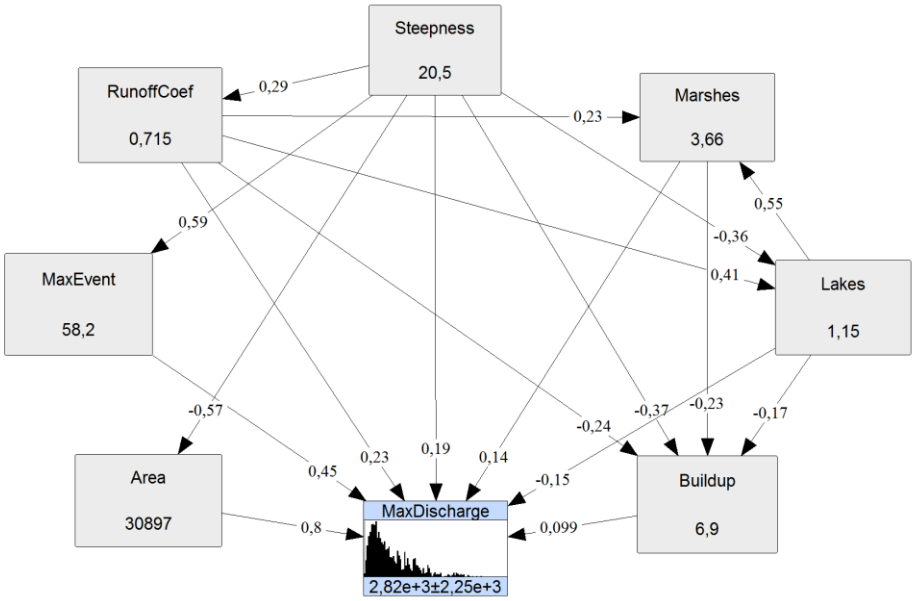

Figure S4. Bayesian Network conditionalized on seven variables.

References

Genest, C., Rémillard, B., Beaudoin, D.: Goodness-of-fit tests for copulas: A review and a power study, Insur. Math. Econ., 44, 199–213, doi:10.1016/j.insmatheco.2007.10.005, 2009.

Hanea, A. M., Morales Nápoles, O., Ababei, D.: Non-parametric Bayesian networks: Improving theory and reviewing applications, Reliab. Eng. Syst. Safety, 144, 265–284, doi:10.1016/j.ress.2015.07.027, 2015.

Joe, H.: Dependence Modeling with Copulas, Chapman & Hall/CRC, London, 2014.

Morales Nápoles, O., Hanea, A. M., Worm, D. T. H.: Experimental results about the assessments of conditional rank correlations by experts: Example with air pollution estimates, in: Steenbergen, R. D. J. M., van Gelder, P. H. A. J. M., Miraglia, S., Vrouwenvelder, A. C. W. M. (Eds.), Safety, Reliability and Risk Analysis: Beyond the Horizon, CRC Press/Balkema, Leiden, The Netherlands, 1359–1366, 2013.