Supplementary Material for "Multivariate statistical modelling of extreme coastal water levels and the effect of climate variability: a case study in the Netherlands"

Victor M. Santos[1,*], Mercè Casas-Prat[2,*], Benjamin Poschlod[3,*], Elisa Ragno[4], Bart van den Hurk[5], Zengchao Hao[6], Tímea Kalmár[7], Lianhua Zhu[8], and Husain Najafi[9]

[1]Department of Civil, Environmental and Construction Engineering, and National Center for Integrated Coastal Research, University of Central Florida, Orlando, Florida, USA
[2]Climate Research Division, Science and Technology Directorate, Environment and Climate Change Canada, Toronto, Ontario, Canada
[3]Department of Geography, Ludwig-Maximilians-University, Munich, Germany
[4]Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands
[5]Deltares, Delft, The Netherlands
[6]College of Water Science, Beijing Normal University, Beijing, China
[7]Department of Meteorology, Faculty of Science, Institute of Geography and Earth Sciences, Eötvös Loránd University, Budapest, Hungary
[8]Key Laboratory of Meteorological Disaster, Ministry of Education; Joint International Research Laboratory of Climate and Environment Change, Nanjing University of Information Science & Technology, Nanjing, China
[9]Department of Computational Hydrosystems, Helmoltz Centre for Environmental Research-UFZ, Leipzig, Germany
[*]These authors contributed equally to this work.

Correspondence: V.M. Santos (vmalagon@Knights.ucf.edu)

Introduction

This Supplementary Material is structured in three sections. Section 1 covers relevant aspects of copula theory to better understand how the 2D and 3D copulas are obtained in this study. Section 3 presents a theoretical example to illustrate the impact that conditioning the predictors on the impact variable has on the correlation coefficient and dependence patterns. Finally, Section 3 includes supplementary figures (S1-S14) that were referred to in the manuscript.

1 Copula theory

Sklar (1959) describes the connection between a Copula C and a bivariate cumulative distribution function (CDF) $F_{XY}(x,y)$ of any pair of variables (x,y) as:

$$F_{XY}(x,y) = C[F_X(x), F_Y(Y)] \tag{1}$$

where $F_X(x)$ and $F_Y(y)$ are the univariate marginal distributions. The bivariate probability density function (PDF) has the following form:

$$f_{XY}(x,y) = C[F_X(x), F_Y(Y)] f_X(x) f_Y(y) \tag{2}$$

where $f_X(x)$ and $f_X(x)$ represent the marginal PDFs. Let u and v be uniformly distributed random variables defined as $u = F_X(x)$ and $v = F_Y(y)$, then the function $c(u,v)$ (sometimes referred to as the copula density function) is given by:

$$c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v} \tag{3}$$

from which sampling can be performed through a Monte Carlo procedure , obtaining synthetic sets that preserve the dependence structure of the original data. In higher dimensions, a joint probability distribution is obtained via pair-copula constructions, i.e., Vine Copulas (Aas et al., 2009; Schepsmeier et al., 2018). This construction is hierarchical in nature and provides higher flexibility than other multivariate distribution functions since different dependence structures between pairs of variables can be adopted. Assuming the joint CDF continuous with strictly increasing marginal CDFs, vine copulas allow for the decomposition of an n-dimensional copula density into the product of $n(n-1)/2$ bivariate copulas. A decomposition example of a three-dimensional $f(x_1,x_2,x_3)$ is given as:

$$f(x_1,x_2,x_3) = f_{3|12}(x_3|x_2,x_1) f_{2|1}(x_2|x_1) f_1(x_1) \tag{4}$$

where:

$$f_{2|1}(x_2,x_1) = c_{12}(F_1(x_1), F_2(x_2)) f_2(x_2) \tag{5}$$

$$f_{3|12}(x_3|x_2,x_1) = c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) f_{3|2}(x_3|x_2) \tag{6}$$

$$f_{3|2}(x_3,x_2) = c_{23}(F_2(x_2), F_3(x_3)) f_3(x_3) \tag{7}$$

$$\tag{8}$$

Therefore, $f(x_1,x_2,x_3)$ can be expressed as:

$$f(x_1,x_2,x_3) = f_3(x_3) f_2(x_2) f_1(x_1) \text{ (marginals)} \tag{9}$$

$$\cdot\, c_{12}(F_1(x_1), F_2(x_2)) f_2(x_2) * c_{23}(F_2(x_2), F_3(x_3)) f_3(x_3) \text{ (unconditional pairs)} \tag{10}$$

$$\cdot\, c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) f_{3|2}(x_3|x_2) \text{ (conditional pair)} \tag{11}$$

$$\tag{12}$$

2 Impact of predictors conditioning on correlation

Here we present a simple theoretical test to illustrate the impact that the predictors' conditioning (on the impact variable) has on the kendall's coefficient (τ) (Kendall, 1938). Let A and B be two variables with standard normal distribution representing daily values of climate drivers. Let C be the impact variable driven by A and B by means of this simple impact function $C = A + B$. The dependence pattern between A and B is modelled by a Gaussian copula with associated τ ranging from -0.9 to 0.9. We define the predictand (impact variable) as the annual maximum of C, noted as $C_{\max}$, and the conditioned predictors as the values of A and B, respectively, when $C_{\max}$ occurs, noted as $A_{C_{\max}}$ and $B_{C_{\max}}$.

Figure S1 shows the scatter plots of randomly generated 50-year time series for each case and the corresponding conditioned predictors (in red). The corresponding empirically estimated τ are also indicated. We can observe that when we condition on the impact variable (for which both drivers positively contribute) we extract a sub-sample of the drivers realizations for which the correlation experiences a negative shift (as compared to the original sample). It can be observed, for instance, that for independence drivers (τ between A and B equals to zero, framed case in Figure S1) we obtain a a negative τ between the corresponding $A_{C_{\max}}$ and $B_{C_{\max}}$. From the latter we could not conclude, therefore, that the drivers are negatively correlated (i.e. the probability of concurrent large values of both A and B being lower than what we would randomly get by chance). In that particular case, the drivers are actually independent. Only in the cases with very strong positive correlation between A and B, the associated τ between $A_{C_{\max}}$ and $B_{C_{\max}}$ remains positive (in the example of Figure S1 for $\tau \geq 0.6$ between A and B). Therefore, a weak correlation between $A_{C_{\max}}$ and $B_{C_{\max}}$ does not necessarily imply a weak correlation between the underlying drivers. In fact, we argue that, for a given case, if the τ between $A_{C_{\max}}$ and $B_{C_{\max}}$ is larger than the τ between $A_{C_{\max}}$ and $B_{C_{\max}}$ as obtained from the corresponding independent case (A and B with the same marginal distributions but being the dependence pattern between them removed), then the underlying drivers A and B have a positive dependence pattern (i.e. concurrent large values of A and B are more likely to happen than by chance).

3 Figures

Included here are additional Supplementary Figures S1-S12 that were referred to in the manuscript.

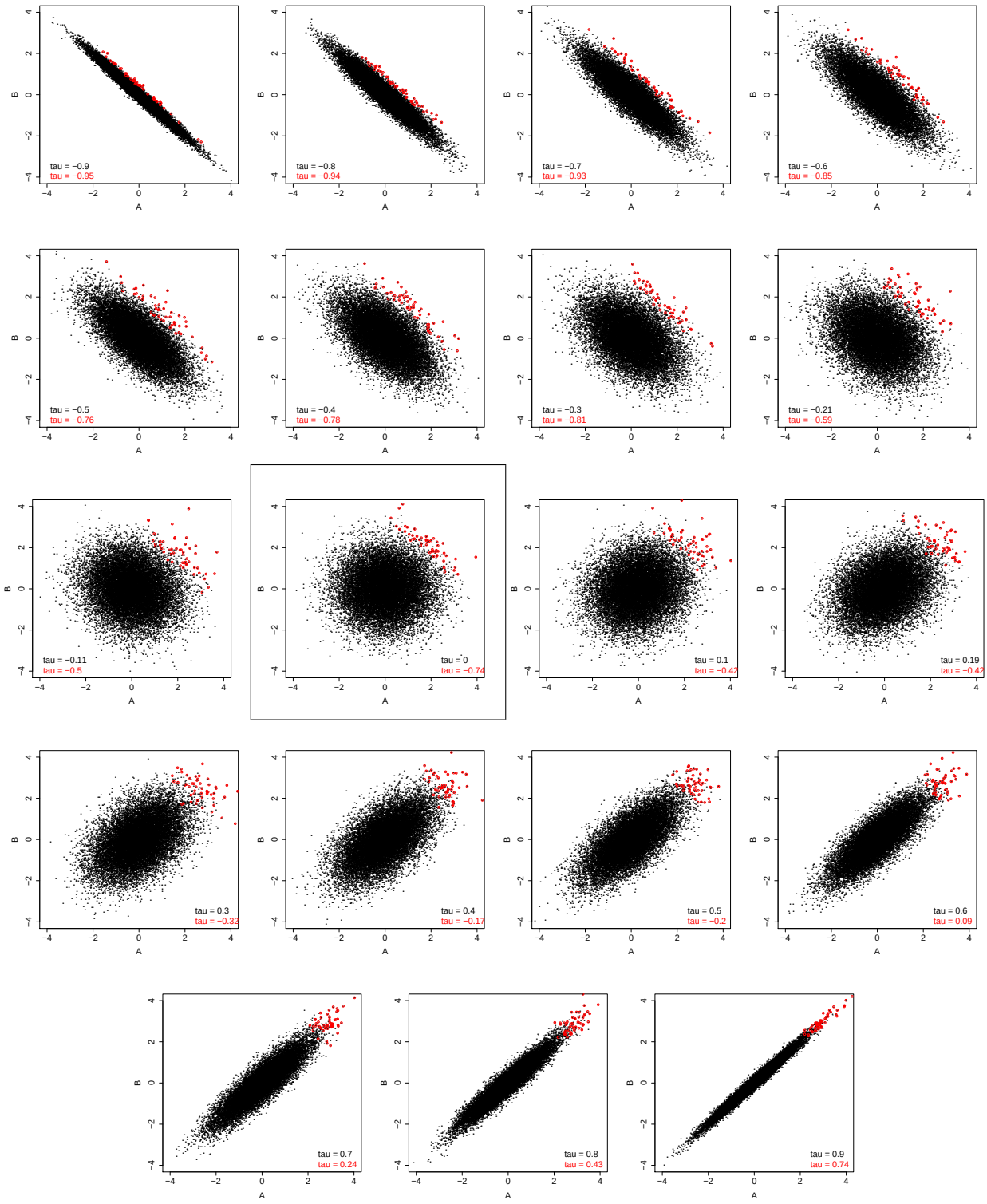

Figure S1. Scatter plot and estimated kendall's τ correlation coefficient for simulated drivers (A and B) for different degrees of dependence (in black), and for conditioned predictors ($A_{C_{\max}}$ and $B_{C_{\max}}$) (in red). The independent case ($\tau = 0$ between A and B) is highlighted with a frame for reference.

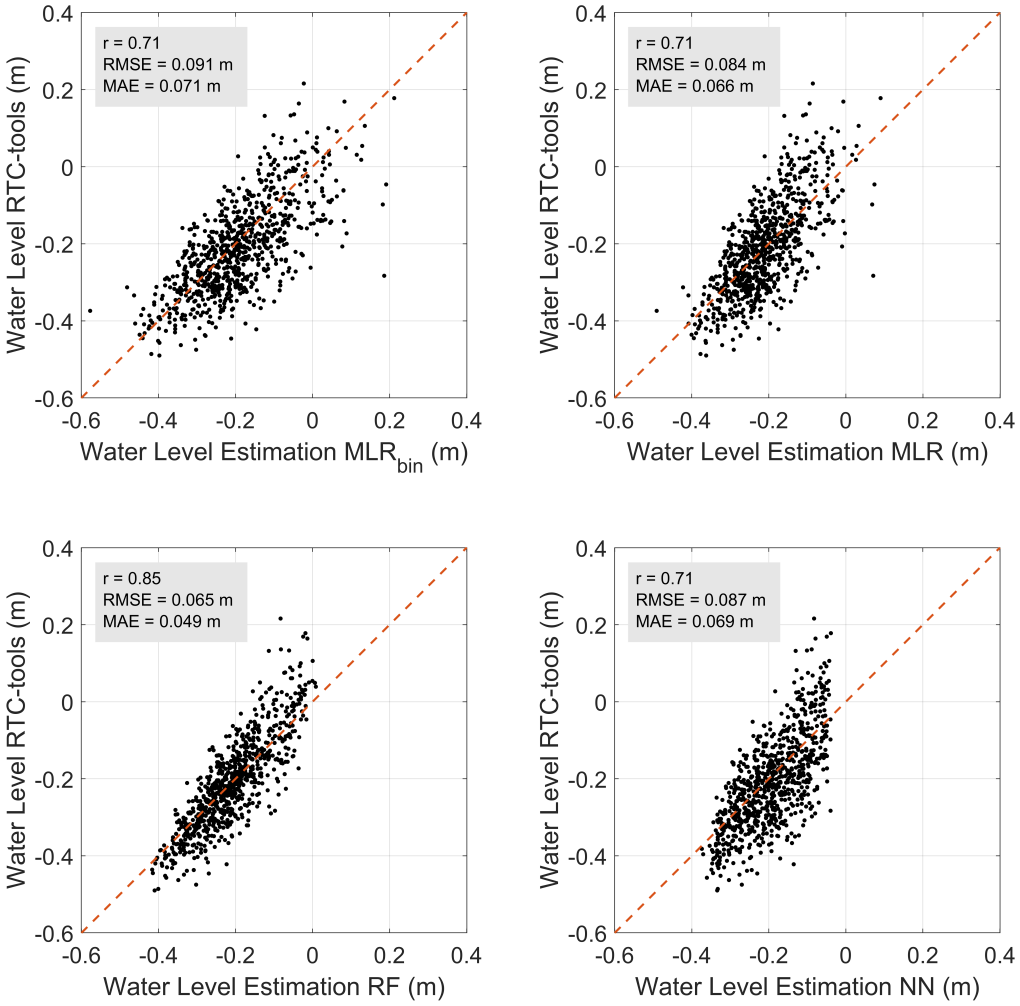

Figure S2. $WL_{\max}$ obtained by RTC-Tools vs. $WL_{\max}$ obtained using the following impact functions with $S^{T}_{36h,\min}$ and $P_{12d,\mathrm{acum}}$ (2D case): Multiple Linear Regression with bin-sampling ($\mathrm{MLR_{bin}}$), Multiple Linear Regression (MLR), Random Forest (RF) and artificial Neural Networks (NN).

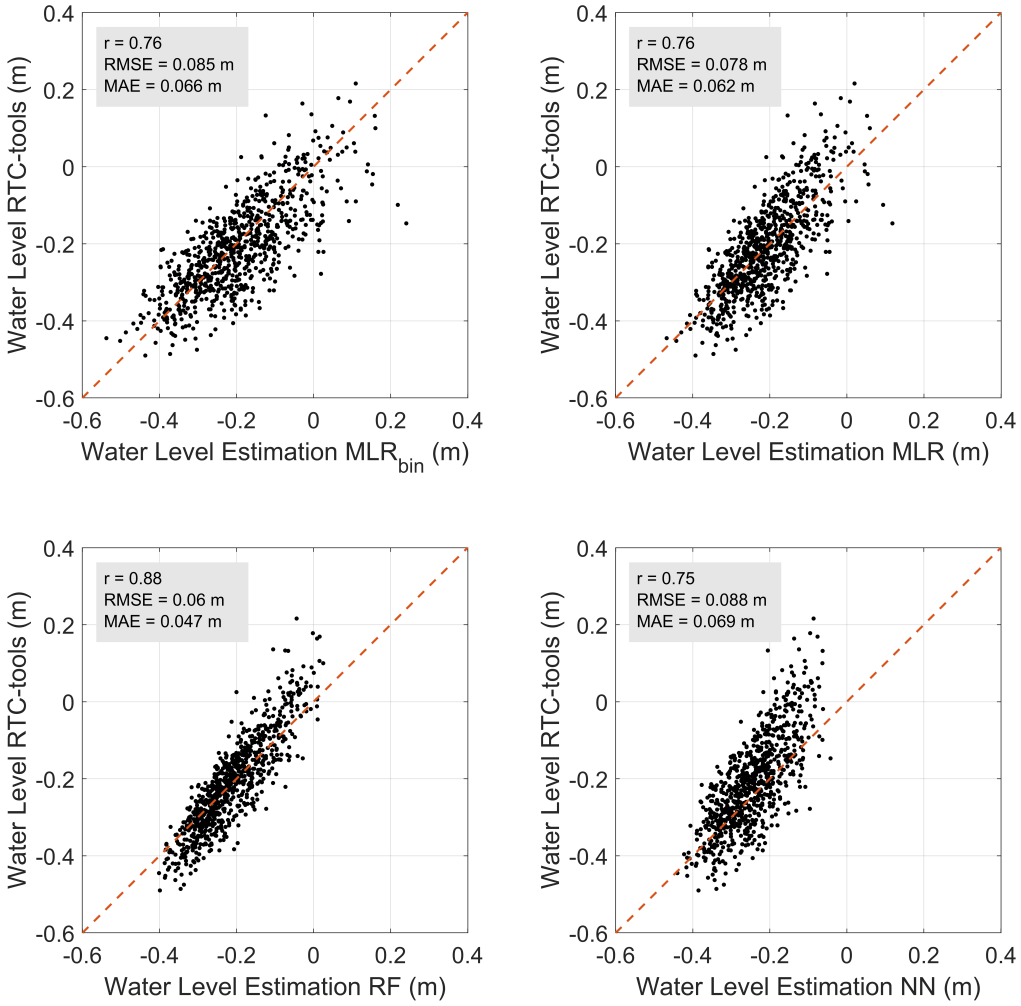

Figure S3. $WL_{\max}$ obtained by RTC-Tools vs. $WL_{\max}$ obtained using the following impact functions with $S_{72h,\text{mean}}$, $T_{12h,\min}$ and $P_{12d,\text{acum}}$ (3D case): Multiple Linear Regression with bin-sampling (MLR_{bin}), Multiple Linear Regression (MLR), Random Forest (RF) and artificial Neural Networks (NN).

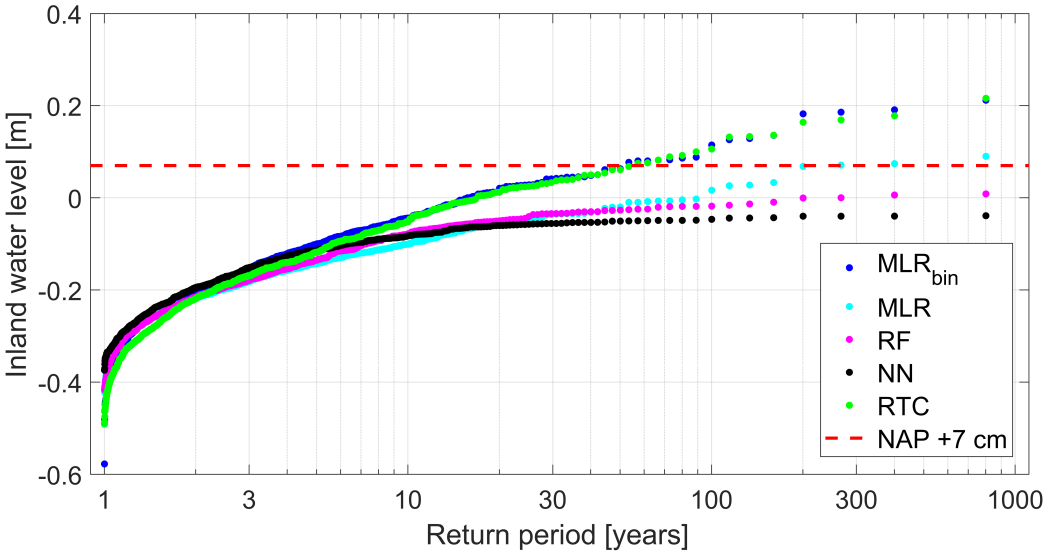

Figure S4. Inland return water level as generated from the RTC-Tools (green) and as obtained from the 2D case using the indicated impact functions: Multiple Linear Regression with bin-sampling (MLR_{bin}), Multiple Linear Regression (MLR), Random Forest (RF) and artificial Neural Networks (NN). The red dashed line (NAP + 7 cm) represents the flood warning level.

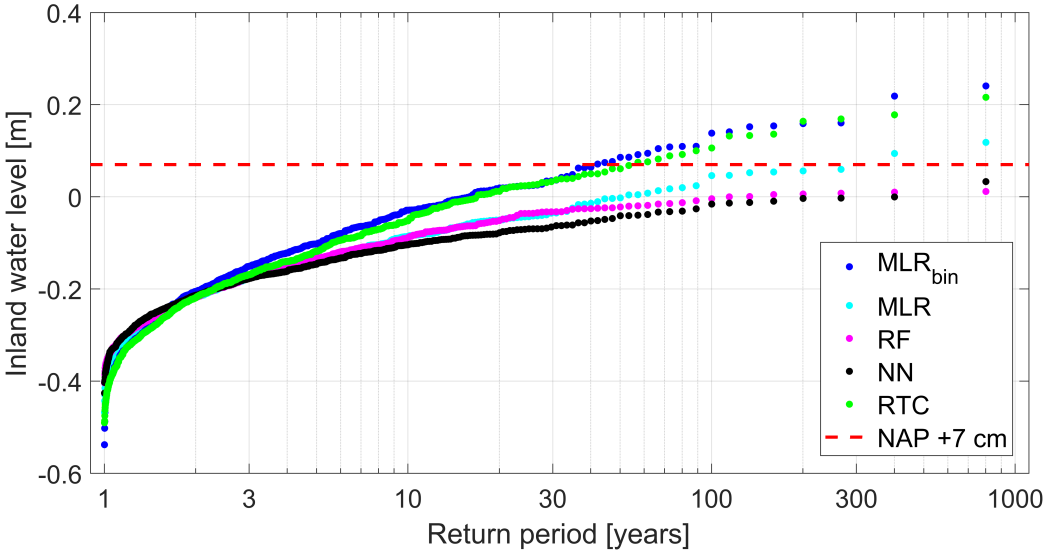

Figure S5. Inland return water level as generated from the RTC-Tools (green) and as obtained from the 3D case using the indicated impact functions: Multiple Linear Regression with bin-sampling (MLR_{bin}), Multiple Linear Regression (MLR), Random Forest (RF) and artificial Neural Networks (NN). The red dashed line (NAP + 7 cm) represents the flood warning level.

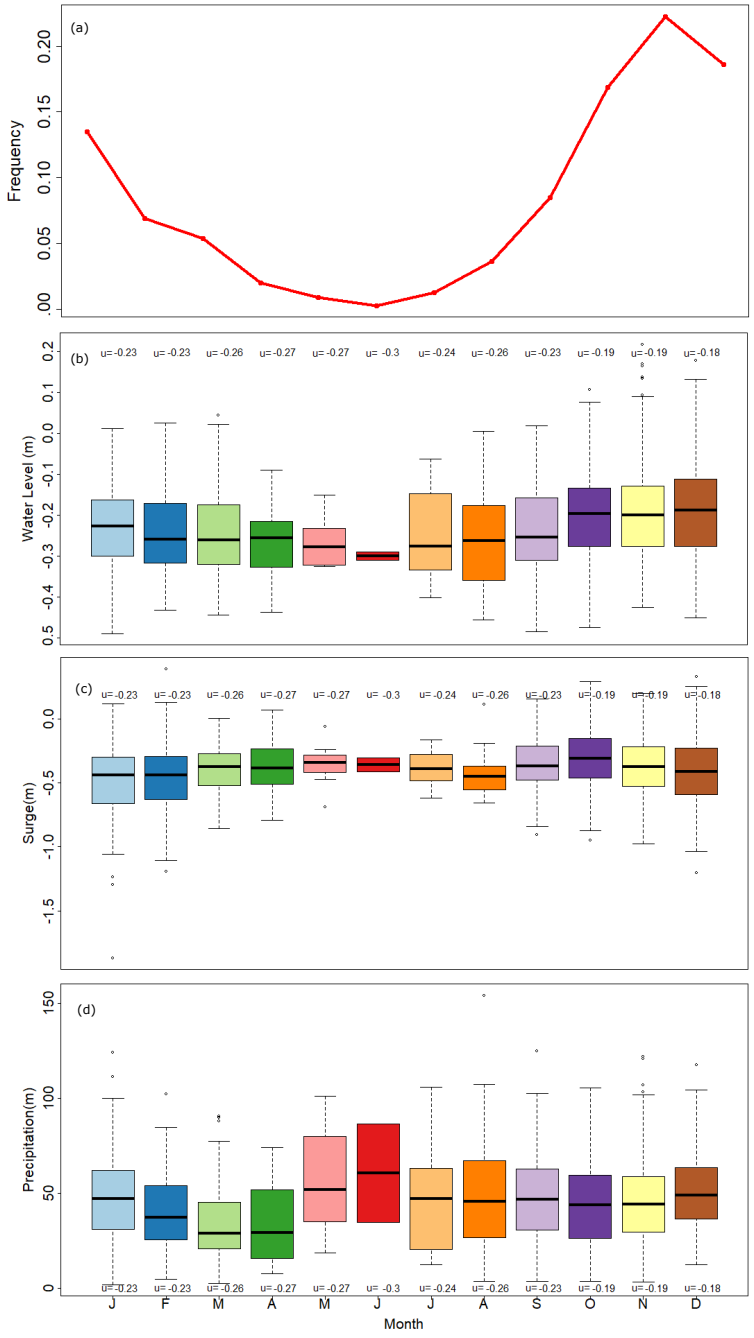

Figure S6. (a) Frequency of $WL_{\max}$ events occurring at each month. (c-d) show, respectively, the monthly mean of $WL_{\max}$, $S^{T}_{36h,\min}$ and $P_{12d,\text{acum}}$ conditioned to $WL_{\max}$.

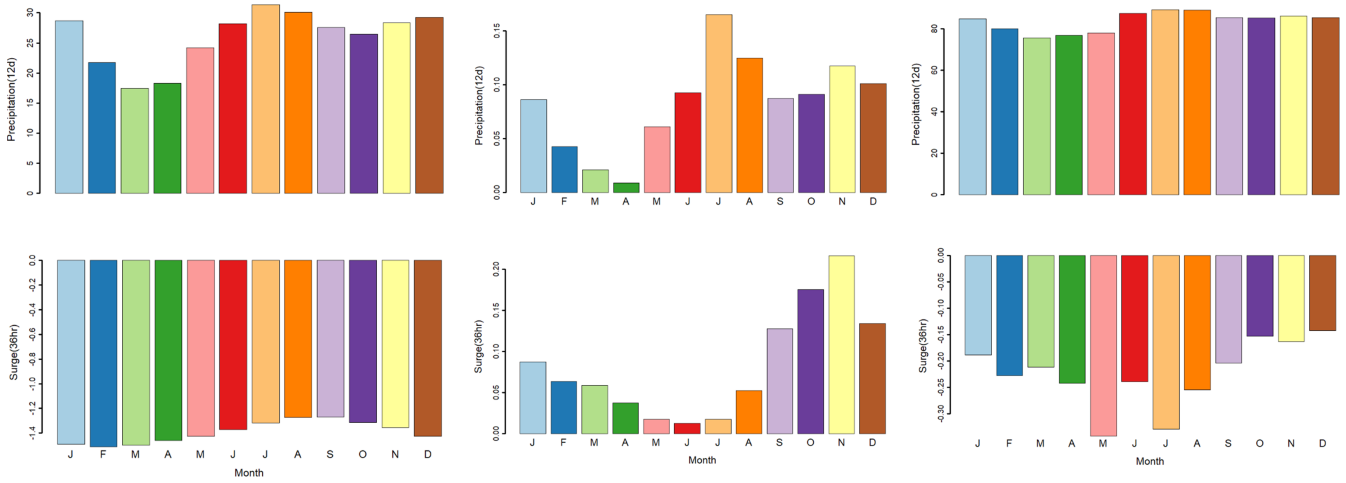

Figure S7. (a) Monthly mean surge and precipitation, (b) Frequency of the (univariate) annual maximum surge and precipitation occurring at each respective month, (c) monthly mean of the (univariate) annual maximum surge and precipitation. Note that that here neither surge nor precipitation is conditioned to the annual maximum water level.

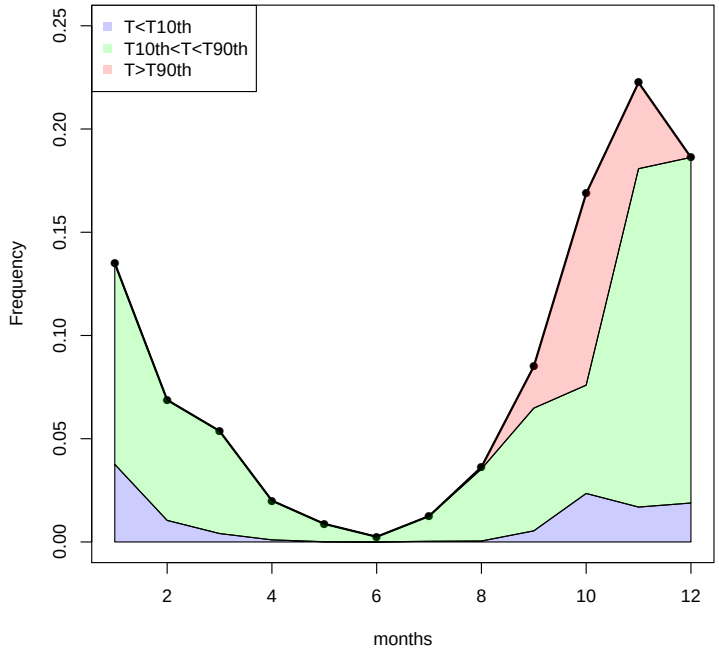

Figure S8. Monthly frequency of the tidal ranges indicated in the legend (shaded areas) relative to the total monthly frequency of $WL_{\max}$ events occurring each month (thick black line).

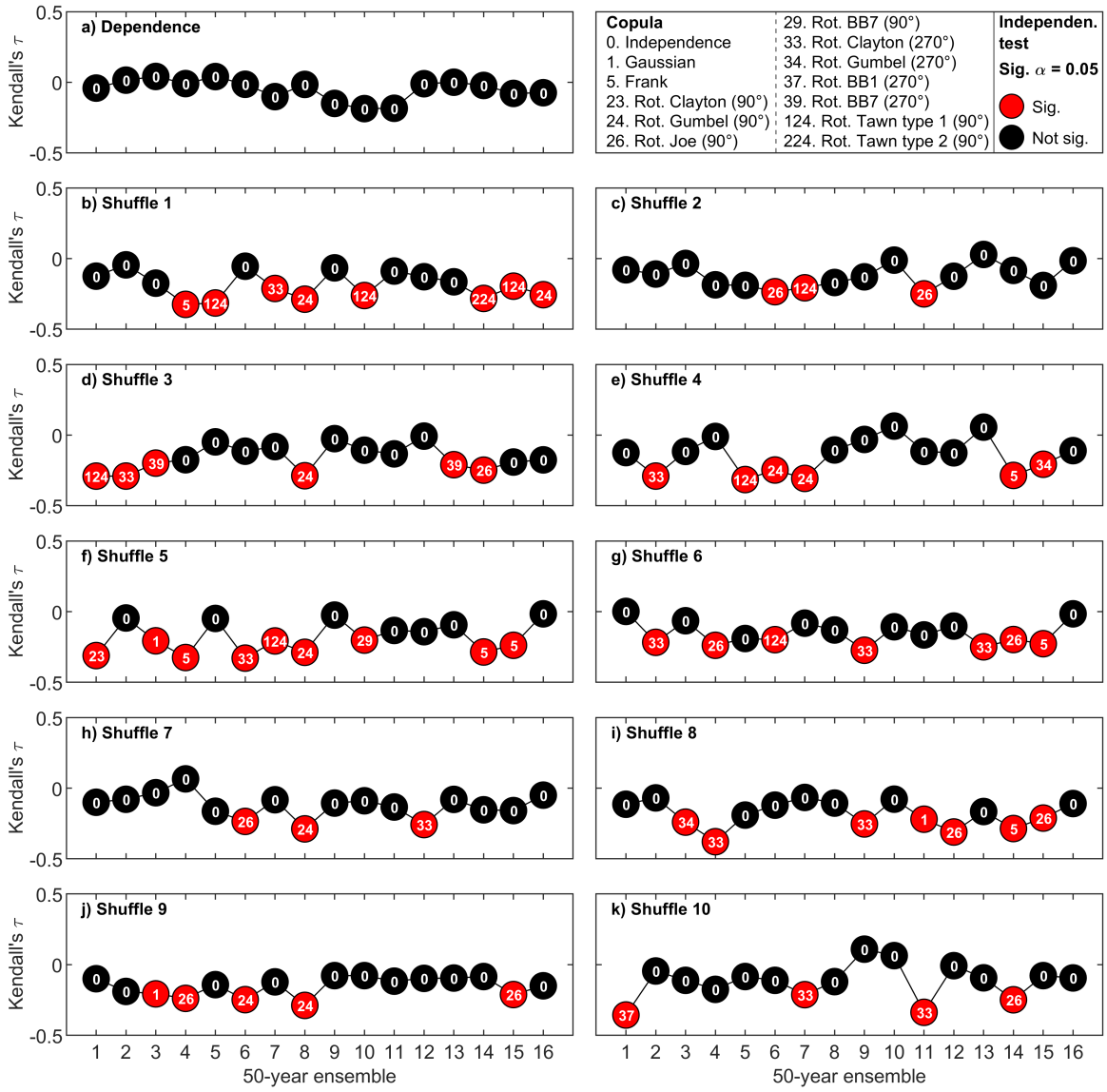

Figure S9. Variability of copula fitting for 50-year runs for original (a) and shuffled data (b-k). Red dots indicate the independence test is rejected.

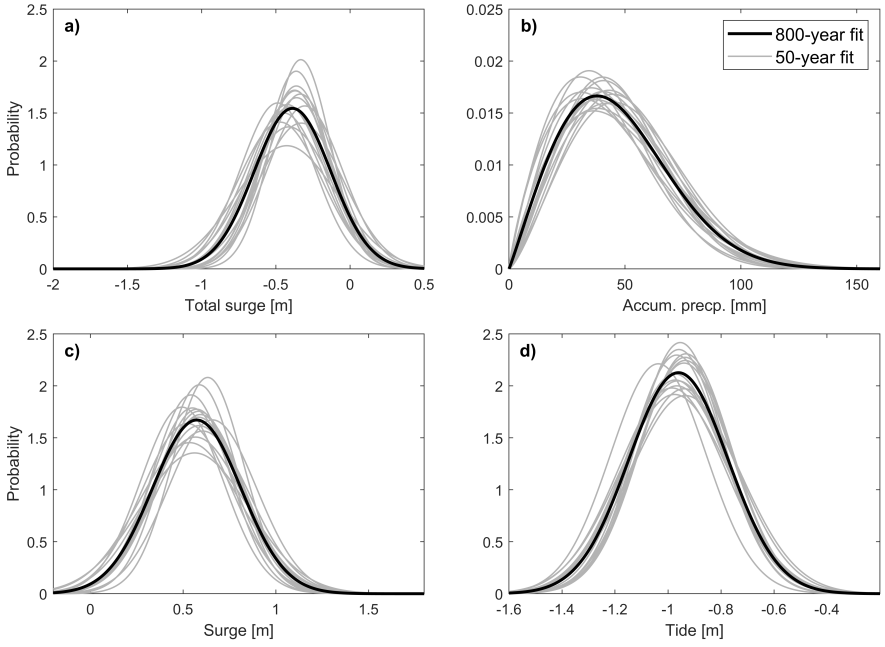

Figure S10. Variability of marginal probability density function for 50-year runs (gray lines) and 800-year ensemble (black line) for the following predictors (original data): (a) $S^{T}_{36h,\min}$, (b) $P_{12d,\mathrm{acum}}$, (c) $S_{72h,\mathrm{mean}}$, (d) $T_{12h,\min}$.

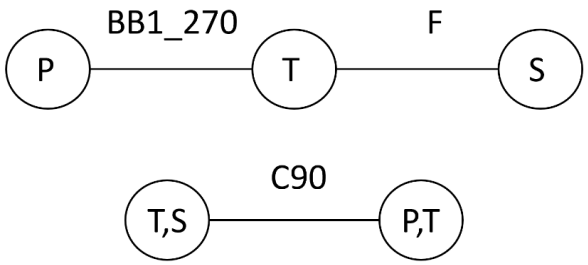

Figure S11. Structure of the regular vine obtained for the 3D dependence case.

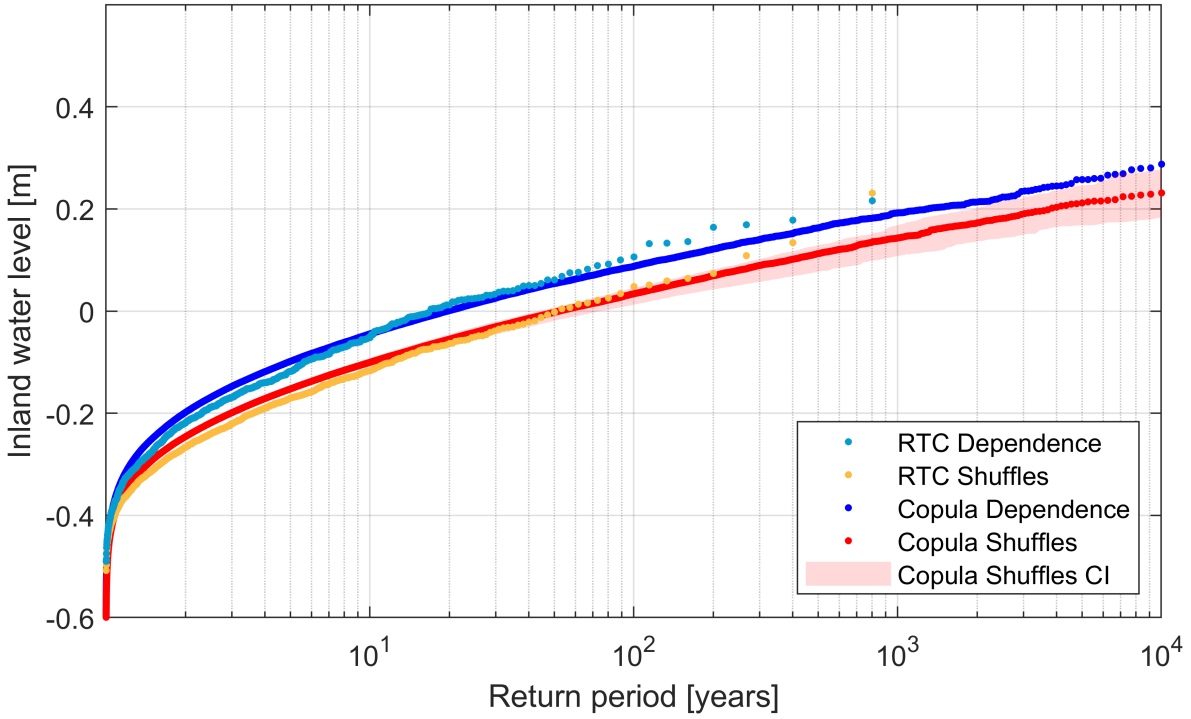

Figure S12. Inland water level return level against estimate return period using a trivariate copula model (3D case). Blue and red dotted lines depict the dependence and independence case, respectively. Transparent red denotes confidence intervals, which account for the uncertainty range between the 5^{th} and 95^{th} percentiles, as computed from all shuffles. Gray and black lines represent the curves empirically obtained.

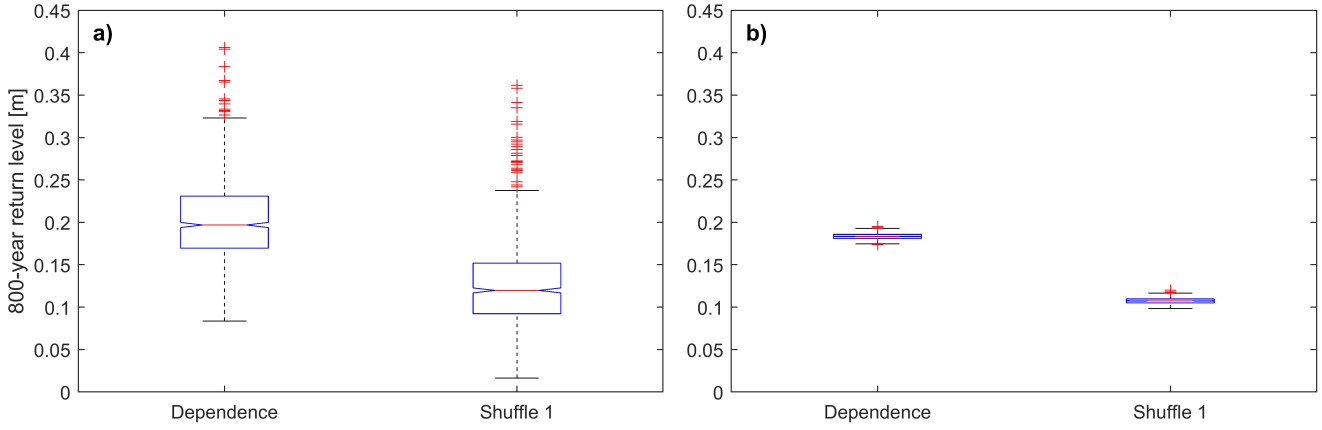

Figure S13. 800-year return water level as obtained by calibrating the proposed statistical framework with original and shuffled data, respectively, and by simulating 800-yr(a) and 100,000-yr(b) records, respectively. Only the first 800-year permutation of the shuffled data is used here for illustration purposes.

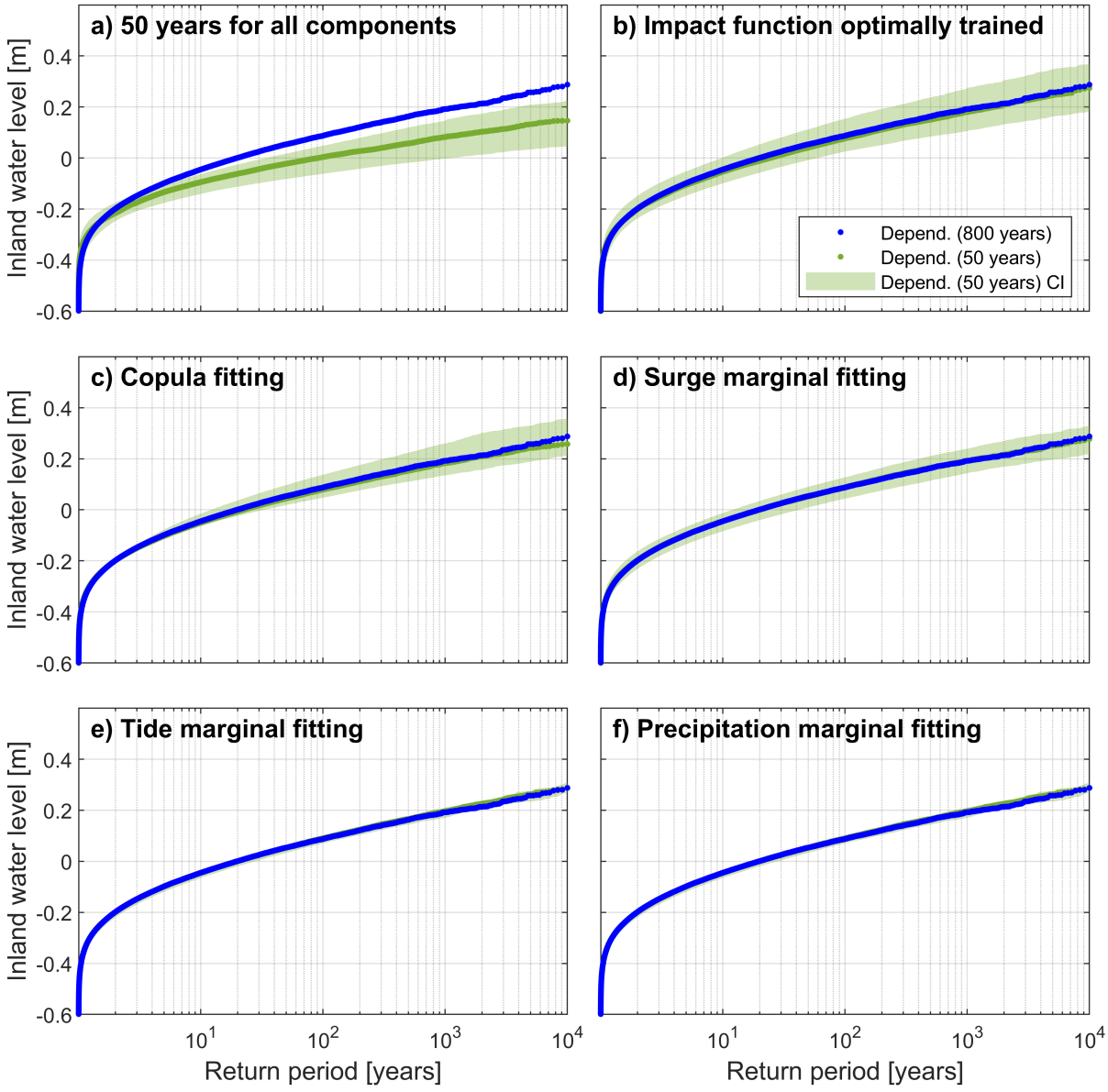

Figure S14. Inland water level return level against estimated return period using a trivariate copula. Gray and black dots depict the return level estimates obtained for the dependence and independence cases, respectively, using the proposed statistical framework. Blue and red illustrate the uncertainty associated to internal climate variability, represented by bounds computed using the 5th and 95th percentiles from all 50-year ensembles, and the median value (dots). Uncertainty is assessed for each component of the methodology: a) 50-year ensembles are used for all components; b) same as a) but impact function is optimally trained with 800 years of data; c) 50-year runs are used for copula fitting only; d) 50-year runs are used for total surge marginal fitting only; e) 50-year runs are used for tide marginal fitting only; and f) 50-year runs are used for precipitation marginal fitting only.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H.: Pair-copula constructions of multiple dependence, Insurance: Mathematics and economics, 44, 182–198, 2009.

Kendall, M. G.: A new measure of rank correlation, Biometrika, 30, 81–93, 1938.

Schepsmeier, U., Stoeber, J., Brechmann, E., Graeler, B., Nagler, T., and Erhardt, T.: Vinecopula: Statistical inference of vine copulas [Computer software manual], (R package version 2.1. 8), 2018.

Sklar, M.: Fonctions de repartition an dimensions et leurs marges, Publ. inst. statist. univ. Paris, 8, 229–231, 1959.