Journal Pre-proof

Imputation of missing values in environmental time series by D-vine copulas

Antoine Chapon, Taha B.M.J. Ouarda, Yasser Hamdi

PII: S2212-0947(23)00044-0
DOI: https://doi.org/10.1016/j.wace.2023.100591
Reference: WACE 100591

To appear in: *Weather and Climate Extremes*

Received date : 18 January 2023
Revised date : 12 June 2023
Accepted date : 22 June 2023

Please cite this article as: A. Chapon, T.B.M.J. Ouarda and Y. Hamdi, Imputation of missing values in environmental time series by D-vine copulas. *Weather and Climate Extremes* (2023), doi: https://doi.org/10.1016/j.wace.2023.100591.

Imputation of missing values in environmental time series by D-vine copulas

Antoine Chapon[1, 2], Taha B. M. J. Ouarda[1], and Yasser Hamdi[2]

[1]Institut National de la Recherche Scientifique, Quebec City, QC, Canada
[2]Institut de Radioprotection et de Sûreté Nucléaire, Fontenay-aux-Roses, France

June 23, 2023

Abstract

Missing values in environmental time series are common and must be imputed before carrying out an analysis requiring complete data. We propose an imputation method for the time series of a target station using information of neighboring stations measuring the same variable. The method allows these neighboring stations to have missing values themselves. The multivariate dataset comprising the time series of the target station and its neighboring stations is jointly modeled by a vine copula and parametric margins. Multiple imputation takes into account the uncertainty of missing data by generating several plausible values for each missing value in the time series of the target station. This is done in a Bayesian framework by sampling the posterior distribution of a missing value, which is conditional on the observed stations for the date. The method is suitable for extremes because the vine copula can model the eventual tail dependence between stations. The application to a skew surge time series is presented, with cross-validated results and a focus on the performance for the upper extremes.

keywords: missing value, multiple imputation, extreme value, vine copula, Bayesian inference.

1 Introduction

The presence of missing values in time series is a recurring issue in environmental sciences and many other disciplines. Data may be missing for various reasons, such as measurement device failure or measurement error (Kalteh and Hjorth, 2009). Many common analyses applied to environmental variables, such as spectral analysis or extreme value analysis, require complete time series (Gao et al., 2018). The analysis can be performed on a subset of the dataset for which there are no missing values, but this can severely limit the length of the time series and thus negatively impact the results. Values can also be missing in a systematic way and introduce bias in an analysis. As an example, the probability of missingness can be higher during extreme events due to measurement device failure, which would artificially reduce the frequency of extremes in the recorded time series. Therefore, a prior imputation of the missing data is often necessary. Even for an analysis that can accommodate missing values, imputing rather than ignoring them can improve results by increasing the length of the time series and reducing the potential bias caused by the missingness mechanism.

Our interest is in the imputation of the time series at a given station (referred to as the target station thereafter), using the information of other stations (the neighboring stations) measuring the same variable in an homogeneous region. The imputation method must allow neighboring stations

to have missing values, as they are also subject to missingness. This homogeneous region is defined according to the objective of the subsequent analysis of the imputed dataset. Hamdi et al. (2019) and Andreevsky et al. (2020) defined this region based on the ratio of common extreme events between the target station and each neighboring station, which is adapted if the interest is in the extremes. Since the analysis of extreme values is of particular interest in environmental sciences, e.g., for risk assessment, the imputation must retain good performance for the tails of the distribution in addition to its bulk, and must take into account the uncertainty associated with missing data (Serinaldi and Kilsby, 2015).

The imputation methods available for environmental sciences have been extensively reviewed (Kalteh and Hjorth, 2009; Ben Aissia et al., 2017; Gao et al., 2018; Hamzah et al., 2020). The time series of the target station and its neighboring stations constitute a multivariate dataset. Joint modeling of this dataset with a parametric distribution is a suitable approach if the extremes are of interest, as the tails of the distribution are explicitly modeled. Furthermore, the uncertainty can be accounted for with multiple imputation by repeated sampling from the multivariate distribution. Multiple imputation involves replacing missing values with several plausible values of what could have been observed (Little et al., 2014). In a Bayesian framework, multiple imputation amounts to sampling several values from the posterior distribution of a missing value.

Copulas are a popular option for constructing multivariate distributions in many fields, including environmental sciences (Tootoonchi et al., 2022). A copula is a multivariate distribution with uniform margins on $[0, 1]$ which models a dependence structure (Größer and Okhrin, 2022). For a d-dimensional random vector $U \in [0, 1]^d$, a copula C is defined by:

$$C(u_1, \ldots, u_d) = \Pr(U_1 \leq u_1, \ldots U_d \leq u_d). \tag{1}$$

Each dimension of a dataset can be transformed to be uniform over $(0, 1)$ with its probability integral transformation (Yan, 2007). A d-dimensional joint distribution F with margins $F_1, \ldots, F_d$ has a unique copula C, such that:

$$F(x_1, \ldots, x_d) = C\{F_1(x_1), \ldots, F_d(x_d)\}. \tag{2}$$

Thus, copulas allow separate modeling of the margins and the dependence between dimensions. Many parametric families of two-dimensional copulas (referred to as pair-copulas thereafter) exist to model a variety of dependence structures, but the choice becomes much more restricted for copulas in higher dimensions (Aas et al., 2009). Furthermore, the existing families of copulas usable in three or more dimensions, such as the Gaussian or Student t copulas, can impose a too restrictive dependence structure for a given dataset.

Copulas have already been applied to multiple imputation. Hollenbach et al. (2014) used Gaussian copulas in two or more dimensions, but this is not appropriate for the extremes because the dimensions are asymptotically independent in the Gaussian copula (i.e., it cannot model the eventual tail dependence). Di Lascio et al. (2015) considered the Gaussian and t copulas in high dimension, or several families of pair-copulas. The high-dimensional t copula has the inverse drawback compared to the Gaussian copula, as it forces a positive tail dependence. The imputation with a single pair-copula (i.e., not several pair-copulas organized as a vine copula) is restricted to two-dimensional datasets. Vine copulas solve both limitations of copulas by constructing high-dimensional dependence structures as assemblages of pair-copulas, taking advantage of the rich diversity of pair-copula families and allowing the dependence structure (including the tail dependence) to vary for each pair of dimensions. Ahn (2021) applied a D-vine copula to estimate streamflow at a partially gauged site conditionally on observations at neighboring stations, with the latter having complete records

in their time series. Ahn (2021) compared the performance of the D-vine with six other imputation methods, including geostatistical methods with inverse distance weighting and Kriging, and found the D-vine to outperform them. Hasler et al. (2018) developed the imputation of a multivariate dataset with a D-vine copula for the case where each dimension may have missing values. The approach of Hasler et al. (2018) was restricted to monotone missingness patterns (i.e., only the lower and/or upper dimensions of the ordered dimensions of the D-vine are missing), and assumed the data to be missing completely at random (MCAR, meaning that the probability of missingness is independent on the actual data value). Hasler et al. (2018) compared the performance of their method with five alternatives, different from the six alternatives tested by Ahn (2021), and also found the D-vine to outperform them. In particular, the D-vine outperforms alternatives when the extremes are considered because it accounts for the eventual tail dependence. Jane et al. (2016) applied copulas to extend wave height time series beyond their measurement period using regional information. This can be achieved by considering the dates outside of the observation period as missing values, which does not require additional development of the multiple imputation model. Valle and Kaplan (2019) applied a Gaussian copula for a counterfactual analysis of a dataset where every dimension can have missing values, which shows that the ability of copulas to handle missing values has applications beyond imputation.

Since our objective is to have a multiple imputation method suitable for the extremes, we followed the methodologies of Ahn (2021) and Hasler et al. (2018) by using a D-vine copula to model the joint distribution of the target station and its neighboring stations. The selection of the family for each pair-copula of the D-vine is critical to account for the eventual tail dependence, or the tail independence, so the parametric families are selected with a Bayesian framework (Min and Czado, 2011). The generation of plausible values for multiple imputation is also performed in a Bayesian framework to account for uncertainty through credible intervals.

This paper is organized as follows. Section 2 presents the methodology. An application to a skew surge station located on the French Atlantic coast is presented in Section 3. The methodology and results are discussed in Section 4, along with a conclusion.

2 Methods

The time series at the target station and its neighbors constitute a multivariate dataset. The margins of this dataset are modeled with univariate parametric distributions. The marginal nonexceedance probabilities of the observations at each station are obtained from their respective margins. The dependence structure of the multivariate probabilities are then modeled with a vine copula, each station corresponding to a dimension of this multivariate distribution. For a given date, values are sampled from the vine copula for the missing dimensions, conditionally on the observed one. The multiple imputation accounts for the uncertainty of the values generated (Little et al., 2014). Finally the multiple imputed probabilities are transformed back to quantiles.

2.1 Marginal distribution

Joint modeling with copulas is often performed in a semiparametric way through the pseudo-likelihood method, which uses the rank of observations obtained from the empirical distribution of the margins (Genest and Favre, 2007). The pseudo-likelihood method is not suitable when the extremes are of interest, because the empirical distribution is not precise enough in the tails where there are few observations, so a parametric distribution for the margins is required.

For time series with support on $\mathbb{R}$, the skewed generalized t distribution can offer a good fit in most cases and is used in this study. Its distribution is given by:

$$f_{SGT}(y|\mu,\sigma,\lambda,p,q) = p\left[2\sigma q^{\frac{1}{p}} B\left(\frac{1}{p},q\right)\right]^{-1}\left(1+\frac{1}{q}[1+\lambda\,\mathrm{sign}(y-\mu)]^{-p}\left|\frac{y-\mu}{\sigma}\right|^{p}\right)^{-(q+\frac{1}{p})}, \quad (3)$$

where μ is a location parameter, σ is a positive scale parameter, λ controls the skewness with $\lambda \in (-1,1)$, p is a positive parameter controlling the peakedness of the density, q is a positive parameter controlling the tails and $B(\cdot,\cdot)$ is the beta function (Kerman and McDonald, 2013). This distribution is implemented in the *sgt* R package (Davis, 2015).

The adequacy of the margins is assessed with quantile-quantile plots. Note that different parametric distributions could be used to model the margins of the different dimensions of the joint distribution.

2.2 Vine copulas

A d-dimensional vine copula is composed of $n_c = d(d-1)/2$ pair-copulas. For the case of a 3-dimensional distribution, the density can be modeled with:

$$\begin{aligned} f(x_1,x_2,x_3) = & f(x_1)\,f(x_2)\,f(x_3) \\ & c_{12}\{F(x_1),F(x_2)\}\,c_{23}\{F(x_2),F(x_3)\} \\ & c_{13|2}\{F(x_1|x_2),F(x_3|x_2)\}, \end{aligned} \quad (4)$$

where c_{12} and c_{23} are the densities of the pair-copulas between the corresponding dimensions, and $c_{13|2}$ is the density of the pair-copula between dimensions 1 and 3, conditional on dimension 2. $F(x_1|x_2)$ and $F(x_3|x_2)$ are the marginal conditional distributions given by:

$$F(u|v) = \frac{\partial\, C_{uv}\{F(u),F(v)\}}{\partial\, F(v)}, \quad (5)$$

where C_{uv} is the distribution of a pair-copula (Aas et al., 2009). For vine copulas in dimension higher than three, (5) is used recursively to obtain the marginal distributions conditional on more than one dimension. A vine copula can be represented as a graph with nodes and edges, the latter corresponding to the pair-copulas (Figure 1). These nodes and edges are organized by trees, with the two dimensions and conditioning dimensions of a pair-copula in a given tree being specified by nodes on the lower tree. Thus, a tree contains the pair-copulas conditional on the same number of dimensions (with unconditional pair-copulas in the first tree). The D-vine is a special case of vine copulas which is fully specified by the order of the dimensions in its first tree (Figure 1). A joint density $f(x_1,\ldots,x_d)$ with a D-vine for the dependence between variables is given by:

$$\begin{aligned} f(x_1,\ldots,x_d) = & \prod_{m=1}^{d} f(x_m) \\ & \prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{i,i+j|i+1,\ldots,i+j-1}\{F(x_i|x_{i+1},\ldots,x_{i+j-1}),F(x_{i+j}|x_{i+1},\ldots,x_{i+j-1})\}, \end{aligned} \quad (6)$$

where $c_{i,i+j|i+1,\ldots,i+j-1}$ is the pair-copula density between the dimensions x_i and x_{i+j} transformed to probabilities with their respective margins $F(x_i|x_{i+1},\ldots,x_{i+j-1})$ and $F(x_{i+j}|x_{i+1},\ldots,x_{i+j-1})$ (Hasler et al., 2018).

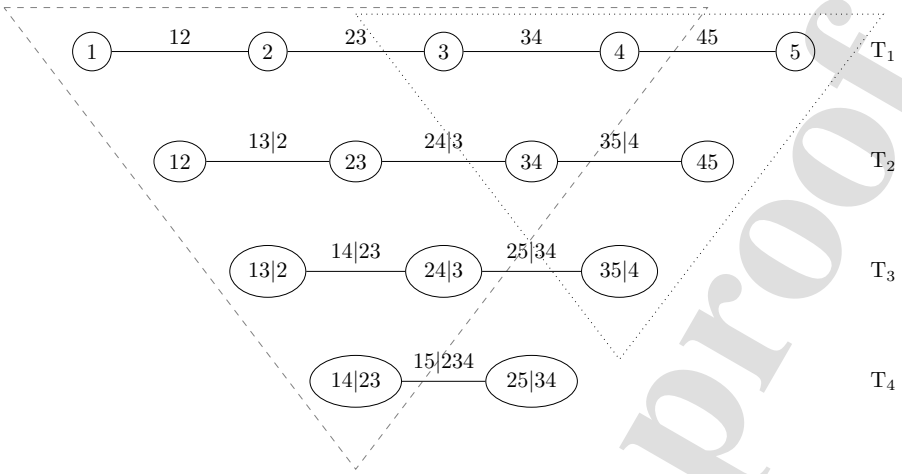

Figure 1: 5-dimensional D-vine, with four trees from top to bottom (T_1 to T_4). The dimensions of the pair-copulas are indicated by the edges between the nodes, with unconditional pair-copulas in T_1 and conditional ones in the subsequent trees. As an example, the edge labeled 13|2 on the second tree represents the pair-copula between dimensions 1 and 3, conditional on dimension 2. The labels of the nodes show how the construction of a vine copula is systematic, with the two dimensions of the pair-copulas on a given tree and their set of conditioning dimensions depending on the previous tree. The dashed and dotted areas give examples of subsets of the D-vine to smaller ones with dimensions 1 to 4 including the six pair-copulas 12, 23, 34, 13|2, 24|3 and 14|23, and dimensions 3 to 5 including the three pair-copulas 34, 45 and 35|4.

2.3 D-vine with missing data

The construction of vine copulas by an assemblage of pair-copulas makes them nested models. Depending on its structure, a vine copula of a given dimension can be reduced to a smaller one of lesser dimension if the remaining pair-copulas are not conditional on the removed dimensions. In the case of the D-vine structure, this subsetting of the model results in a smaller D-vine. Figure 1 presents a 5-dimensional example, with the dashed and dotted lines delineating smaller D-vines obtained when the dimensions 1 and 2, or 5, respectively, are removed. Hasler et al. (2018) exploited this property of the D-vine to compute the likelihood corresponding to each date (i.e., to each multivariate observation). Assuming the data MCAR, the contribution of each date to the likelihood is the observed-data likelihood, given by:

$$f(x_{obs}) = \int f(x_{obs}, x_{mis})\, dx_{mis}, \tag{7}$$

where x_{obs} and x_{mis} are the observed and missing dimensions of this date, respectively. For the example of a 4-dimensional D-vine, the contribution to the likelihood of a date having the last fourth dimension (i.e., the rightmost) missing is:

$$f(x_1, x_2, x_3) = \int f(x_1, \ldots, x_4)\, dx_4,$$

which after integrating (6) results in the joint density of (4). This is applied recursively if the third dimension is missing along the fourth one. The same applies if one or several leftmost dimensions are missing. Note that dimensions could be missing on either side of the D-vine, but the remaining dimensions need to be continuously ordered (e.g., if only the third dimension of a four-dimensional D-vine is missing, a valid smaller D-vine cannot be obtained by removing this third dimension only).

The purpose of subsetting the full d-dimensional D-vine into smaller ones is to use more dates to compute the likelihood. In the case of Hasler et al. (2018), their dataset only had a monotone missingness pattern, so these subsets allowed every observation to contribute to the likelihood. In our case, the missingness patterns of some dates do not correspond to a valid D-vine subset, when the observed dimensions are not ordered continuously in the full D-vine, therefore these dates cannot contribute to the likelihood. Despite not being able to use every observation for the likelihood of the D-vine, this approach still uses much more information compared to only using dates without any missing value. Subsetting the full D-vine is also useful to reduce the computational requirement when sampling from the model for imputation, as will be presented in Section 2.5.

Let S be the set of missingness patterns corresponding to a subsettable D-vine, including the full D-vine. Let $x_{.,s}$ be the observations for which a valid subset $s \in S$ exists, and $m_s \subseteq 1, \ldots, d$ the observed dimensions of this subset. The likelihood of the D-vine computed with its subsets S is given by:

$$L_S(\theta|x) = \prod_{s \in S} \left[\prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1,\ldots,i+j-1}\{F(x_{i,s}|x_{i+1,s},\ldots,x_{i+j-1,s}), F(x_{i+j,s}|x_{i+1,s},\ldots,x_{i+j-1,s})\} \Big| i, i+j \in m_s \right], \tag{8}$$

where θ is the parameter vector of all the n_c pair-copulas of the D-vine. Note that compared to the density in (6), the likelihood in (8) only concerns the dependence structure modeled by the D-vine and does not include the margins.

The target station and its $d-1$ neighboring stations each correspond to a dimension of the d-dimensional D-vine. The dimensions (i.e., stations) in the D-vine are ordered by considering two criteria. The first criterion is to order the dimensions with the highest pairwise tail dependence next to each other, so that the corresponding pair-copulas are in the first trees of the D-vine. The first tree has unconditional pair-copulas and the lower trees are conditional on fewer dimensions than the higher trees (Figure 1). The pairwise tail dependence is estimated by fitting the t pair-copula on the observations of two stations transformed to probabilities with their respective margins. For the t copula, the lower and upper tail dependence λ is the same, given by:

$$\lambda = 2\, t_{\eta+1}\left(-\sqrt{\eta+1}\sqrt{\frac{1-\omega}{1+\omega}}\right), \tag{9}$$

where $t_{\eta+1}$ is the t distribution with $\eta+1$ degrees of freedom, ω is the parameter of the t pair-copula and η its degrees of freedom (Nagler et al., 2022). The estimate of λ given by (9) is always positive, therefore, the tail dependence is also tested with the method based on the Neyman–Pearson lemma described in Reiss and Thomas (2007). Since the time series in the application have serial correlation and that this test requires the data to be independent and identically distributed, it is applied to

100 resamples of 1 000 values for each pair of stations to break the serial correlation, and the mean of the 100 p-values is used.

The second criterion for this ordering is placing the dimensions with the highest ratio of missing values on the edges of the D-vine, to allow for more dates to be used in the likelihood computation with (8). Furthermore, less observations are unusable for the likelihood computation if the dimensions closer to the edge of the D-vine have their missing values at the same date.

2.4 Selection and adjustment of pair-copulas by reversible jump Markov chain Monte Carlo (RJMCMC)

Once the dimensions are ordered, the selection of the pair-copula families in the D-vine and the adjustment of their parameters is done jointly via reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). This RJMCMC algorithm applied to a D-vine is a modified version of the one developed by Min and Czado (2011). It alternates between *Stay* steps, during which the parameters of every pair-copula are updated with a Metropolis-Hastings move, and *Jump* steps, during which the family of one pair-copula changes with a reversible jump move. Only uniform priors are specified. The D-vine is fitted on the nonexceedance probabilities of the observations, which are obtained with the probability integral transformation of the skewed generalized t margins.

Six families of pair-copula are considered in the D-vine to model different types of dependence between its dimensions (Table 1). The independence pair-copula accounts for the independence between two dimensions of the D-vine, or for conditional independence in trees higher than the first one. The Gaussian copula covers cases when two dimensions are dependent for the bulk of the data but asymptotically independent (i.e., absence of tail dependence), and allows for a positive or negative correlation. Both the survival Clayton and Gumbel copulas have upper tail dependence. The BB1 and survival BB1 copulas are mixtures based on the Gumbel copula with dependence in both lower and upper tails. These pair-copulas are selected to allow a good fit of the D-vine for the upper extremes in particular, but if the lower extremes were of interest instead, the survival Clayton and Gumbel copulas could be swapped for the Clayton and survival Gumbel copulas, which have a positive lower tail dependence. If the Gaussian copula was not considered, all the families other than the independence copula would have a positive upper tail dependence, which could force the model to have asymptotic dependence between some dimensions.

More families could be considered but it is preferable to limit the set of families so that relevant pair-copulas are proposed more often during the *Jump* steps of the RJMCMC algorithm. The Student t copula is not considered because the evaluation of its likelihood takes much more time than for the six families mentioned previously (with the *VineCopula* R package). If the method is applied to a smaller dataset, the RJMCMC can be run for more iterations, so a larger set of pair-copula families can be tested. Likewise the t copula could be included for a small dataset, since the difference in computation time would then be negligible.

The D-vine is initialized with the selection method described in Hasler et al. (2018), where the pair-copula in each tree is selected recursively from the first to the last tree. The estimate of the parameter vector θ of the D-vine and the family of each pair-copula obtained by this initial selection are used as starting values in the RJMCMC. The algorithm of Hasler et al. (2018) is adapted to our setting by using only the observations with missingness patterns corresponding to valid D-vine subsets $s \in S$, as for the likelihood in (8).

Table 1: Parameter boundaries, lower and upper tail dependence of the six families of pair-copula (with 0 and + indicating the absence and positive tail dependence, respectively). The boundaries of the copula parameters and the code for each family in the *VineCopula* R package are also provided.

code	family	l_ρ	u_ρ	l_ν	u_ν	lower t.d.	upper t.d.
0	Independence					0	0
1	Gaussian	-1	1			0	0
13	survival Clayton	0	28			0	+
4	Gumbel	1	17			0	+
7	BB1	0	5	1	6	+	+
17	survival BB1	0	5	1	6	+	+

2.4.1 *Stay* step: updating of pair-copula parameters

During the *Stay* step, the parameters of each pair-copula are updated sequentially. Let $i \in 1, \ldots, n_c$ be the index of the pair-copulas of the D-vine. Let θ be the parameter vector of the entire D-vine. Let k be the index of the pair-copula for which new parameter values are proposed, with $\theta_k = \{\rho_k\}$ or $\theta_k = \{\rho_k, \nu_k\}$ the parameter vector of this kth pair-copula, for a family with one or two parameters, respectively. We denote by $l_{\rho,i}$ and $u_{\rho,i}$ the lower and upper bounds, respectively, for the first parameter of the family of the ith pair-copula. We use the similar notation $l_{\nu,i}$ and $u_{\nu,i}$ for the eventual second parameter. These bounds are specific to each pair-copula family (Table 1).

Let *old* and *new* refer to the current and proposed states of the chains, respectively. A value θ_k^{new} is drawn from an adaptive proposal $N(\theta_k^{old}, \Sigma_k)$ in one or two dimensions, truncated to $(l_{\rho,k}, u_{\rho,k})$ and $(l_{\nu,k}, u_{\nu,k})$ for each dimension, respectively. The variance or covariance matrix of this proposal is given by:

$$\Sigma_k = \begin{cases} (1-\beta)\, 2.4^2\, SamVar_k + \beta\, (u_{\rho,k} - l_{\rho,k})/1000, & \text{if } \theta_k = \{\rho_k\} \\ (1-\beta)\, 2.4^2/2\, SamVar_k + \beta\, \mathrm{diag}(u_{\rho,k} - l_{\rho,k}, u_{\nu,k} - l_{\nu,k})/1000, & \text{if } \theta_k = \{\rho_k, \nu_k\} \end{cases} \tag{10}$$

where $SamVar_k$ is the sample variance or covariance matrix of the chains for the current family sampled so far for the kth pair-copula, and $\mathrm{diag}(a, b)$ is a two-dimensional diagonal matrix with a and b on the main diagonal. This adaptive proposal is used when the one or two chains of θ_k contain at least 50 sampled values each, with $\beta = 0.01$. Up until this point, the proposal is nonadaptive with $\beta = 1$ (Roberts and Rosenthal, 2009; Craiu and Rosenthal, 2014).

The uniform prior of the D-vine is given by:

$$\pi(\theta) = \prod_{i=1}^{n_c} (u_{\rho,i} - l_{\rho,i})^{-1} (u_{\nu,i} - l_{\nu,i})^{-1}, \tag{11}$$

where $(u_{\nu,i} - l_{\nu,i})^{-1} = 1$ if the family of the ith pair-copula has only one parameter and $(u_{\rho,i} - l_{\rho,i})^{-1} = 1$ as well for the independence copula.

The new parameter vector θ^{new} of the entire D-vine is assembled with:

$$\theta_i^{new} = \begin{cases} \theta_i^{new}, & \text{if } i = k \\ \theta_i^{old}. & \text{if } i \neq k \end{cases} \tag{12}$$

The acceptance probability of this Metropolis-Hastings move is given by:

$$\alpha_{stay} = \min\left\{1, \frac{L_S(\theta^{new}|x)}{L_S(\theta^{old}|x)} \frac{\pi(\theta^{new})}{\pi(\theta^{old})} \frac{\phi(\theta_k^{old})}{\phi(\theta_k^{new})}\right\}, \tag{13}$$

where $L_S(\theta|x)$ is the likelihood of the D-vine computed with its subsets given by (8), and $\phi(\theta_k)$ is the density of the one or two dimensional truncated normal proposal (depending if the kth pair-copula has one or two parameters). The uniform priors in (13) cancel out since the families of pair-copulas are unmodified during the *Stay* step.

Note that the Gibbs sampler can not be used instead of the Metropolis-Hastings sampler because the full conditional distribution of the D-vine is not known (Min and Czado, 2010).

2.4.2 *Jump* step: modification of pair-copula families

During the *Jump* step, one pair-copula of the D-vine is uniformly selected to propose a modification of its family. Let n_f be the number of pair-copula families considered and f_k the family of the kth pair-copula selected. Each family other than the current one has a $1/(n_f - 1)$ probability of being proposed for a *Jump* step.

Once the pair-copula and the proposed family are selected, new parameter values θ_k^{new} are generated from a one or two dimensional normal distribution $N(\theta_k^*, \Sigma_k)$ truncated to $(l_{\rho,k}, u_{\rho,k})$ and $(l_{\nu,k}, u_{\nu,k})$, respectively. Σ_k is defined as (10) for each combination of pair-copula of the D-vine and family. If the family f_k has not been sampled at least 50 times for the kth pair-copula the proposal is nonadaptive, with $\beta = 1$.

θ_k^* is obtained by adjusting the proposed pair-copula to its marginal probabilities, which are given by $F(x_i|x_{i+1}, \ldots, x_{i+j-1})$ and $F(x_{i+j}|x_{i+1}, \ldots, x_{i+j-1})$ for a copula between dimensions i and $i+j$ of the D-vine (as was mentioned in Section 2.3). The marginal probabilities are obtained recursively from conditional distributions given by (5). As an example, the marginal probabilities for the pair-copula $C_{14|23}$ on the third tree of the D-vine would be conditional on two dimensions (Figure 1). The marginal probabilities for the first dimension of this pair-copula are obtained with:

$$F(x_1|x_2, x_3) = \frac{\partial\, C_{13|2}\{F(x_1|x_2), F(x_3|x_2)\}}{\partial\, F(x_3|x_2)},$$

where $C_{13|2}$ is the pair-copula on the second tree between dimensions 1 and 3 and conditional on dimension 2, and $F(x_1|x_2)$ and $F(x_3|x_2)$ are marginal probabilities obtained from the copulas C_{12} and C_{23} on the first tree, respectively (Hasler et al., 2018). The proposed pair-copula is adjusted on its two marginal probabilities time series by maximum likelihood. The marginal probabilities are obtained with the complete observations of the concerned dimensions, e.g., for the $C_{14|23}$ pair-copula, the marginal probabilities are obtained from the observations without missing values in dimensions $1, \ldots, 4$. Min and Czado (2011) estimated the location parameter of the proposal by adjusting the entire D-vine with the modified pair-copula family by maximum likelihood. Instead, estimating θ_k^* only by adjusting the proposed pair-copula on its marginal probabilities does not required evaluating the likelihood of the entire D-vine, which is the costliest part of the algorithm. This modification of the algorithm makes a significant difference in computation time for applications with a large dataset.

As in the *Stay* step with (12), the full parameter vector θ^{new} is assembled with θ_k^{new}, whose family and parameters are modified, and the family and parameters of the pair-copulas other than the kth that remain unmodified. The Jacobian of this bijection is equal to 1.

The acceptance probability of the jump is given by:

$$\alpha_{jump} = \min\left\{1, \frac{L_S(\theta^{new}|x)}{L_S(\theta^{old}|x)} \frac{\pi(\theta^{new})}{\pi(\theta^{old})} \frac{g(f_k^{new} \to f_k^{old})}{g(f_k^{old} \to f_k^{new})} \frac{\phi(\theta_k^{old})}{\phi(\theta_k^{new})}\right\}, \tag{14}$$

where $g(a \to b)$ is the probability of proposing a jump from family a to b (which in our case is the same for each family, thus the corresponding ratio cancels out), and $\phi(\theta_k)$ is the density of the one or two dimensional truncated normal proposal. $\pi(\theta)$ is the prior defined in (11), which does not cancel out in (14), compared to (13) (Min and Czado, 2011; Gruber and Czado, 2018).

2.5 Sampling the D-vine conditionally on the observed dimensions

Multiple imputation is performed by sampling from the missing dimensions of the D-vine conditionally on the observed dimensions. The sampled values are nonexceedance probabilities, which are transformed to quantile with the margins.

The D-vine is subsetted before sampling values according to a date missingness pattern, similarly to the subsets for the likelihood with (8). The purpose of sampling from subsets of the D-vine instead of the full d-dimensional one is to reduce the computational requirement. For a given date, the continuously missing dimensions of neighboring stations from each end of the D-vine can be removed without losing any usable dependency from a pair-copula.

Formally, let $i \in 1, \ldots, d$ be the index of the ordered dimensions of a D-vine, with i_z the dimension of the target station to be imputed. For a given observation x_t (i.e., a d-dimensional vector corresponding to the date t), let:

$$a_{t,i} = \begin{cases} 1, & \text{if } x_{t,i} \text{ is observed or if } i = i_z \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

and

$$b_{t,i} = \sum_{1}^{i} a_{t,i} \sum_{d}^{i} a_{t,i}. \tag{16}$$

The valid D-vine subset for x_t corresponds to the dimensions for which $b_{t,i} > 0$.

As an example, let us consider a 5-dimensional D-vine, as in Figure 1, where the third dimension would be the target station. If the dimensions 1, 2 and 4 were observed and the fifth dimension missing along the third, sampling only the third from a D-vine subsetted to dimensions $1, \ldots, 4$ (the dashed area in Figure 1) would be equivalent and faster compared to sampling jointly the third and fifth dimensions from the full D-vine, considering that only the third is of interest. However, if the second and fifth dimensions were missing, the valid subset would remain the same, with dimensions $1, \ldots, 4$, and sampling jointly from both the second and third dimensions would be required, because the pair-copulas between the first dimension and the third and fourth are conditional on the second one. Thus, the computational requirement for sampling an observation depends on its missingness pattern, and subsetting the D-vine reduces this requirement for most patterns.

For a given date t with a missing value for the target station, let y_t be the subset of x_t for which $b_{t,i} > 0$ (i.e., the subset of the observation vector corresponding to the dimension of the valid D-vine subset for this date). This subsetted observation vector y_t contains $m \geq 1$ missing values, with at least the missing value of the target station. Nonexceedance probabilities are sampled for these m dimensions of the D-vine subset (or full D-vine if $y_t = x_t$) with a Metropolis-Hastings algorithm.

The adaptive proposal is a m-dimensional normal distribution $N_m(y_{t,m}^{old}, \Sigma_m)$ truncated to the hypercube $[0,1]^m$, where $y_{t,m}$ corresponds to the m missing dimensions of y_t. The variance or covariance matrix Σ_m is defined similarly to (10), with:

$$\Sigma_m = (1-\beta)\, 2.4^2/m\, SamVar_m + \beta\, \mathrm{diag}_m(10^{-4}), \tag{17}$$

where $\mathrm{diag}_m(10^{-4})$ is a $m \times m$ diagonal matrix with 10^{-4} on the main diagonal and $SamVar_m$ is the sample variance or covariance matrix of the m chains sampled so far. $\beta = 0.01$ when at least 100 values have been sampled for each m dimension, otherwise $\beta = 1$. The acceptance probability is given by:

$$\alpha = \min\left\{1, \frac{L_s(\theta|y_t^{new})}{L_s(\theta|y_t^{old})} \frac{\pi(y_t^{new})}{\pi(y_t^{old})} \frac{\phi(y_{t,m}^{old})}{\phi(y_{t,m}^{new})}\right\}, \tag{18}$$

where y_t is the vector of observation assembled from the non missing values of y_t and the values $y_{t,m}^{old}$ or $y_{t,m}^{new}$ sampled for the previous or current iteration of the algorithm, respectively, $L_s(\theta|y_t)$ is the likelihood of the D-vine given by (8) for the subset $s \in S$ corresponding to y_t, and $\pi(y_t)$ is a uniform prior whose corresponding ratio cancels.

Convergence of the Markov chains is assessed following Gelman et al. (2015) with the $\widehat{R}$ test by running two or more times the m chains for a given date. The $\widehat{R}$ value indicates the potential scale reduction of the posterior distribution if the chains of length n were ran further, declining to 1 as $n \to \infty$. As a rule of thumb, the convergence is considered satisfactory if $\widehat{R} < 1.1$. For simplicity, this convergence test is performed only for the chains corresponding to the target station.

2.6 Model validation

The performance of the imputation is assessed by k-fold cross-validation, with $k = 5$ (James et al., 2017). The k training and validation sets are obtained by repeating and nonoverlapping blocks of size b, using $b = 70$ because the autocorrelation of the daily skew surge approaches nonsignificative levels at this lag (not shown). The value of k is kept low to reduce computation time.

For each k model, the observations of the validation set at the target station are considered missing and are imputed by MCMC. For the sake of simplicity, the convergence of the chains were not assessed for the cross-validation. A highest posterior density credible interval is computed for each imputed date. These credible intervals are computed on the quantiles rather than their nonexceedance probabilities because the posterior distributions of the latter are skewed, with a negative (positive) skew for the upper (lower) extremes. The highest density intervals would be affected by the skewness of the probabilities (Hyndman, 1996). This skewness disappears when the sampled probabilities are transformed back to quantile. The ratios of observations falling within their respective 90% credible interval, or below and above, indicate the validity of the imputation's uncertainty obtained by the D-vine. A perfect model would have 90% of observations inside their respective credible intervals. The highest density intervals are computed with the *hdrcde* R package (Hyndman et al., 2021).

The validity of the model is further assessed with the Nash–Sutcliffe efficiency (NSE) score, given by:

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}(X_t^{imp} - X_t^{obs})^2}{\sum_{t=1}^{T}(X_t^{obs} - \overline{X^{obs}})^2}, \tag{19}$$

where T is the total number of timesteps of the observed values, X_t^{imp} is the mean of the values generated by MCMC for the date t, X_t^{obs} is the observed value for the date t and $\overline{X^{obs}}$ is the mean

of the observed values (Knoben et al., 2019). NSE = 1 indicates that the model perfectly reproduces the observations, while NSE = 0 indicates that it has the same explanatory power as the mean. The NSE is also computed through the k-fold cross-validation, with k NSE values.

3 Application

3.1 Data and case study

The imputation method is applied to a dataset of skew surge time series for nine tide gauges located along the French Atlantic coast (Figure 2). The skew surge is defined as the difference between the maximal observed sea level and the highest predicted astronomical high tide for a single tidal cycle, which may occur at different times in that cycle (Saint Criq et al., 2022), resulting in an approximately 12 hours and 25 minutes timestep. Extreme events are of interest for this variable as they can contribute to coastal flooding. Data availability varies by station, with measurements starting during the 1970s for most. The method is applied to a period of 46 years, from 1971 to 2016. This dataset is characterized by numerous missing values, ranging from 4.4% to 61.8% of the time series depending on the station (Figure 2). These missing values are sometimes isolated but at other times extend over long periods, spanning several years in the worst cases. This dataset was already used in the previous studies of Hamdi et al. (2019) and Andreevsky et al. (2020) (albeit with a different selection of neighbor stations).

The skew surge data is assumed to be MCAR. This simplifying assumption is made to follow the methodology of Hasler et al. (2018) and to avoid having to model the missingness mechanism. As a result, the probability of missingness is assumed independent of the value of the skew surge. For the skew surge this assumption may not always be valid, as an extreme sea level or tidal event could increase the chance of measurement device failure. Nonetheless, the methodology is tested with this assumption to assess the performance offered by the D-vine.

Among the tide gauge dataset spanning the French Atlantic coast, the target station of La Rochelle is chosen to test the model. This station has 42.4% missing values over the 46 years of the dataset (Figure 2). Eight other neighbor stations for La Rochelle are selected using the method presented in Hamdi et al. (2019) and Andreevsky et al. (2020). This method selects neighboring stations according to their ratio of common extreme events with the target station. The threshold for this ratio, above which a station is included as a neighboring station, is chosen to be sufficiently low so that there is enough regional information to impute each missing date from the target station. However, some stations with too few observations were not included in the D-vine to keep its dimensionality low enough for computational reasons, or because including these stations reduced the number of dates for which a valid subset can be obtained to compute the D-vine likelihood. The selection of these neighbors is not the focus of the present study, and another method of defining a homogeneous region from the standpoint of extremes could be used instead, e.g., a measure of the pairwise tail dependence between the target station and each other station, a canonical correlation approach (Cavadias et al., 2001), or with the theory of complex networks (Han et al., 2020). Furthermore, the threshold for the metric defining the homogeneous region is chosen so as to have enough neighboring stations, but some stations are left out because of practical reasons related to fitting the D-vine with missing data. As a result, the set of neighboring stations used for imputation is not particularly sensitive to this metric, as long as it is adapted to the objective.

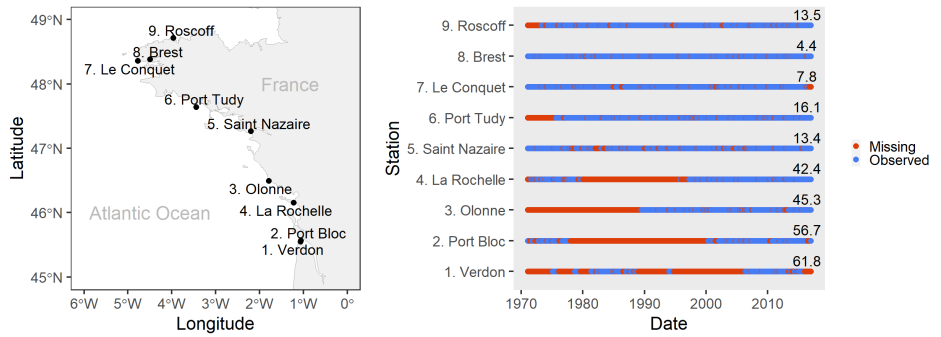

Figure 2: Location of the target station (station number 4, La Rochelle) and its eight neighboring stations along the French Atlantic coast (left). Missing and observed dates for each station, with the percentage of missing values per time series indicated on the right side (right). The station numbers correspond to their order as dimensions of the D-vine copula.

3.2 Regional skew surge modeled by D-vine

Figure 3 presents the quantile-quantile plots of the skewed generalized t margins for the nine stations. These plots show that this distribution provides good fits for the skew surge time series, even for upper extremes which are of particular interest here. The largest observation is underestimated by the models for some stations, but this observation is much larger than the second largest observation (in particular for the target station of La Rochelle). The points deviate from the main diagonal in the lower tail for some stations (most visibly for Olonne, bottom left subplot), but this is not a great concern for the skew surges as the lower extremes are not of interest.

Figure 2 presents the observed and missing dates for each of the nine stations. The order of the stations from bottom to top corresponds to the order of the dimensions of the D-vine. Ordering the stations to maximize the pairwise tail dependence and Kendall's τ (Figure 4) in the first tree of the D-vine results in this order being in accordance with the spatial organization of the stations (i.e., the order of the stations along the coast, Figure 2). However, the periods of missing values of the stations need also to be considered so that a large part of the observations is usable when computing the likelihood of the D-vine with its subsets. Furthermore, it is preferable to allow most of the observations at the target station to be included in the likelihood computation, as this dimension of the D-vine is of particular interest.

In the case of La Rochelle (station number 4), the six stations to its North (stations 5 to 9) have few missing dates, except for Olonne (station 3) during the 1970s and 1980s, while the two stations to its South have long periods of missingness (stations 1 and 2). If the stations were ordered solely on the basis of relative spatial position, Olonne (station 3) should be placed between La Rochelle (station 4) and Saint Nazaire (station 5, Figure 2). But doing so would result in a D-vine that could not be subsetted to include both the target station La Rochelle and the observed stations to its North when Olonne is missing. Instead, the Olonne station is placed to the other side of the D-vine relatively to the target station of La Rochelle (Figure 2).

Figure 4 shows the pairwise tail dependence between stations. In this figure the stations follow their

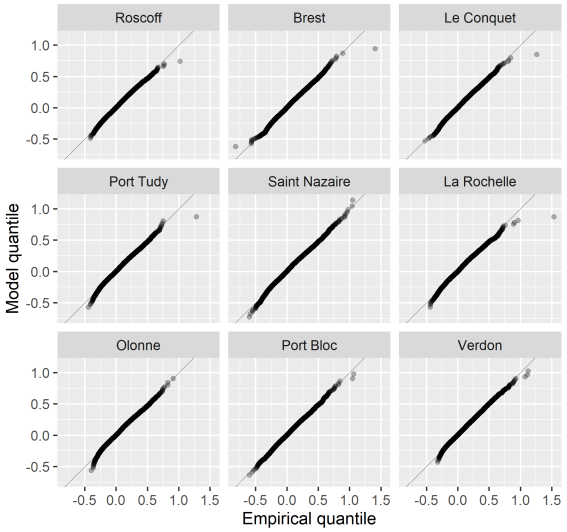

Figure 3: Quantile-quantile plots of the skewed generalized t margins for the nine stations.

ordering as dimensions of the D-vine. The highest tail dependence values are close to the main diagonal, indicating that these highest values correspond to pair of stations ordered next or close to each other in the D-vine. The stations ordered next to each other have corresponding pair-copulas on the first tree of the D-vine, which are not conditional on other stations (Figure 1). Similarly, the stations that are two orders apart correspond to a pair-copula on the second tree, solely conditional on one other station. Figure 4 also presents the pairwise Kendall's τ, for which the same conclusion can be drawn. Overall the highest values of τ are found close to the main diagonal, which indicates that the ordering of the stations is adequate.

The RJMCMC algorithm selecting and adjusting the pair-copulas is run for 5 000 iterations. The subsets of the D-vine according to the missingness patterns allow 38.18% of the dates to be used in the D-vine likelihood computation (Equation 8). Figure 5 presents the trace plots of the pair-copulas first parameter. For each pair-copula of the D-vine, the family most sampled by the RJMCMC is the most probable among the six considered, and is kept for the final model used for the subsequent imputation. The first 1 000 sampled values for these most visited families are discarded as warm-up (lighter color in Figure 5). Table 2 gives the acceptance ratios of the RJMCMC for the pair-copula parameters, with most of them being inside the 20 to 80% range recommended by Min and Czado (2010). The mean acceptance ratio of 2.98% for the jump steps is satisfactory.

The final D-vine obtained by RJMCMC is presented in Table 3. The BB1 family is selected for seven out of eight unconditional pair-copulas in the first tree of the D-vine (bottom row of Table 3.c), with only the pair-copula between station 1 and 2 having no upper tail dependence. The BB1 family has a positive upper tail-dependence (Table 1). Similarly in the second tree (second row from the bottom of Table 3.c), the only pair-copula without upper tail dependence is between dimensions 1 and 3 (conditional on dimension 2). The two families without upper pair-dependence (coded

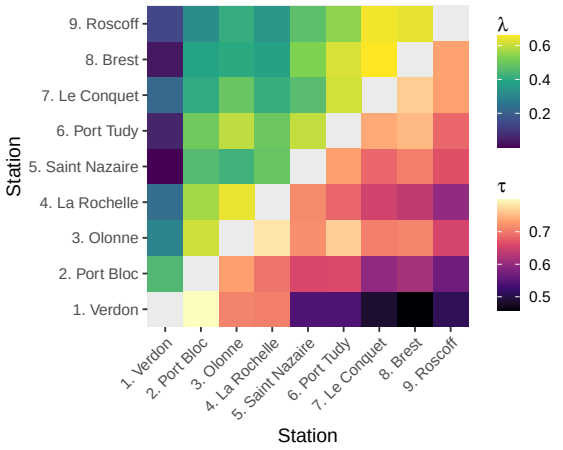

Figure 4: Pairwise tail dependence λ and Kendall's τ between stations. The tail dependence is significant at the level $\alpha = 0.05$ for every pair of stations (p-values not shown).

0 for the independence pair-copula and 1 for the Gaussian pair-copula, Table 1) are more often selected in higher trees, which are conditional on several dimensions (Figure 1). Having pair-copulas with a positive tail dependence in the lower trees of the D-vine is consistent with the pairwise tail dependence estimates of Figure 4 and shows that the final model obtained by RJMCMC is appropriate for imputation of upper extremes. For comparison, if the Gaussian family was the most selected in the lower trees of the D-vine, this would indicate that the regional information becomes less relevant for imputation as the values increase, making the model inappropriate for the extremes.

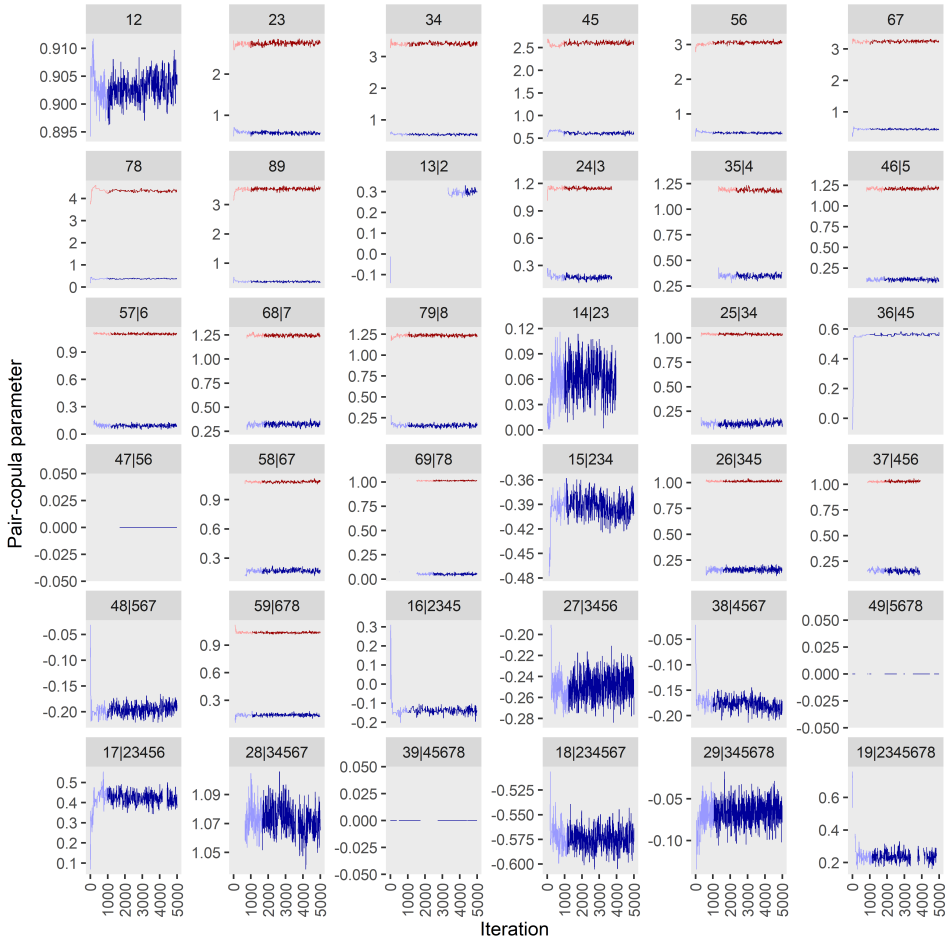

Figure 5: Trace plot of the parameters for each pair-copula. The name of each subplot indicates the two dimensions of the pair-copula and its conditioning dimensions. For each subplot, only the trace of the pair-copula family in which the algorithm stayed the longest is displayed, which explains why not all traces start at the beginning and why some are discontinuous. The blue line corresponds to the first parameter ρ of the family and the red line corresponds to the eventual second parameter ν. The independence copula has no parameter but is nonetheless indicated by a constant value of 0. The lighter part of the traces indicate the first 1000 warm-up values for this family, which are discarded.

Table 2: Acceptance ratios (in percentage) of the RJMCMC after the warm-up period, for the (a) first parameter of the pair-copulas, (b) second parameter and (c) the jumps between families. The reader is referred to Appendix A for explanations of the matrix representation of a vine copula used in this Table, with the actual matrix of the D-vine in Table 3.a. The dashes in the subtable (b) indicate that a two-parameter family has never been accepted for this pair-copula.

(a) parameter 1

26.2							
38.3	35.7						
47.3	34.1	25.9					
41.0	26.6	40.5	14.0				
23.1	24.9	33.4	33.5	25.7			
27.5	24.5	41.8	4.4	31.0	39.4		
23.4	25.1	26.2	26.8	23.1	22.4	17.7	
19.9	8.3	15.2	19.4	17.5	19.9	23.4	26.1

(b) parameter 2

13.1							
–	–						
–	21.9	28.7					
–	–	–	–				
23.1	–	33.7	35.7	–			
28.8	27.1	–	–	31.1	–		
23.3	28.8	26.0	28.0	23.2	22.3	20.2	
19.9	8.3	15.2	19.4	17.5	19.9	23.4	–

(c) jump

4.4							
0.7	1.2						
6.2	4.1	2.2					
19.8	0.0	2.5	0.0				
2.0	0.0	4.8	4.3	1.7			
8.0	2.1	4.1	0.0	5.7	1.8		
3.3	1.9	4.8	3.7	6.1	5.3	7.0	
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 3: D-vine for the skew surge with La Rochelle as target station (dimension 4), with (a) the vine copula matrix indicating the conditioning set of dimensions for each pair-copula, (b) their first parameter, (c) their families (see Table 1 for the code of each family) and (d) their second parameter. The reader is referred to Appendix A for explanations on the matrix representation of a vine copula used in this table. The values in the subtables (b, c, d) correspond to the dimensions of the subtable (a); for example the pair-copula between dimensions 1 and 9 has the family number 13, with a first parameter of 0.23 and no second parameter. The reader is referred to Figure 2 for the neighbor stations corresponding to the other dimensions. Note that the main diagonal is empty in the subtables other than (a), which is indicated by the dots. The dashes in the subtables (b) and (d) indicates that the pair-copula has no first and/or second parameter.

(a) vine copula matrix

9								
1	8							
2	1	7						
3	2	1	6					
4	3	2	1	5				
5	4	3	2	1	4			
6	5	4	3	2	1	3		
7	6	5	4	3	2	1	2	
8	7	6	5	4	3	2	1	1

(b) pair-copulas parameter 1

.								
0.23	.							
-0.07	-0.58	.						
–	1.07	0.42	.					
–	-0.18	-0.25	-0.14	.				
0.14	-0.20	0.15	0.16	-0.39	.			
0.05	0.17	–	0.56	0.12	0.06	.		
0.16	0.32	0.09	0.11	0.34	0.17	0.30	.	
0.37	0.38	0.45	0.46	0.60	0.52	0.57	0.90	.

(c) pair-copulas families

.								
13	.							
1	1	.						
0	4	13	.					
0	1	1	1	.				
17	1	7	7	1	.			
17	7	0	1	7	13	.		
17	7	7	7	17	17	1	.	
7	7	7	7	7	7	7	1	.

(d) pair-copulas parameter 2

.								
–	.							
–	–	.						
–	–	–	.					
–	–	–	–	.				
1.03	–	1.03	1.01	–	.			
1.01	1.08	–	–	1.03	–	.		
1.24	1.25	1.10	1.21	1.18	1.14	–	.	
3.54	4.35	3.24	3.05	2.60	3.42	2.74	–	.

3.3 Multiple imputation of the skew surge

The performance of the multiple imputation is evaluated by considering the observed values at the target station La Rochelle as missing. The method is first tested on a period of 100 time steps, from 2011-10-31 to 2011-12-09. This period is chosen because it has almost no missing values in the nine stations and includes the second highest observation at La Rochelle, which happened on 2011-12-16. Figure 6 presents the multiple imputation of the skew surge for four different combinations of neighboring stations considered to be missing in addition to the target station, in order to evaluate the performance of the method with respect to the information available in the region for a given date. Figure 6.a considers that only the target station is missing and the eight neighboring stations are observed. Figure 6.b considers the neighboring stations on either side of the target station (dimension 4) in the D-vine to also be missing (dimensions 3 and 5). Figure 6.c considers the stations corresponding to dimensions 2 to 6 to be missing. Lastly, Figure 6.d considers only the neighbors on either side of the target station in the D-vine to be observed (dimensions 3 and 5). For the four cases, the 100 time steps are imputed by sampling 25 000 values by MCMC for each (5 chains of 10 000 values, with the first half of each chain being discarded as warm-up). Figure 6 presents pseudo-histograms (color coded) of the posterior density for each date, the NSE and the mean of the variance of each date sample for the four combinations of observed and missing neighbors. The NSE is greater than 0.9 when all eight neighbors are observed but degrades as less regional information is available. Likewise, the mean variance and the spread of the posterior both increase as more stations are missing, indicating greater uncertainty. The order of the missing dimensions in the D-vine affect the performance. The case of Figure 6.b has only two neighbors missing, but these are ordered directly on each side of the target station, and the remaining six neighbors are observed. Figure 6.d presents an opposite case where only the two neighbors missing in Figure 6.b are observed and the remaining six are missing. Both the NSE and the mean variance show that the performance is better in the case of Figure 6.d rather than Figure 6.b, despite more regional information being available in the latter case. This comparison shows that the performance of the imputation is conditional on having the information for neighboring stations with high tail dependence and Kendall's τ with the target station (Figure 4). The $\widehat{R}$ convergence criterion is less than 1.1 for the 100 dates of the four cases (not shown). Note that the results of Figure 6 are not computed through cross-validation.

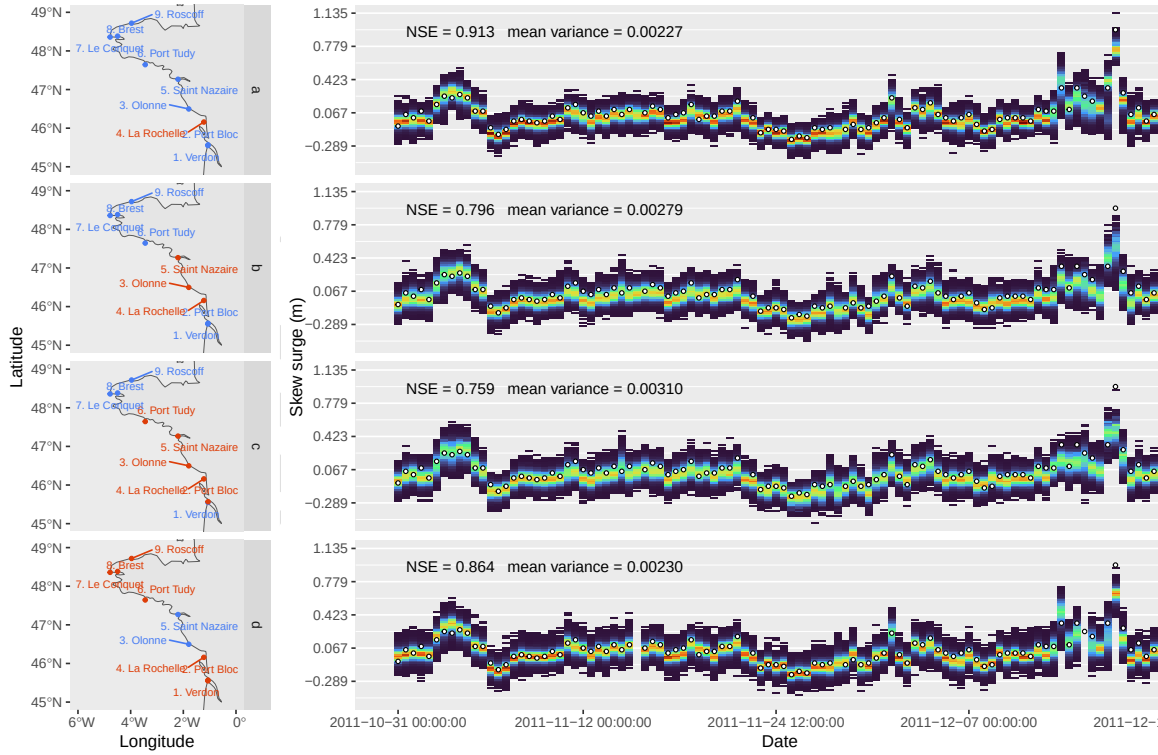

Figure 6: Multiple imputation of the skew surge at La Rochelle for four missingness patterns (rows), from 2011-10-3
The maps on the left column show which stations are considered observed (blue) and missing (red) for each test. T
shows pseudo-histograms of the multiple imputation for each date (color coded, with 25 000 sampled values per da
missingness pattern. The white dots indicate the observation value of each date. For each test, the NSE is computed
value of each date sampled values. Likewise, the mean of the variance of each date sample is computed.

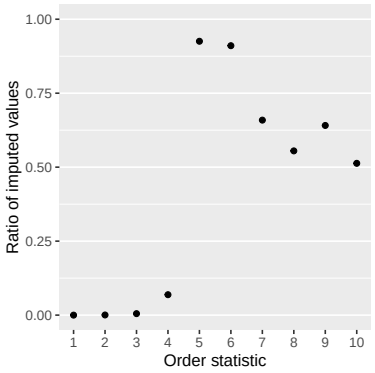

Figure 7: Ratio of imputed values in the ten largest order statistics for 2 000 replicates of the completed skew surge time series at La Rochelle.

It can be assumed that the largest order statistics are among the 42.4% missing values at La Rochelle, which the multiple imputation should be able to recover. Figure 7 shows the ratio of imputed values among the ten largest order statistic for 2 000 completed time series (each missing date being imputed with two chains of length 2 000, the first halves being discarded as warm-up). The first to fourth largest values remain observations in almost all of the completed time series, but the fifth and sixth largest values are imputed in more than 90% of cases, and this ratio is above 50% for the seventh to tenth largest values. This demonstrates that the method is suitable to impute extremes as it consistently generates values of the largest order. Common extreme value analysis involves the block maxima approach in which the extremes are defined as the largest values in blocks of time (typically years), and the threshold-based approach in which extremes are defined as exceedances of a high threshold (Coles, 2001). For both approaches, analyzing the multiple imputed time series would give a better result than restricting the analysis to the observations, as more extreme values would be included, resulting in better estimates and reduced uncertainties (but note that the uncertainty of the multiple imputation must be propagated to the uncertainty of the extreme value analysis).

For the k-fold cross-validation, the observations of the target stations are also treated as missing and are imputed. The mean of the MCMC sample of each date is used to compare the imputed values to the observation and to compute the NSE through the cross-validation. Note that these mean values are not meant to be used for a subsequent analysis of the imputed time series, as doing so would not take into account the uncertainty of the imputed values. Figure 8 shows that there is an overall good correspondence between the observations and the means of the imputed values. The differences seem to increases for the upper extremes, but not to a large extent. The NSE values for the k-fold cross-validation are $\{0.852, 0.862, 0.851, 0.844, 0.876\}$, with a mean value of 0.857. This indicates a good performance of the model when considering the bulk of the data (since this criterion is not specific to extremes).

The validity of the uncertainty obtained by the multiple imputation framework is also assessed through cross-validation. When all the observations are considered, 86.6% of them actually fall within their respective 90% credible intervals, with 6.3% and 7.1% of observations being below and above their interval bounds, respectively (Table 4). Therefore, when the whole distribution is considered, the uncertainty of the multiple imputation is only slightly underestimated, with credible

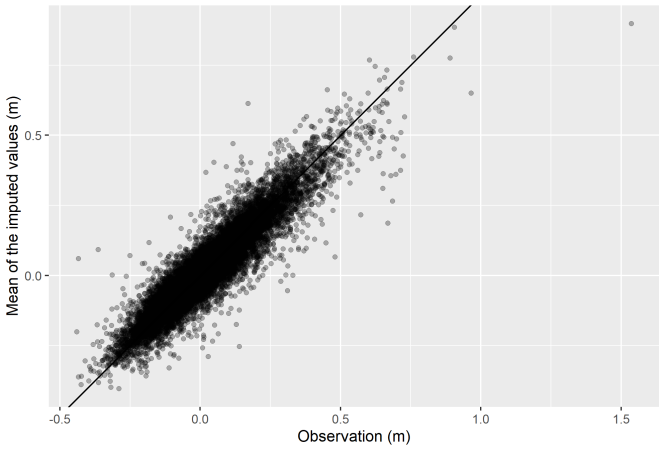

Figure 8: Observations of the skew surge at La Rochelle compared to the means of the multiple imputed values. This imputation is done through the k-fold cross-validation.

intervals that are a bit narrower than what they should be. When considering the observations above the quantile 0.9, the proportion falling inside the intervals is smaller with only 75.3% of values, indicating an underestimation of the uncertainty, and the ratio of observations above the upper bound of the credible intervals increases to 19.5%, which indicates an underestimation of the upper extremes. Although the lower extremes are not of interest for the skew surge, their imputation performance deteriorates in a manner comparable to that of the higher extremes.

Table 4: Comparison of the credible intervals of sampled values from the cross-validated model with observations. The table indicates the percentage of observations falling inside, above and below their respective 90% credible intervals. These ratios are indicated for all the observations as well as for those below the 0.1 and above the 0.9 quantiles of the skewed generalized t margin, to assess the extent to which the model performance deteriorates for extremes.

	$<$ lower bound	inside 90% CI	$>$ upper bound
$<$ quantile 0.1	17.6	78.0	4.3
all observations	6.3	86.6	7.1
$>$ quantile 0.9	5.2	75.3	19.5

4 Discussion and conclusions

A method is developed to impute the missing values of a time series at a target station using the information of neighboring stations measuring the same variable, when these neighbors can themselves have missing values. The core of the proposed approach is to model the joint distribution of the time series of the target station and its neighbor stations by a D-vine copula. The uncertainty

is accounted for by multiple imputation in a Bayesian framework.

The method is tested with the imputation of a skew surge time series at a station on the French Atlantic coast. The overall performance of the model is good, with a cross-validated NSE of 0.857. When the upper extremes are considered, the method consistently generates new values in the ten largest orders in each replicate of the imputed time series, which indicates that it is able to impute the missing extremes. The completed time series could subsequently be used for any analysis—including extreme value analysis—, with the uncertainty of the missing values being accounted for by the multiple imputation approach. However the cross-validated credible intervals reveal that the uncertainty of the imputed values is underestimated for the extremes. The performance of the model decreases for the upper extremes (which are of interest for the skew surge), but not to an extent suggesting that it is not suitable for extreme values.

The scope of this study is limited to the classical assumption of stationarity. Removing this assumption in future work would require nonstationary models for both the margins and the dependence structure. The latter could be achieved with a dynamic (i.e., nonstationary) vine copula, allowing the dependence between stations to vary in time according to covariates (Chebana and Ouarda, 2021). These covariates could be climatic variables, such as atmospheric pressure or wind speed for skew surge, with the parameters of the pair-copulas depending on them. The fitting of the D-vine by RJMCMC could be expanded to include the selection of the covariates and the adjustment of their hyperparameters (El Adlouni and Ouarda, 2009). A similar approach could be used for nonstationary models of the margins.

The imputation model could also be expanded by adding in the D-vine variables different from the one measured at the target station. These additional variables could be any that are sufficiently correlated with the time series to impute. If the imputation of the extremes is of interest, it would be preferable to use variables that are tail dependent with the one to be imputed. This could be useful in a situation where not enough neighbor stations measuring the same variable are available, or if they have too many missing values to impute every date of the target station. Adding these variables of a different nature would not require additional development of the present model, as long as they have the same timestep than the variable to impute.

Accounting for the autocorrelation of the time series and the eventual time lag of their correlation could further improve the imputation. Moreover these autocorrelation and time lag could themselves be dependent on covariates, such as storm related variables in the case of the skew surge.

The MCAR assumption may not be always valid in the case of the skew surge time series analyzed, as an extreme event can increase the chance of failure in measurement. Thus the missingness mechanism should be accounted for in the model.

Only the dates with a monotone missingness pattern are included in the computation of the D-vine likelihood (Equation 8), but more information could be used by allowing several D-vine subsets for the same date. For instance, a 6-dimensional D-vine with only the third dimensions missing could be subsetted to a pair-copula for dimensions $\{1, 2\}$ and a 3-dimensional D-vine for dimensions $\{4, 5, 6\}$. This would result in a likelihood value for each subset of a given date, which could be weighted to obtain a single likelihood value for the date.

As mentioned in the introduction, Ahn (2021) and Hasler et al. (2018) have both found that the imputation with a D-vine outperforms alternative methods, and particularly so for the extremes as a vine copula can model the eventual tail dependence between dimensions. Thus, although a proper comparison of the performance of our method with other imputation methods was outside

the scope of the article, we can assume that ours would at least offer a similar performance with alternatives for the bulk of the data and would outperform those that do not account for the eventual tail dependence for the extremes.

Funding: The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs Program, and the French Institute for Radioprotection and Nuclear Safety (IRSN) for funding this research. The authors would like also to extend their gratitude to Prof. Xuebin Zhang, the Editor, and two anonymous reviewers for their comments and suggestions, which significantly enhanced the quality of the paper.

Author contributions: Antoine Chapon: Conceptualization, Methodology, Software, Data Curation, Writing - Original Draft. Taha B. M. J. Ouarda: Writing - Review & Editing, Supervision, Project administration, Funding acquisition. Yasser Hamdi: Funding acquisition.

Appendix A Matrix representation of a vine copula

The matrix representation of a vine copula (Morales Napoles, 2009; Dißmann et al., 2013) is used for the results of Tables 2 and 3. This representation is a $d \times d$ matrix indicating which dimensions correspond to each pair-copula of the vine, as well as their conditional dimensions for the trees other than the first one. Table 5.a gives an example for a 5-dimensional D-vine. The matrix is to be read by columns, with a number on the diagonal $\{d, \dots, 1\}$ denoting a dimension of a pair-copula and another number below in the same column denoting the second dimension of this pair-copula. In this same column, any other number below both dimensions of the pair-copula indicates a conditioning dimension.

For instance, the pair-copula between dimensions 2 and 5 is indicated in the first column of Table 5.a (in bold), with every dimension below them both being the conditioning ones, here dimensions 3 and 4 (italicized). The 5-dimensional D-vine of Figure 1 also shows this 25|34 pair-copula in tree T_3. Table 5.b is the corresponding representation of a value for each pair-copula (which could be a pair-copula parameter value, the Kendall's τ of the pair-copula, etc.). The subscript notation of each value denotes the two dimensions of the pair-copula and its set of conditioning dimensions. The dots on the diagonal highlight the fact that it is empty, compared to Table 5.a. The value corresponding to the pair-copula between dimensions 2 and 5 taken as example is Table 5.a is indicated in bold.

Table 5: Matrix representation of a 5-dimensional D-vine.

(a) vine copula matrix

5				
1	4			
2	1	3		
3	2	1	2	
4	3	2	1	1

(b) value x for each pair-copula

.				
$x_{15\|234}$	.			
$\boldsymbol{x_{25\|34}}$	$x_{14\|23}$	.		
$x_{35\|4}$	$x_{24\|3}$	$x_{13\|2}$	.	
x_{45}	x_{34}	x_{23}	x_{12}	.

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). "Pair-copula constructions of multiple dependence". In: *Insurance: Mathematics and Economics* 44.2, pp. 182–198. DOI: 10.1016/j.insmatheco.2007.02.001.

Ahn, K.-H. (2021). "Streamflow estimation at partially gaged sites using multiple-dependence conditions via vine copulas". In: *Hydrology and Earth System Sciences* 25.8, pp. 4319–4333. DOI: 10.5194/hess-25-4319-2021.

Andreevsky, M., Y. Hamdi, S. Griolet, P. Bernardara, and R. Frau (2020). "Regional frequency analysis of extreme storm surges using the extremogram approach". In: *Natural Hazards and Earth System Sciences* 20.6, pp. 1705–1717. DOI: 10.5194/nhess-20-1705-2020.

Ben Aissia, M.-A., F. Chebana, and T. B. M. J. Ouarda (2017). "Multivariate missing data in hydrology – review and applications". In: *Advances in Water Resources* 110, pp. 299–309. DOI: 10.1016/j.advwatres.2017.10.002.

Cavadias, G. S., T. B. M. J. Ouarda, B. Bobée, and C. Girard (2001). "A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins". In: *Hydrological Sciences Journal* 46.4, pp. 499–512. DOI: 10.1080/02626660109492846.

Chebana, F. and T. B. M. J. Ouarda (2021). "Multivariate non-stationary hydrological frequency analysis". In: *Journal of Hydrology* 593. DOI: 10.1016/j.jhydrol.2020.125907.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Craiu, R. V. and J. S. Rosenthal (2014). "Bayesian computation via Markov chain Monte Carlo". In: *Annual Review of Statistics and Its Application* 1.1, pp. 179–201. DOI: 10.1146/annurev-statistics-022513-115540.

Davis, C. (2015). *sgt: Skewed generalized t distribution tree*. Version 2.0.

Di Lascio, F. M. L., S. Giannerini, and A. Reale (2015). "Exploring copulas for the imputation of complex dependent data". In: *Statistical Methods and Applications* 24.1, pp. 159–175. DOI: 10.1007/s10260-014-0287-2.

Dißmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013). "Selecting and estimating regular vine copulae and application to financial returns". In: *Computational Statistics and Data Analysis* 59, pp. 52–69. DOI: 10.1016/j.csda.2012.08.010.

El Adlouni, S. and T. B. M. J. Ouarda (2009). "Joint Bayesian model selection and parameter estimation of the generalized extreme value model with covariates using birth-death Markov chain Monte Carlo". In: *Water Resources Research* 45.6. DOI: 10.1029/2007WR006427.

Gao, Y., C. Merz, G. Lischeid, and M. Schneider (2018). "A review on missing hydrological data processing". In: *Environmental Earth Sciences* 77. DOI: 10.1007/s12665-018-7228-6.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2015). *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC. 675 pp.

Genest, C. and A.-C. Favre (2007). "Everything you always wanted to know about copula modeling but were afraid to ask". In: *Journal of Hydrologic Engineering* 12.4, pp. 347–368. DOI: 10.1061/(ASCE)1084-0699(2007)12:4(347).

Green, P. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 84.4, pp. 711–732. DOI: 10.1093/BIOMET/82.4.711.

Gruber, L. F. and C. Czado (2018). "Bayesian model selection of regular vine copulas". In: *Bayesian Analysis* 13.4. DOI: 10.1214/17-BA1089.

Größer, J. and O. Okhrin (2022). "Copulae: An overview and recent developments". In: *WIREs Computational Statistics* 14.3. DOI: 10.1002/wics.1557.

Hamdi, Y., C.-M. Duluc, L. Bardet, and V. Rebour (2019). "Development of a target-site-based regional frequency model using historical information". In: *Natural Hazards* 98.3, pp. 895–913. DOI: 10.1007/s11069-018-3237-8.

Hamzah, F. B., F. M. Hamzah, S. F. M. Razali, O. Jaafar, and N. A. Jamil (2020). "Imputation methods for recovering streamflow observation: A methodological review". In: *Cogent Environmental Science* 6.1. Ed. by F. Li, p. 1745133. DOI: 10.1080/23311843.2020.1745133.

Han, X., T. B. M. J. Ouarda, A. Rahman, K. Haddad, R. Mehrotra, and A. Sharma (2020). "A network approach for delineating homogeneous regions in regional flood frequency analysis". In: *Water Resources Research* 56.3. DOI: 10.1029/2019WR025910.

Hasler, C., R. V. Craiu, and L.-P. Rivest (2018). "Vine copulas for imputation of monotone non-response". In: *International Statistical Review* 86.3, pp. 488–511. DOI: 10.1111/insr.12263.

Hollenbach, F. M., I. Bojinov, S. Minhas, N. W. Metternich, S. Minhas, M. D. Ward, and A. Volfovsky (2014). "Multiple imputation using Gaussian copulas". In: Publisher: arXiv Version Number: 3. DOI: 10.48550/ARXIV.1411.0647.

Hyndman, R. J. (1996). "Computing and graphing highest density regions". In: *The American Statistician* 50.2, pp. 120–126. DOI: 10.2307/2684423.

Hyndman, R. J., J. Einbeck, and M. P. Wand (2021). *hdrcde: Highest density regions and conditional density estimation*. Version 3.4.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2017). *An Introduction to Statistical Learning*. New York: Springer. 434 pp.

Jane, R., L. Dalla Valle, D. Simmonds, and A. Raby (2016). "A copula-based approach for the estimation of wave height records through spatial correlation". In: *Coastal Engineering* 117, pp. 1–18. DOI: 10.1016/j.coastaleng.2016.06.008.

Kalteh, A. M. and P. Hjorth (2009). "Imputation of missing values in a precipitation–runoff process database". In: *Hydrology Research* 40.4, pp. 420–432. DOI: 10.2166/nh.2009.001.

Kerman, S. C. and J. B. McDonald (2013). "Skewness–kurtosis bounds for the skewed generalized t and related distributions". In: *Statistics and Probability Letters* 83.9, pp. 2129–2134. DOI: 10.1016/j.spl.2013.05.028.

Knoben, W. J. M., J. E. Freer, and R. A. Woods (2019). "Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores". In: *Hydrology and Earth System Sciences* 23.10, pp. 4323–4331. DOI: 10.5194/hess-23-4323-2019.

Little, T. D., T. D. Jorgensen, K. M. Lang, and E. W. G. Moore (2014). "On the joys of missing data". In: *Journal of Pediatric Psychology* 39.2, pp. 151–162. DOI: 10.1093/jpepsy/jst048.

Min, A. and C. Czado (2010). "Bayesian inference for multivariate copulas using pair-copula constructions". In: *Journal of Financial Econometrics* 8.4, pp. 511–546. DOI: 10.1093/jfinec/nbp031.

– (2011). "Bayesian model selection for D-vine pair-copula constructions". In: *Canadian Journal of Statistics* 39.2, pp. 239–258. DOI: 10.1002/cjs.10098.

Morales Napoles, O. (2009). "Bayesian belief nets and vines in aviation safety and other applications". PhD thesis. Technische Universiteit Delft.

Nagler, T., U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, and T. Erhardt (2022). *VineCopula: Statistical inference of vine copulas*. Version 2.4.4.

Reiss, R. and M. Thomas (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. 3rd ed. Birkhäuser. 516 pp.

Roberts, G. O. and J. S. Rosenthal (2009). "Examples of adaptive MCMC". In: *Journal of Computational and Graphical Statistics* 18.2, pp. 349–367. DOI: 10.1198/jcgs.2009.06134.

Saint Criq, L., E. Gaume, Y. Hamdi, and T. B. M. J. Ouarda (2022). "Extreme sea level estimation combining systematic observed skew surges and historical record sea levels". In: *Water Resources Research* 58.3. DOI: 10.1029/2021WR030873.

Serinaldi, F. and C. G. Kilsby (2015). "Stationarity is undead: Uncertainty dominates the distribution of extremes". In: *Advances in Water Resources* 77, pp. 17–36. DOI: 10.1016/j.advwatres.2014.12.013.

Tootoonchi, F., M. Sadegh, J. O. Haerter, O. Räty, T. Grabs, and C. Teutschbein (2022). "Copulas for hydroclimatic analysis: A practice-oriented overview". In: *WIREs Water* 9.2. DOI: 10.1002/wat2.1579.

Valle, D. and D. Kaplan (2019). "Quantifying the impacts of dams on riverine hydrology under non-stationary conditions using incomplete data and Gaussian copula models". In: *Science of The Total Environment* 677, pp. 599–611. DOI: 10.1016/j.scitotenv.2019.04.377.

Yan, J. (2007). "Enjoy the joy of copulas: With a package copula". In: *Journal of Statistical Software* 21.4. DOI: 10.18637/jss.v021.i04.

Author Statement

Author contributions: Antoine Chapon: Conceptualization, Methodology, Software, Data Curation, Writing - Original Draft. Taha B. M. J. Ouarda: Writing - Review & Editing, Supervision, Project administration, Funding acquisition. Yasser Hamdi: Funding acquisition.

Conflict of Interest

Conflicts of Interest Statement

Manuscript title: Imputation of missing values in environmental time series by D-vine copulas

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Antoine Chapon, Taha B. M. J. Ouarda, Yasser Hamdi