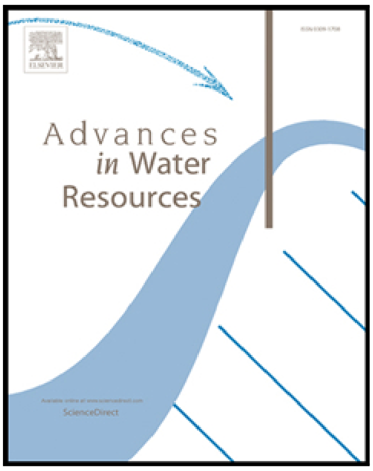

Multivariate missing data in hydrology – Review and applications

M.-A. Ben Aissia , F. Chebana , T.B.M.J. Ouarda

PII: S0309-1708(16)30623-6
DOI: 10.1016/j.advwatres.2017.10.002
Reference: ADWR 2964

To appear in: *Advances in Water Resources*

Received date: 6 November 2016
Revised date: 25 April 2017
Accepted date: 3 October 2017

Please cite this article as: M.-A. Ben Aissia , F. Chebana , T.B.M.J. Ouarda , Multivariate missing data in hydrology – Review and applications, *Advances in Water Resources* (2017), doi: 10.1016/j.advwatres.2017.10.002

Highlights:

- Highlight the importance of treating MD in multivariate hydrological frequency analysis
- Reviewing and applying multivariate imputation methods and by comparing univariate and multivariate imputation methods.
- An application is carried out for multiple flood attributes on three sites in order to evaluate the performance of the different methods based on the leave-one-out procedure.
- The results indicate that, the performance of imputation methods can be improved by adopting the multivariate setting, compared to mean substitution and interpolation methods, especially when using the copula-based approach.

Multivariate missing data in hydrology – Review and applications

M.-A. Ben Aissia[1]

F. Chebana[1]*

T. B. M. J. Ouarda[1,2]

[1] Centre Eau Terre Environnement (ETE), Institut national de la recherche scientifique (INRS), 490, Rue de la Couronne, Quebec, Qc, Canada, G1K 9A9.

[2] Institute Center for Water and Environment (iWATER), Masdar Institute of Science and Technology. PO Box 54224, Abu Dhabi, UAE.

* Corresponding author: fateh.chebana@ete.inrs.ca

May 2017

Abstract

Water resources planning and management require complete data sets of a number of hydrological variables, such as flood peaks and volumes. However, hydrologists are often faced with the problem of missing data (MD) in hydrological databases. Several methods are used to deal with the imputation of MD. During the last decade, multivariate approaches have gained popularity in the field of hydrology, especially in hydrological frequency analysis (HFA). However, treating the MD remains neglected in the multivariate HFA literature whereas the focus has been mainly on the modeling component. For a complete analysis and in order to optimize the use of data, MD should also be treated in the multivariate setting prior to modeling and inference. Imputation of MD in the multivariate hydrological framework can have direct implications on the quality of the estimation. Indeed, the dependence between the series represents important additional information that can be included in the imputation process. The objective of the present paper is to highlight the importance of treating MD in multivariate hydrological frequency analysis by reviewing and applying multivariate imputation methods and by comparing univariate and multivariate imputation methods. An application is carried out for multiple flood attributes on three sites in order to evaluate the performance of the different methods based on the leave-one-out procedure. The results indicate that, the performance of imputation methods can be improved by adopting the multivariate setting, compared to mean substitution and interpolation methods, especially when using the copula-based approach.

1 Introduction

The availability of hydrological data of adequate quality and length is vital for optimal water resources planning and management. In practice, hydrological studies suffer from missing data (MD) caused for instance by equipment failures, errors in measurements, budget cuts, and natural hazards (Kalteh and Hjorth 2009). This is generally the case for hydrological variables such as rainfall and streamflow, particularly for extreme conditions such as in remote watersheds where equipment failures are often detected and fixed with a significant delay.

Generally, hydrological data are characterized by several correlated variables, such as *Q* and *V* (e.g. Ouarda et al. 2000; Zhang and Singh 2006; Chebana and Ouarda 2011a). These correlated variables are considered simultaneously in a multivariate framework, see e.g. Chebana (2013) for an explanation of the importance and the justification of jointly considering all variables associated to an event such as in hydrological frequency analysis (HFA).

HFA, is an essential and commonly used approach for the analysis and prediction of hydrological extreme events. In HFA, we are frequently faced with the MD problem which can affect the reliability of the results if it is not correctly handled. Generally, before proceeding with any hydrological analysis, it is relevant to ensure that the quality of the data is adequate through an exploratory analysis, outlier detection and MD estimation. The presence of MD was highlighted for several hydrometeorological variables such as streamflow (Ng et al. 2009) and precipitation (Makhnin and McAllister 2009).

Given two hydrological variables *X* and *Y*, in general three multivariate MD situations may occur: (i) only the *X* set contains MD and the *Y* set is complete and vice-versa; (ii) both *X* and *Y* series contain MD but not for the same event; and (iii) *X* and *Y* contain MD for the same event. In addition, when a data gap exists in a given station, associated data are generally observed in one or more neighboring stations, such as in other tributaries of the same river. However, for particular applications and variables, some of the situations listed above are more common. A typical case is for flood peak (Q) and volume (V) where MD are usually caused by missing streamflow observations during the flood event. Therefore, the situation where only Q is missing is less common.

Multivariate HFA is composed of four main steps: (a) carry out the exploratory analysis including outlier detection, MD estimation and descriptive analysis, (b) verify HFA assumptions, (c) model the extreme events and estimate the corresponding parameters, and (d) estimate and analyze the risk (see Table 1 for an overview). Recently, Chebana et al. (2013) described these steps and focused on testing multivariate trends in HFA. Under the framework of step (a), the statistical features and the shape of the data are investigated in a multivariate setting by Chebana and Ouarda (2011b). However, MD estimation is generally ignored in multivariate HFA. Consequently, MD estimation in the multivariate setting of HFA is currently a missing step.

Ignoring the MD estimation in multivariate HFA may lead to a loss of information which may result in inappropriate decisions regarding, for instance, the design of hydraulic structures. Consequently, it is necessary to estimate MD in multivariate HFA to avoid or reduce the unnecessary construction costs associated with overestimation and the potential loss of human lives associated with underestimation.

The present paper is organized as follows. The literature review concerning missing data is presented in Section 2. Section 3 deals with the general technical considerations associated to MD including the uncertainty in their estimation. Descriptions and definitions of the considered imputation methods are presented in Section 4. Section 5 contains the applications of these methods. Conclusions are reported in Section 6.

2 Literature review

The MD estimation is also called infilling (e.g. Abudu et al. 2010), reconstruction (e.g. Kim and Pachepsky 2010), completion (e.g. Ramos-Calzado et al. 2008), patching (e.g. Hughes and Smakhtin 1996) or imputation (e.g. Schneider 2001). It is largely studied in the time domain analysis, i.e. analyzing the data over a time period (see e.g. Gyau-Boakye and Schultz 1994; Hughes and Smakhtin 1996; Abebe et al. 2000; Han and Li 2010; Marlinda et al. 2010). However, in frequency analysis, the MD handling problem has received less attention (e.g. Peterson et al. 2011). Table 2 summarizes the different MD frameworks.

Several imputation methods have been developed to treat MD in both time domain analysis and frequency analysis. In time domain analysis, the use of imputation methods has received considerable attention in hydrology and elsewhere in statistics (see e.g. Gleason and Staelin 1975; Jeffrey et al. 2001; Ng et al. 2009; Honaker and King 2010). However, the imputation of MD in frequency analysis has received less attention (see e.g. Kelly et al. 2004; Erol 2011; Peterson et al. 2011). In frequency analysis studies, MD estimation is largely treated in the univariate setting (e.g. Kodituwakku et al. 2011, in a health study) whereas in the multivariate setting, studies are relatively rare (e.g. Kelly et al. 2004, in a biological study).

Handling MD in multivariate HFA is generally ignored or treated separately for each series. The most common practices in HFA are to ignore missing observations (see e.g. Overeem et al. 2009; Westra et al. 2012) or to impute each missing value by the mean of the variable (see e.g. Özçelik and Benzeden (2010) and Peterson et al. (2011)). Fleig et al. (2011) used more sophisticated univariate methods, such as, interpolation and regression to estimate MD in HFA. Consequently, MD estimation in multivariate HFA has not yet been adequately studied. Multivariate imputation methods are useful, in particular in hydrology, to improve the quality of the estimation and to provide more accurate imputed values that take variable dependence into account.

In hydrology, imputation techniques of time domain analysis are extensively treated in the univariate and multivariate setting. Table 3 shows an overview of the main imputation techniques in missing hydrological data with a number of references, as well as the advantages and disadvantages of each method. Univariate methods are largely treated in hydrology and include mean or subgroup mean imputation (e.g. Linacre 1992), time series analysis (e.g. Lettenmaier 1980), spatial or temporal interpolation (e.g. Filippini et al. 1994), regression (e.g. Kuligowski and Barros 1998), hot-deck imputation (Srebotnjak et al. 2012) and inverse distance heightening method (ASCE 1996).

In time domain analysis, multivariate techniques of MD estimation are largely considered in hydrology and can be gathered in three groups: (1) multivariate versions of univariate methods including, for instance, the multivariate version of the regression model (e.g. Simonovic 1995) or the time series analysis approach (e.g. Bennis et al. 1997); (2) data driven methods including Artificial Neural Networks (ANNs), e.g. Raman and Sunilkumar (1995) and the k-nearest neighborhood (K-NN) approach, e.g. Kalteh and Hjorth (2009); and (3) model-based approaches including the Expectation-Maximization (EM) algorithm and the Multiple Imputation (MI) approach (e.g. Ng et al. 2009). In HFA, handling MD is generally treated in the univariate setting using, for instance, mean substitution (MS) or linear interpolation (LI). A number of multivariate imputation methods have been applied in hydrological time domain analysis but they have not been used in the HFA multivariate setting. In the case of streamflows, some authors resorted to streamflow estimation techniques at ungauged sites (Vogel and Fennessey, 1994; Shu and Ouarda, 2012) to fill in missing values, based on a number of source stations, and into several target stations. The approaches are often based on the use of the Flow-Duration-Curve (FDC) approach. These methods are not considered in the present study because of their lack of generality (specific to streamflows). Copula-based methods are well known to be very efficient in modeling the dependence structure and have been used in various applications, in particular in hydrology (e.g. Dupuis 2007, Chebana and Ouarda 2011, Requena et al. 2013, Hamdi et al. 2016). In the multivariate context of MD imputation, one can find a number of recent studies (Bárdossy and Pegram 2014, Ding et al. 2016, Käärik et al. 2009 and Marta L. Di Lascio et al. 2015). Several families of copulas are developed, such as Gaussian and Archimedean copulas, and can be used to impute missing values for an incomplete dataset.

Several studies focused on comparing MD imputation methods in multivariate time series analysis, such as Kalteh and Hjorth (2009) and Coulibaly and Evora (2007). Kalteh and Hjorth (2009) compared five multivariate methods to impute missing values in precipitation-runoff databases. The considered methods are self-organizing maps (SOM) which is an unsupervised ANNs method, multilayered ANN, multivariate K-NN, regularized EM algorithm (REGEM) and MI method. They found that SOM and the multivariate K-NN methods provide the most robust and reliable results. The ability of SOM to produce reliable estimates of missing hydrological data is also demonstrated in Adebayo and Rustum (2012) and Mwale et al. (2012). On the other hand, Coulibaly and Evora (2007) compared six ANN methods to impute missing daily weather records. These methods are the multilayer perceptron (MLP) network, the time-lagged feed forward network (TLFN), the generalized radial basis function (RBF) network, the recurrent neural network (RNN) and its variant the time delay recurrent neural network (TDRNN), and the counter propagation fuzzy-neural network (CFNN). They found that MLP, TLFN and CFNN methods can provide the most accurate estimates of the missing precipitation values.

In the present study several univariate and multivariate methods are used to investigate the performance of multivariate methods against univariate ones in the case of several types of MD patterns and different dependence levels. In the univariate context several methods can be used (Table 3). The MS and LI methods have been used in HFA studies such as Fleig et al. (2011) and Peterson et al. (2011). Another method that is used is the stepwise regression tree method (SRT) which is a regression model in several nodes. This method was shown to be an efficient technique to impute univariate MD (see e.g. Kim and Pachepsky 2010).

According to Table 3, five multivariate imputation methods are generally used in hydrology in time domain analysis. The first method is the ANN method and its several variants (see e.g. Coulibaly and Evora 2007). According to Kalteh and Hjorth (2009), Adebayo and Rustum (2012) and Mwale et al. (2012), among the ANN methods, the SOM method leads to good performances. The second one is K-NN which is not recommended in the context of this paper since it consists in replacing MD by observed data from the same vector of the series (i.e. in (Q,V,D) series, replace MD in Q by observed values in V or D is not realistic). As a third method, we have the EM algorithm. It was originally developed by Dempster et al. (1977) and received some modifications such as the Expectation Conditional Maximization (e.g. Meng and Rubin 1993), the Expectation Conditional Maximization Either (e.g. Liu and Rubin 1994), Alternating Expectation Conditional Maximization (e.g. Meng and Van Dyk 1997), Parameter-Expanded expectation-maximization (e.g. Liu and Rubin 1998) and the REGEM algorithm (e.g. Schneider 2001). The latter is the most commonly used in hydrology (e.g. Kalteh and Hjorth 2009). The next method is the MI which consists in the imputation of several values (usually 3-5 times) for each MD using an appropriate imputation model (e.g. Patrician 2002). The MI method is rarely used in hydrology (e.g. Kalteh and Hjorth 2009). Finally, the copula-based method is among the recent ones and required the fitting of a multivariate distribution to the available data (including the copula and the margins).

Several softwares handling MD imputation in the multivariate context are available. In particular, a number of R-packages can be used depending on the imputation method, for instance *AMELIA*, *CLASS*, *MICE*, *NORM*, *VIM*, *MI or CoImp*. A number of other packages have also been developed for other environments for example: *S+MissingData* for S-PLUS, *ice* for Stata, *PROCMI* for SAS and the *SOM toolbox* for Matlab.

3 General MD considerations

Based on the previous literature review, seven imputation methods are used in the HFA framework in this paper (MS, LI, SRT, SOM, REGEM, MI and Copulas MD). These methods are described in the next section.

Let $\left(X_i\right)_{i=1,\ldots,n} = \left(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(\mathrm{d})}\right)'_{i=1,\ldots,n}$ be a continuous *d*-dimensional sample from a stochastic process $(d \geq 1, n \geq d)$, where “,” denotes the matrix transpose. Let $\left(x_i\right)_{i=1,\ldots,n} = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)}\right)'_{i=1,\ldots,n}$ be an observation from X_i, such as flood peak *Q* and volume *V*, at time *i*. Each series $X^{(k)}, k=1,\ldots,d$ can be written as $X^{(k)} = \left(X_{obs}^{(k)}, X_{mis}^{(k)}\right)$where $X_{obs}^{(k)}$ represents the observed part and $X_{mis}^{(k)}$ denotes the missing part. Before imputing MD it is important to know how and where the MD occurred in the series. For this we refer respectively to MD mechanisms and MD patterns.

3.1 Missing data mechanisms

The MD mechanism determines how the MD is produced. It is a potential factor that could affect the imputation results (Zhu et al. 2012). There are three types of MD mechanisms (Little and Rubin 2002):

- *Missing completely at random (MCAR)*

In this case, MD is unrelated to both the observed or unobserved values in the series. Let $x_i^{(k)}$ be the value of $X^{(k)}, k=1,...,d$, at time i and $p(x_i^{(k)})$ the probability that $x_i^{(k)}$ is missing. Under MCAR assumption, $p(x_i^{(k)})$ can be expressed as

$$p\left(x_i^{(k)} \middle| X_{obs}^{(k)}, \mathrm{X}_{mis}^{(k)}\right) = p\left(x_i^{(k)}\right) \tag{1}$$

meaning that $p\left(x_i^{(k)}\right)$ is independent of both the observed $X_{obs}^{(k)}$ and unobserved $X_{mis}^{(k)}$ parts of $X^{(k)}$.

- *Missing at random (MAR)*

It refers to the case where the incomplete data depends on the observed values but not on the unobserved ones. The probability $p(x_i^{(k)})$ can be expressed as

$$p\left(x_i^{(k)} \middle| X_{obs}^{(k)}, \mathrm{X}_{mis}^{(k)}\right) = p\left(x_i^{(k)} \middle| X_{obs}^{(k)}\right) \tag{2}$$

The MAR mechanism occurs when the probability of an observation having a missing value for a component may depend on the available values, but not on the MD themselves.

- *Not missing at random (NMAR)*

In this mechanism, the probability of an observation having a missing value could depend on the observed values as well as the unobserved ones.

Most of MD in hydrological modeling may be attributed to MCAR or MAR cases (Gill et al. 2007; Kalteh and Hjorth 2009). They are also called ignorable response mechanisms because the reasons for MD can be ignored during the analysis. Model-based methods require the MCAR or MAR assumption (Kalteh and Hjorth 2009).

3.2 Missing data pattern

MD imputation methods depend also on the MD pattern which describes where data are observed or missed in the series. Some of these methods apply to any pattern of MD, whereas others are limited to special ones. Several MD patterns exist in the literature such as *multivariate nonresponse* where a set of series are all observed or missing on the same set of cases, *monotone pattern* where the series can be arranged so that all $X^{(j+1)},...,X^{(k)}$ are missing for cases where $X^{(j)}$ is missing, for all j = 1,…, k-1 and *general pattern* where the MD typically have a random pattern (see e.g. Little and Rubin 2002 for more details).

The methods for handling MD in the case of multivariate nonresponse, or monotone patterns can be easier than the methods for general pattern. In the present study, we consider multivariate hydrological datasets where MD are inside the series. Figure 1 illustrates three possibilities of MD patterns in the bivariate case. The three possibilities are: (i) only one missing value is present in one of the two series, (ii) two missing values are present and are located at the same event, and (iii) two missing values are present but are not at the same event. These three different possibilities are not treated using the same methods.

3.3 Uncertainty in MD estimation

As indicated above, missing values are a reality in hydrology and hence they can be either imputed or ignored. In the latter case, given the short data records commonly available in hydrology, not imputing missing data may have more impact on the analysis than in other application fields. The impact of missing data on HFA depends on their frequency and their (unknown) magnitude. For instance, a single ordinary missing value can probably have little impact on an analysis. In addition, the magnitude of the impact may depend on the objectives of the study. For instance, in a regional HFA a missing value in a given site may lead to discarding the corresponding values in all sites.

Not imputing missing data may have a negative impact on the analysis. On the other hand, imputed values have an uncertainty associated to their estimates. The issue of uncertainty in the context of missing data is important and should not be neglected as demonstrated in a number of studies. For instance, recently, Frazier et al. (2016) indicated that the process of estimating missing data is a major source of uncertainty. Indeed, the imputed values are not observations but a statistically plausible set of estimated values based on other information (Dziura et al., 2013). Bárdossy and Pegram (2014) indicated that uncertainty on those estimated values should be considered for any subsequent application. They highlighted the importance of quantifying this uncertainty. Indeed, handling inappropriately missing observations can bias the statistical inference and lead to possibly incorrect hydrological models because the obtained inference fails to reflect any uncertainty due to missing data (Schafer, 1997 and Ng et al., 2009). Schafer and Olsen (1998) argued that any analysis that ignores the uncertainty of missing-data prediction will lead to standard errors that are too small, p-values that are artificially low, and rates of Type I error that are higher than nominal levels.

Even though the missing data uncertainty is important, it is generally not treated or only mentioned in a number of studies. In some papers, the issue is discussed very briefly such as in the above cited references. However, this topic is rarely treated in depth. For instance, to the best knowledge of the authors, Little and Rubin (2002) may be the only reference dedicating a whole chapter to this issue. In the multivariate setting, the topic of the uncertainty associated to missing data is not mentioned, and even less in the copula-based approach, except in Bárdossy and Pegram (2014).

Evaluating uncertainty due to missing data can be done through bias, standard deviation (variance) or confidence intervals. The multiple imputation methods are advantageous in this way. The copula method can also provide uncertainty since it is based on a distribution (copula). However, not all imputation methods allow to get uncertainty. In Little and Rubin (2002, chapter 5), the authors focused on deriving estimates of uncertainty that incorporate the added variance due to missing values. However, they indicated that in many applications the missing value bias is often more crucial than that of the variance. They presented four general approaches to account for the additional uncertainty. The first approach employs explicit variance formulas whereas the second one involves modified imputations. The third approach is represented by the use of resampling methods. In this approach, uncertainty is estimated from the variability of point estimates of the parameters from a suitable set of samples drawn from the original sample. It includes in particular the bootstrap and jackknife methods. Finally, multiple imputed data sets, such as a multiple imputation (MI) method, is the approach consisting in creating multiple completed data sets. This idea provides consistent standard errors under broad classes of imputation procedures. Because of the uncertainty issue, the MI approach is preferred in a number of studies, e.g. Schafer and Graham (2002) and Dziura et al. (2013).

Bárdossy and Pegram (2014) discussed briefly the importance of the uncertainty issue and its evaluation using copulas. One of the most important advantages of the copula based approach, compared to the other procedures, is that it delivers not only expected values but also full conditional distributions for the missing values.

It is worth it to underline that the uncertainty can be assessed for the missing value itself as an estimated quantity, but also for any other statistical parameter or any subsequent inference. In the multivariate setting, the uncertainty can be obtained either for each component separately as in the univariate setting, or jointly for all components including their dependence. Even though, the latter is more realistic, it may be more complex to obtain since it involves complex notions without a unique definition, such as multivariate versions of quantiles (see Chebana and Ouarda 2011a) or multivariate scales (see Chebana and Ouarda 2011b). The results would be confidence zones instead of intervals.

3.4 Employed Softwares

In the present paper, Matlab codes are developed for MS, Li and SRT methods. The SOM imputation method was carried out by the SOM toolbox which can be downloaded from the site: http://research.ics.aalto.fi/software/somtoolbox/. The Matlab code used for the REGEM method can be downloaded from the site: http://www.clidyn.ethz.ch/imputation/ index.html. Finally, for the MI and Copula methods, the R-packages *NORM* and *CoImp* are used respectively.

3.5 Performances of imputation methods

To evaluate the accuracy of the imputation methods, their performances are evaluated through a jackknife resampling procedure. It consists in considering each value as a missing one by removing it temporarily from the series. The criteria employed to evaluate the performances are the Relative Root-Mean Squared Error $(RRMSE)$(see e.g. Chebana and Ouarda 2008) and the mean relative bias (MRB) (see e.g. Beaulieu et al. 2012) defined by:

$$RRMSE = \frac{100}{n}\sqrt{\sum_{i=1}^{n}\left(\frac{\hat{x}_i - x_i}{x_i}\right)^2}, \quad x_i \neq 0 \tag{3}$$

$$MRB = \frac{100}{n}\sum_{i=1}^{n}\left(\frac{\hat{x}_i - x_i}{x_i}\right), \quad x_i \neq 0 \tag{4}$$

where $\hat{x}_i$ is the imputed value and x_i is the observed one.

These performance measures were chosen to provide a measure for the deviation of the estimated values from the observations (RRMSE) and to indicate whether the imputation method may tend to overestimate or underestimate the observations (MRB). These measures are also widely used in HFA. However, a variety of other measures can be considered, such as those where one can remove more than one value at a time.

4 Considered imputation methods

4.1 Mean substitution (MS)

The MS method is the simplest imputation technique. It consists in replacing each missing value in the series $X^{(k)}$, k=1,…,d by the corresponding mean of each component. This imputation method has been used in multivariate HFA studies (see for instance Wang et al. (2009) and Kao and Chang (2012)).

4.2 Linear Interpolation (LI)

One of the simplest methods to impute MD is the LI method. It consists of drawing a straight line between observed values before and after the gap and then estimating MD values by interpolation. In univariate regional HFA, this method was used by Fleig et al. (2011). However, to the authors’s best knowledge, it was not used to estimate MD in multivariate HFA.

4.3 Stepwise Regression Trees (SRT)

The SRT algorithm developed in Huang and Townshend (2003) consists in fitting, in each node of a regression tree, a stepwise regression model (e.g. Miller 2002). Initially, all the data are in the first node of the tree and the partition of the samples into subsets is made recursively until no remaining nodes can be further split. The split of a node into two subsets is made when splitting reduces the residual sum of squares (RSS), such that:

$$RSS = \sum_{i=1}^{n} \left(\hat{x}_i - x_i \right)^2, x \in R^d \qquad (5)$$

where *n* is the number of observations in the subset; and x_i and $\hat{x}_i$ represent the observed and predicted series from fitting a stepwise regression model. The *RSS* in a given node, before splitting, is noted *RSSN*. The *RSS* of the left and right node after splitting are computed and denoted *RSSL* and *RSSR*, respectively. The sum of *RSSL* and *RSSR* is the total residual sum of squares denoted by *RSST*. The RSST is computed for all possible splits, and the one leading to the smallest *RSST* is conserved and noted *RSSM*. If the split improves the predictions, it will be conserved. Therefore, a measure of the improvement (*I*) from splitting is

$$I = \frac{RSSN - RSSM}{RSSN} 100\% \qquad (6)$$

The split is conserved if *I* is larger than the fixed minimum improvement values (I_{min}) and if there are as many observations in the nodes resulting from splitting as the predefined minimum node size (n_{min}). This procedure of splitting continues recursively until all nodes are considered terminal, i.e. the number of observations in that node (*n*) is smaller than n_{min} or *I* is smaller than I_{min}. To split a node, I_{min} is fixed to 1%. This value was also used in Huang and Townshend

(2003) and Beaulieu et al. (2012). We use n_{min} of : 3, 4, 5, 6, 7, 10 or 15 observations. The value of n_{min} leading to the model with the best performance is chosen.

When SRT is used, the MDs are estimated using a regression model into the corresponding node.

4.4 Self-Organizing map (SOM)

The SOM method, also called feature map or Kohnen map, is the most widely used of the ANN algorithms designed for unsupervised pattern recognition applications (Kohonen et al. 1996). The ability of the SOM technique in the estimation of missing univariate and multivariate hydrological data was demonstrated in several studies, see e.g. Adebayo and Rustum (2012) and Mwale et al. (2012). However, these applications were made in time domain analysis. In the present study, this method is applied in the HFA context. The principal goal of the SOM is to transform, in a nonlinear way, a high dimensional input layer to a two dimensional discrete map. A typical structure of a two-dimensional SOM consists of a multi-dimensional input layer and the competitive or output layer. Both of these layers are fully interconnected. The neurons in the input layer are connected to all output layers via weight vectors. Therefore, similar input patterns are represented by the same output neurons, or by one of their neighbors (Back et al. 1998). The SOM can be viewed as a tool for reducing the amount of data by clustering nonlinear statistical relationships between high dimensional data into a simple relationship on a two dimensional display (Kohonen et al. 1996). This method preserves the most important relationship of the original data elements. This implies that, during the mapping, not much information is lost which makes the SOM method a very good tool for prediction. Note that, for prediction values outside the range used for the extrapolation, the SOM method cannot be used. This is mainly due to the fact that, as it is the case with most data-driven methods, SOM is a very poor extrapolator

(Adeloye et al. 2011). It has a limited capacity to predict values which have not been observed in the past (rarely large magnitude for instance).

The training of the SOM is iterative and is hence similar to a sequential training algorithm. In the training algorithm, the whole database is presented to the map before any updates are made while in the sequential training, the weights are updated vector by vector. The SOM procedure can be summarized as follows: at the beginning of the training, weight vectors must be initialized to each neuron and the input vectors are compared with the SOM neurons to find the closest matches which are called the best matching units (BMUs). The Euclidean distance is the most commonly used criterion. This procedure must be iterated several times until the optimal number of iterations is reached or the specified error criterion is attained. The MDs are obtained as their corresponding values in the BMU.

4.5 Regularized Expectation-Maximization algorithm (REGEM)

The Expectation Maximization (EM) algorithm is a very general iterative method for Maximum Likelihood (ML) estimation in MD problems (Dempster et al. 1977). The EM algorithm is proposed for several contexts. The REGEM method (Schneider 2001), as a particular form of the EM algorithm, is based on estimated regression models between missing and available data. The REGEM method is an iterative algorithm based on E step (Expectation) and M step (Maximization). This method consists of: (1) replacing MD by estimated values; (2) given the observed data and current estimated regression parameters, estimating the mean vector and the covariance matrix of the data; (3) Re-estimating the MD assuming the new parameters are correct. The algorithm consists in iterating these 3 steps until convergence i.e. when the variations of the mean vector and the covariance matrix are lower than a predefined threshold.

The initial estimates of the model parameters are obtained from the complete database after substituting the missing values with the mean.

During the past few decades, the REGEM algorithm was intensively used for MD imputation on multivariate normally distributed series (e.g. Little and Rubin 2002). However, the literature dealing with the application of the REGEM algorithm in hydrology is very sparse (e.g. Kalteh and Hjorth 2009). The REGEM algorithm has not yet been applied in HFA.

4.6 Multiple imputation (MI)

MI is a fairly straightforward procedure for imputing multivariate MD (Rubin 1987). It provides a useful strategy for dealing with datasets that have MD (Klebanoff and Cole 2008; Sterne et al. 2009). It has been and continues to be developed theoretically and adapted and implemented in numerous statistical problems such as measurement error (e.g. Yucel and Zaslavsky 2005; Reiter and Raghunathan 2007). The basic idea of this method is to first generate several completed data sets by generating several possible values for each MD, and then to analyze each dataset separately. The number of completed datasets to be generated depends on the extent of the missing data. However, according to Schafer (1997), five completed datasets typically provide unbiased estimates. The Schafer’s (1999) NORM software, which was used in the present study, uses the data augmentation algorithm to generate five possible values for each MD. The multivariate normal distribution is used to generate imputations. The data augmentation algorithm treats parameters and MD as random variables and simulates random values of parameters and MD from their conditional distribution.

Like the REGEM method, the MI technique has been used intensively for MD imputation on multivariate normally distributed variables (e.g. Little and Rubin 2002). However, its application in hydrology remains very limited (e.g. Kalteh and Hjorth 2009) especially in multivariate HFA.

4.7 Copula-based methods

An appropriate approach to deal with this MD issue is using the conditional distribution because it contains all information about the history of measurements and about marginal distributions. Let $H_{(k-1)} = (X_1, X_2, \ldots, X_{(k-1)})$ be the history data which has no missing values and assume that there is a missing value X_k at time point k. The conditional density function of X_k given the history is

$$f_{X_k|H_{(k-1)}}\left(x_k \,|x_1, x_2, \ldots, x_{k-1}\right) = \frac{f_{H_{(k-1)},X_k}\left(x_1, x_2, \ldots, x_{k-1}, x_k\right)}{f_{H_{(k-1)}}\left(x_1, x_2, \ldots, x_{k-1}\right)} \tag{7}$$

where $f_{H(k-1),Xk}$ is the joint density function of the data and $f_{H(k-1)}$ is the density function of history. The MD are imputed from generated values from (7) as their mean or mode (see *Käärik et al.* 2009, for more details). However, it is very difficult to obtain the joint and conditional functions. Therefore, the use of copulas is necessary where the joint distribution can be decomposed into a copula to deal with the dependence structure as well as marginal distributions for each variable.

Käärik et al. (2009) handled the missing values with the Gaussian copula which is part of the elliptical family of copulas. The use of this type of copula is justified by its common usage for simple dependence structure, its simplicity and its analytical expression. The Gaussian copula for missing data imputation has been also considered, for instance, by Bárdossy and Pegram (2014) to the infilling of precipitation records. They have found that, for the daily data, the copula-based imputation method is clearly unbiased and superior to other methods (e.g. linear regression, multilinear regression) in terms of point estimation based on the mean absolute error and the root mean squared error.

Alternative models to the Gaussian copula, such as the Archimedean family, have been examined in Di Lascio et al. (2015). It was shown that the copula-based method can be successfully applied on multivariate missing data, independently if the missing pattern is monotone or non-monotone and where the data is characterized by a complex dependence structure. It also outperforms other methods (e.g. EM algorithm).

5 Applications

In this section, the previously developed imputation methods are applied to a case study of three stations dealing with a number of hydrological variables. It is given for illustrative purposes in order to emphasize the MD aspects.

In this application, the main flood characteristics are considered, i.e. Q, V and D (duration) on three stations characterized by their natural regime. These stations are located in the Cote Nord region of the province of Quebec, Canada. The first station, namely *Moisie* station (reference number 072301), is located on the Moisie River at 1.5 km upstream of the QNSLR bridge with a drainage area of 19 012 km^2. Data series of Q, V and D are available from 1979 to 2004 with missing values in 1999 and 2000. The *Magpie* station (reference number 073503) is the second station and is located at the outlet of Magpie Lake. The drainage basin of the *Magpie* station has an area of 7 201 km^2 and complete data are available from 1979 to 2004. The third station is the *Romaine* station (reference number 073801) located at 16.4 km from the Chemin-de-fer bridge on the Romaine River, with a drainage area of 12 922 km^2. The Q, V and D series are available from

1979 to 2004 with no MD. Figure 2 and Table 4 present respectively the location and general information about the considered stations. The correlations between *Q*, *V* and *D* for each station are presented in Table 5.

For each station, the univariate imputation methods, i.e. MS, LI and SRT are applied to each series of *Q*, *V* and *D*. For the multivariate imputation methods, the considered series are (Q,V), (V,D) and (Q,D). The performance of the imputation methods is evaluated using the two stations with no MD, i.e. *Magpie* and *Romaine* stations, and the imputation methods are applied to estimate MD in *Moisie* station. The latter is considered not to evaluate the methods but rather as a real situation with MD. Note that even though it is mentioned above that the situation (*V* known and *Q* unknown) is not realistic, here these variables are treated as if they are generic. This is done to ensure that we represent the general situation.

To evaluate the performance of an imputation method, it is assumed that only one value can be missing in each series. For the copula-based method, in order to determine the best copula to fit the data, we considered the appropriate goodness-of fit test and the AIC criterion (see Chebana 2013). To keep the focus on MD, the corresponding detailed results are not presented. Briefly, the best copulas for (Q,V), (Q,D) and (V,D) are respectively Gumbel, Clayton and Gumbel.

The obtained RRMSE and MRB values of imputation methods are given in Table 6. The table indicates that the MRB is generally positive and relatively low. Hence, the imputation methods can be seen as slightly overestimating the MD. In terms of the RRMSE, the methods can be gathered in three categories. The best results are obtained with the copula-based method where it ranges from 6 to 11% for one copula and from 6 to 14% for the other one. This performance can be attributed to the fact this methodology employs all the available information through the conditional joint distribution (copula and margins). The second category is composed of the three other multivariate methods (SOM, REGEM and MI) as well as the univariate SRT approach where the RRMSE is ranging from 13 to 18% except for SOM which reaches 28%. The exception to find the univariate approach SRT in this category could be related to the recursive error minimizing process of the approach. The SOM as a multivariate approach is not performing well especially for high dependencies (from 18 to 28%). A reasonable explanation could be that the SOM, as other ANN method, is not performing well in extrapolating or when the missing data are out of the range of those employed in the training phase. The last category is composed of the remaining univariate approaches (LI and MS) where the RRMSE is very high (from 22 to 46%) compared to the other two categories without no crossing (except with SOM). Their lower RRMSEs correspond to low dependencies (also higher RRMSEs for higher dependencies). This is not intuitive since we expect better imputation when variables are more dependent. This show the lack of the univariate methods to take into account the dependence information. Note that for the copula, REGEM, MI and SRT, it is not possible to discriminate high and low dependencies mainly because of the short RRMSE variation range. Figure 3 summarizes the above obtained results.

Given the small sample size usually encountered in HFA applications (here *n* = 26), the selected copula may have an impact on the obtained results. In order to check this point, four different copulas are considered (all commonly employed in HFA) and the corresponding performance measures (RRMSE and MRB) are evaluated. The obtained results in Table 7 show that the performance measures are almost in the same range as those of the selected copulas in Table 6. Hence, in the present study, the choice of the copula seems to have little impact on the performance of the copula-based imputation method. However, this result is not general and requires further developments.

As an illustration, the missing data for the years 1999 and 2000 from the Moisie station data are imputed by the different methods. They are presented in Figure 4. Even though it is not possible to check which method is providing the right MD, the obtained values seem to be in accordance with the previous leave-one-out results. Indeed, all methods fellow almost the same pattern, except for the two univariate methods (MS and LI) which had the worst performances.

6 Conclusions

The main objectives of this study are to show the importance of MD imputation in multivariate (multi-variable and multi-site) hydrological series, to compare univariate and multivariate imputation methods and to present imputation methods that can be considered in multivariate HFA. Imputation methods reduce the loss of information which may lead to suboptimal results and hence to inappropriate decisions regarding, for instance, risk estimation of extreme event.

A number of univariate and multivariate imputation methods are presented and applied to the multivariate HFA context. These methods are generally used in time domain analysis. The application of these methods on flood variables of Quebec datasets indicates that overall the multivariate approaches are generally performing better than the univariate ones and the copula-based approach presented the clearly best performance especially in terms of the RRMSE.

Note that in the present study we focused on the bivariate case which is the most considered case in multivariate HFA. However, more variables could be employed to characterize hydrological events. In this case, it is expected that the use of high dimensional data can improve the performance of multivariate imputation methods since more information would be available and used for the imputation.

Acknowledgments

This study was sponsored by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada Research Chair Program. The authors are grateful to the reviewers and the Editor for their valuable comments and suggestions. The authors would like to tank Alexandre Angers for his assistance.

References

Abebe, A. J., D. P. Solomatine and R. G. W. Venneker (2000). "Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events." *Hydrological Sciences Journal* 45(3): 425-436.

Abudu, S., A. S. Bawazir and J. P. King (2010). "Infilling missing daily evapotranspiration data using neural networks." *Journal of Irrigation and Drainage Engineering* 136(5): 317-325.

Adebayo, A. J. and R. Rustum (2012). "Self-organising map rainfall-runoff multivariate modelling for runoff reconstruction in inadequately gauged basins." *Hydrology Research* 43(5): 603-617.

Adeloye, A. J., R. Rustum and I. D. Kariyama (2011). "Kohonen self-organizing map estimator for the reference crop evapotranspiration." *Water Resources Research* 47(8): W08523.

ASCE (1996). *Hydrology Handbook*. New York, American Society of Civil Engineers. 784 Pages

Azen, S. P., M. van Guilder and M. A. Hill (1989). "Estimation of parameters and missing values under a regression model with non-normally distributed and non-randomly incomplete data." *Statistics in Medicine* 8(2): 217-228.

Back, B., K. Sere and H. Vanharanta (1998). "Managing complexity in large data bases using self-organizing maps." *Accounting, Management and Information Technologies* 8(4): 191-210.

Bárdossy A. and G. Pegram (2014), Infilling missing precipitation record - A comparison of a new copula-based method with other techniques, J. Hydrol., 519, 1162-1170.

Beaulieu, C., S. Gharbi, T. B. M. J. Ouarda, C. Charron and M. Aissia (2012). "Improved Model of Deep-Draft Ship Squat in Shallow Waterways Using Stepwise Regression Trees." *Journal of Waterway, Port, Coastal, and Ocean Engineering* 138(2): 115-121.

Bennis, S., F. Berrada and N. Kang (1997). "Improving single-variable and multivariable techniques for estimating missing hydrological data." *Journal of Hydrology* 191(1–4): 87-105.

Berg, D. (2009). Copula Goodness-of-Fit Testing: An Overview and Power Comparison. The European Journal of Finance, 15, 675-701.

Bobée, B. and F. Ashkar (1991). *The gamma family and derived distributions applied in hydrology*, Water Resources Publications. 203 Pages

Chebana, F. (2013). Multivariate Analysis of Hydrological Variables. *Encyclopedia of Environmetrics*, John Wiley & Sons, Ltd. DOI: 10.1002/9780470057339.vnn044

Chebana, F. and T. B. M. J. Ouarda (2008). "Depth and homogeneity in regional flood frequency analysis." *Water Resources Research* 44(11): n/a-n/a.

Chebana, F. and T. B. M. J. Ouarda (2011a). "Multivariate quantiles in hydrological frequency analysis." *Environmetrics* 22(1): 63-78.

Chebana, F. and T. B. M. J. Ouarda (2011b). "Depth-based multivariate descriptive statistics with hydrological applications." *Journal of Geophysical Research: Atmospheres* 116(D10): D10120.

Chebana, F., T. B. M. J. Ouarda and T. C. Duong (2013). "Testing for multivariate trends in hydrologic frequency analysis." *Journal of Hydrology* 486(0): 519-530.

Cherubini U., E. Luciano, W. Vecchiato (2004) Copula methods in finance, John Wiley & Sons.

Chow, G. C. and A.-L. Lin (1976). "Best Linear Unbiased Estimation of Missing Observations in an Economic Time Series." *Journal of the American Statistical Association* 71(355): 719-721.

Coulibaly, P. and N. D. Evora (2007). "Comparison of neural network methods for infilling missing daily weather records." *Journal of Hydrology* 341(1–2): 27-41.

Cunnane, C. and V. Singh (1987). Review of statistical models for flood frequency estimation. *Hydrologic Frequency Analysis*. V. P. Singh, Reidel**:** 49-95.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1): 1-38.

Di Lascio, F. M. L., Giannerini, S., and Reale, A. (2015). Exploring copulas for the imputation of complex dependent data. Statistical Methods & Applications, 24(1), 159-175.

Ding W. and P. X.-K. Song (2016) EM algorithm in Gaussian copula with missing data, Comp. Stat. Data. Anal., 101, 1-11.

Dupuis DJ (2007) Using copulas in hydrology: Benefits, cautions, and issues. Journal of Hydrologic Engineering, 12(4): 381–393.

Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z., & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J Biol Med*, *86*(3), 343-358.

Erol, S. (2011). "Time-Frequency Analyses of Tide-Gauge Sensor Data." *Sensors* 11(4): 3939-3961.

Filippini, F., G. Galliani and L. Pomi (1994). "The estimation of missing meteorological data in a network of automatic stations." *Transactions on Ecology and the Environmental Modelling &amp; Software* 4: 283-291.

Fleig, A. K., L. M. Tallaksen, H. Hisdal and D. M. Hannah (2011). "Regional hydrological drought in north-western Europe: linking a new Regional Drought Area Index with weather types." *Hydrological Processes* 25(7): 1163-1179.

Frane, J. (1976). "Some simple procedures for handling missing data in multivariate analysis." *Psychometrika* 41(3): 409-415.

Frazier, M., Longo, C., & Halpern, B. S. (2016). Mapping Uncertainty Due to Missing Data in the Global Ocean Health Index. *PLoS One*, *11*(8), e0160377.

Gill, M. K., T. Asefa, Y. Kaheil and M. McKee (2007). "Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique." *Water Resour. Res.* 43(7): W07416.

Gleason, T. and R. Staelin (1975). "A proposal for handling missing data." *Psychometrika* 40(2): 229-252.

Gyau-Boakye, P. and G. A. Schultz (1994). "Filling gaps in runoff time series in West Africa." *Hydrological Sciences Journal* 39(6): 621-636.

Hamdi Y., Chebana F., and Ouarda T.B.M.J. (2016). Bivariate Drought Frequency Analysis in The Medjerda River Basin, Tunisia. Journal of Civil & Environmental Engineering, 6: 227. doi:10.4172/2165-784X.1000227.

Han, Y. and N. Li (2010). Interpolation of missing hydrological data based on BP-Neural Networks. *Information Science and Engineering (ICISE)*.

Honaker, J. and G. King (2010). "What to do about missing values in time-series cross-section data." *American Journal of Political Science* 54(2): 561-581.

Hopke, P. K., C. Liu and D. B. Rubin (2001). "Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic." *Biometrics* 57(1): 22-33.

Huang, C. and J. R. G. Townshend (2003). "A stepwise regression tree for nonlinear approximation: Applications to estimating subpixel land cover." *International Journal of Remote Sensing* 24(1): 75-90.

Hughes, D. A. and V. Smakhtin (1996). "Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves." *Hydrological Sciences Journal* 41(6): 851-871.

Jeffrey, S. J., J. O. Carter, K. B. Moodie and A. R. Beswick (2001). "Using spatial interpolation to construct a comprehensive archive of Australian climate data." *Environmental Modelling &amp; Software* 16(4): 309-330.

Käärik E. (2007), Handling dropouts in repeated measurements using copulas, University of Tartu Press, 51.

Käärik E. and M., Käärik (2009), Modeling dropouts by conditional distribution, a copula-based approach, Journal of Statistical Planning and Inference, 139, 3830-3835.

Kalteh, A. M. and P. Hjorth (2009). "Imputation of missing values in a precipitation–runoff process database." *Hydrology Research* 40(4): 420–432.

Kao, S. C. and N. B. Chang (2012). "Copula-based flood frequency analysis at ungauged basin confluences: Nashville, tennessee." *Journal of Hydrologic Engineering* 17(7): 790-799.

Kelly, E., F. Sievers and R. McManus (2004). "Haplotype frequency estimation error analysis in the presence of missing genotype data." *BMC Bioinformatics* 5(1): 188.

Khaliq, M. N., T. B. M. J. Ouarda, J. C. Ondo, P. Gachon and B. Bobée (2006). "Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review." *Journal of Hydrology* 329(3–4): 534-552.

Kim, J.-W. and Y. A. Pachepsky (2010). "Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation." *Journal of Hydrology* 394(3–4): 305-314.

Kite, G. (1988). *Frequency and Risk Analyses in Hydrology*. Colo., USA, Water Resources Publications. 257 Pages

Klebanoff, M. A. and S. R. Cole (2008). "Use of multiple imputation in the epidemiologic literature." *American Journal of Epidemiology* 168(4): 355-357.

Kodituwakku, S., R. A. Kennedy and T. D. Abhayapala (2011). Time-frequency analysis compensating missing data for Atrial Fibrillation ECG assessment. *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*.

Kohonen, T., E. Oja, O. Simula, A. Visa and J. Kangas (1996). "Engineering applications of the self-organizing map." *Proceedings of the IEEE* 84(10): 1358-1384.

Kuligowski, R. J. and A. P. Barros (1998). "Using artificial neural networks to estimate missing rainfall data 1." *JAWRA Journal of the American Water Resources Association* 34(6): 1437-1447.

Lettenmaier, D. P. (1980). "Intervention analysis with missing data." *Water Resour. Res.* 16(1): 159-171.

Linacre, E. (1992). *Climate Data and Resources - A Reference and Guide*. Routledge, London and New York. 384 Pages

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis With Missing Data*. New Jersey, Wiley Interscience Publication. 381 Pages

Liu, C. and D. B. Rubin (1994). "The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence." *Biometrika* 81(4): 633-648.

Liu, C. and D. B. Rubin (1998). "Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data." *Biometrika* 85(3): 673-688.

Makhnin, O. V. and D. L. McAllister (2009). "Stochastic precipitation generation based on a multivariate autoregression model." *Journal of Hydrometeorology* 10(6): 1397-1413.

Marlinda, A. M., M. S. Siti and H. Sobri (2010). "Restoration of Hydrological Data in the Presence of Missing Data via Kohonen Self Organizing Maps." *InTech*: 223-242.

Meng, X.-L. and D. Van Dyk (1997). "The EM Algorithm—an Old Folk-song Sung to a Fast New Tune." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59(3): 511-567.

Meng, X. L. and D. B. Rubin (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework." *Biometrika* 80(2): 267-278.

Miller, A. (2002). *Subset Selection in Regression*, Taylor & Francis. 256 Pages

Mwale, F. D., A. J. Adeloye and R. Rustum (2012). "Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi – A self organizing map approach." *Physics and Chemistry of the Earth, Parts A/B/C* 50–52(0): 34-43.

Ng, W., U. Panu and W. Lennox (2009). "Comparative Studies in Problems of Missing Extreme Daily Streamflow Records." *Journal of Hydrologic Engineering* 14(1): 91-100.

Ouarda, T. B. M. J., M. Hache, P. Bruneau and B. Bobee (2000). "Regional flood peak and volume estimation in northern Canadian basin." *Journal of Cold Regions Engineering* 14(4): 176-191.

Overeem, A., T. A. Buishand and I. Holleman (2009). "Extreme rainfall analysis and estimation of depth-duration-frequency curves using weather radar." *Water Resources Research* 45(10): W10424.

Özçelik, C. and E. Benzeden (2010). "Regionalization approaches for the periodic parameters of monthly flows: a case study of Ceyhan and Seyhan River basins." *Hydrological Processes* 24(22): 3251-3269.

Patrician, P. A. (2002). "Multiple imputation for missing data†‡." *Research in Nursing & Health* 25(1): 76-84.

Peterson, H. M., J. L. Nieber and R. Kanivetsky (2011). "Hydrologic regionalization to assess anthropogenic changes." *Journal of Hydrology* 408(3–4): 212-225.

Raman, H. and N. Sunilkumar (1995). "Multivariate modelling of water resources time series using artificial neural networks." *Hydrological Sciences Journal* 40(2): 145-163.

Ramos-Calzado, P., J. Gómez-Camacho, F. Pérez-Bernal and M. F. Pita-López (2008). "A novel approach to precipitation series completion in climatological datasets: application to Andalusia." *International Journal of Climatology* 28(11): 1525-1534.

Rao, A. R. and K. H. Hamed (2000). *Flood Frequency Analysis*. Boca Raton, CRC Press. 376 Pages

Reiter, J. P. and T. E. Raghunathan (2007). "The Multiple Adaptations of Multiple Imputation." *Journal of the American Statistical Association* 102(480): 1462-1471.

Requena, A.I., Mediero, L., Garrote, L. (2013) A bivariate return period based on copulas for hydrologic dam design: Accounting for reservoir routing in risk estimation. *Hydrology and Earth System Sciences*, 17 (8), pp. 3023-3038

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc. 358 Pages

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London, Chapman & Hall. 448 Pages

Schafer, J. L. (1999). "NORM: Multiple Imputation of Incomplete Multivariate Data under a Normal Model, version 2. Software for Windows 95/98/NT, available at: http://www.stat.psu.edu/jls/misoftwa.html."

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research, 33*(4), 545-571.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2), 147

Schneider, T. (2001). "Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values." *Journal of Climate* 14(5): 853-871.

Shu, C., and T.B.M.J. Ouarda (2012). Improved methods for daily streamflow estimates at ungauged sites, Water Resources Research, 48, W02523, doi:10.1029/2011WR011501.

Simonovic, S. P. (1995). "Synthesizing missing streamflow records on several Manitoba streams using multiple nonlinear standardized correlation analysis." *Hydrological Sciences Journal* 40(2): 183-203.

Srebotnjak, T., G. Carr, A. de Sherbinin and C. Rickwood (2012). "A global Water Quality Index and hot-deck imputation of missing data." *Ecological Indicators* 17(0): 108-119.

Sterne, J. A., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood and J. R. Carpenter (2009). "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." *BMJ (Clinical research ed.)* 338.

Teegavarapu, R. S. V. and V. Chandramouli (2005). "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records." *Journal of Hydrology* 312(1–4): 191-206.

Vogel, R. M., and N. M. Fennessey (1994), Flow-duration curves. I : New interpretation and confidence intervals, J. Water Resour. Plann. Manage., 120(4), 485–504.

Wang, C., N.-B. Chang and G.-T. Yeh (2009). "Copula-based flood frequency (COFF) analysis at the confluences of river systems." *Hydrological Processes* 23(10): 1471-1486.

Westra, S., R. Mehrotra, A. Sharma and R. Srikanthan (2012). "Continuous rainfall simulation: 1. A regionalized subdaily disaggregation approach." *Water Resources Research* 48(1): W01535.

Yan, J., Kojadinovic, I. (2010). Modeling Multivariate Distributions with Continuous Margins Using the copula R Package. Journal of Statistical Software, 34, 1-20

Yucel, R. M. and A. M. Zaslavsky (2005). "Imputation of Binary Treatment Variables With Measurement Error in Administrative Data." *Journal of the American Statistical Association* 100(472): 1123-1132.

Yue, S., P. Pilon and G. Cavadias (2002). "Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series." *Journal of Hydrology* 259(1–4): 254-271.

Zhang, L. and V. Singh (2006). "Bivariate Flood Frequency Analysis Using the Copula Method." *Journal of Hydrologic Engineering* 11(2): 150-164.

Zhu, B., C. He and P. Liatsis (2012). "A robust missing value imputation method for noisy data." *Applied Intelligence* 36(1): 61-74.

Tables

Table 1: Main HFA steps in the univariate and multivariate frameworks with some references

Main HFA steps	Framework	
	Univariate	Multivariate
(i) Exploratory analysis: - Outlier detection - Missing data imputation - Descriptive analysis	For instance: Cuanne and Singh (1987) Rao and Hamed (2000) Kite (1988)	For instance: Chebana and Ouarda (2011) for outlier detection in descriptive analysis **The specific aim of the present paper: missing data imputation in the multivariate setting**
(ii) Checking the HFA assumptions	For instance: Yue and al. (2002) Khaliq et al. (2006)	For instance: Chebana et al. (2013) for testing multivariate trends in HFA
(iii) Modeling and estimation	For instance: Cuanne and Singh (1987) Bobée and Ashkar (1991)	For instance: Shiau (2006) Zhang and Singh (2006)
(iv) Risk evaluation and analysis	For instance: Rao and Hamed (2000)	For instance: Shiau (2003) Chebana and Ouarda (2011)

Table 2: Summary of missing data frameworks with some references

Framework		Fields	
		Statistic	Hydrology
Univariate	Time series analysis	Large body of literature: Gelason and Staelin (1975) Chow and Lin (1976) Azen et al. (1989)	Large body of literature: Lettenmayer (1980) Jefferey et al. (2001) Teegavarapu and Chandramouli (2005)
	Frequency analysis	Large body of literature: Erol (2011) Kodituwakku et al. (2011)	Sparce body of literature: Fleig et al. (2011) Peterson et al. (2011)
Multivariate	Time series analysis	Large body of literature: Frane (1976) Hopke et al. (2001) Honaker et al. (2010)	Large body of literature: Ng et al. (2009) Kalteh et Hjorth (2009)
	Frequency analysis	Sparce body of literature: Kelly et al. (2004)	**The specific aim of the present paper**

Table 3: Overview of imputation methods in MD context

Techniques	Description	When to be used	Avantages	Disadvantages	References (e.g.)
Univariate setting					
Mean substitution	Missing data are replaced by the mean	Less than 10% of data are missing	Easy to use	Underestimates the variance and the degree of freedom	Linacre (1992)
Subgroup mean substitution	Missing data are replaced by the mean of a subgroup	When it is easy to define subgroups	Gives better estimates when compared to mean substitution	Underestimates the variance, subgroups are defined arbitrarily	Linacre (1992)
Time series analysis	Determines the model and the corresponding parameters and then estimates missing data	High autocorrelation	Takes into consideration the temporal variability in the data	The necessity to define, a priori, the functional form of the relationships	Lettenmaier (1980)
Interpolation	Interpolate two points of data, one immediately before the gap and the other soon alter the gap and interpolating the missing data	Only suitable in stable periods and short length of the gap	Gives better estimates of statistical inference when well used	Limited to special cases that rarely occurs	Filippini et al. (1994)
Regression	Estimate parameters of the regression and use them to estimate tnk«tng data	Data sets exhibiting significant temporal patterns	Estimated data preserves deviation from the mean and the shape of the available	Could distort the number of degrees of freedom. Difficult to use in noisy data sets	Kuligowski and Barros (1998)
Hot-deck imputation	Replace missing data with value from a similar case	Data are missing in certain patterns	Missing values are replaced by real values	Problematic if no other case closely related to the missing value	Srebotnjaketal (2012)
Inverse distance weighting	Define the neighborhood and the weighting parameters. Then estimate missing data by spatial interpolation using weighting	Stations are highly correlated	Gives better estimates of statistical inference when well used	Problematic with the existence of negative autocorrelation	ASCE(1996)
Multivariate setting					
k-nearest neighbor	Estimate the missing data based on the closest training examples in the feature space	When the feature space does not require the selection of a predetermined model	Flexible and missing values are replaced with real values	Low accuracy rate in multidimensional data and computation cost is quite high	Kaltehand Hjorth (2009)
Artificial neural networks	Determine the architecture of the ANN, estimate parameters and estimate missing data	When assumptions about the missing data mechanism cannot be made and in case of nonlinear relationships between variables	Ability to model complex patterns without a prior knowledge of the underlying process	Numerous parameters to estimate and gives unrealistic results when such noise is available in the data	Raman and Sunilkumar (1995)
Expectation minimisation	Estimate model parameters by iterative process that continuous until convergence	When distribution assumptions are realistic	Increased accuracy if model is correct	Strict assumptions, complex algorithm and takes time to converge	Ng et al. (2009)
Multiple	Specify and appropriate	When assumptions are	The variability of the	Strict assumptions and takes	Ng et al. (2009)

imputation	imputation model, estimate more than one imputed value for each of the missing data	realistic	imputed values can be considered	time to converge	
Copula-based	Determine the conditional distribution given the historical data. Generate the missing values from the obtained distribution	Both variables are continuous. Complex pattern of dependance	Flexible. Superior performance with highlighting the dependence	Requires fitting of a copula. Not developed for all copulas	Käärik et Käärik (2009)

Table 4: General characteristics of *Moisie*, *Magpie* and *Romaine* stations

Station name	Station number	Latitude	Longitude	Period of records (#years)	Missing data	Area (Km^2)	Mean streamflow (m^{3}s^{-1})
Moisie	072301	50 21 09	-66 11 12	1979-2004 (26)	1999, 2000	19 012	391.62
Magpie	073503	50 41 08	-64 34 43	1979-2004 (26)	-	7 201	163.56
Romaine	073801	50 18 28	-63 37 07	1979-2004 (26)	-	12 922	282.89

Table 5: Correlations between *Q*, *V* and *D*

Stations		Variables	
	Q	**V**	**D**
Moisie			
Q	1	0.59	-0.07
V		1	0.66
D			1
Magpie			
Q	1	0.70	-0.20
V		1	0.44
D			1
Romaine			
Q	1	0.77	-0.36
V		1	0.18
D			1

Table 6: RRMSE and MRB of univariate and multivariate imputation methods

Method		RRMSE		MRB	
		Magpie	Romaine	Magpie	Romaine
			Q		
Univariate	**MS**	41.09	46.26	9.73	11.93
	LI	32.33	30.73	5.65	4.63
	SRT	16.65	17.68	2.94	2.12
Multivariate	**SOM**	28.24	27.40	5.13	4.18
	REGEM	16.85	17.77	3.05	2.27
	MI	16.14	17.31	0.83	-4.08
	Copula (Q,V)	8.93	8.34	4.26	14.36
	Copula (Q,D)	12.62	8.54	9.28	15.18
			V		
Univariate	**MS**	42.54	39.25	12.81	11.49
	LI	37.64	31.08	8.94	6.52
	SRT	17.31	15.26	1.63	2.10
Multivariate	**SOM**	19.37	17.88	-1.07	2.69
	REGEM	17.33	15.31	1.71	2.16
	MI	16.98	14.69	-0.17	-2.20
	Copula (Q,V)	10.67	7.91	12.88	4.38
	Copula (V,D)	13.59	12.43	34.62	25.52
			D		
Univariate	**MS**	28.21	22.59	5.70	6.41
	LI	28.67	22.58	5.48	5.01
	SRT	17.54	13.81	1.94	3.31
Multivariate	**SOM**	15.04	17.52	-1.58	5.05
	REGEM	17.75	13.77	2.07	3.45
	MI	17.21	13.38	-0.12	-2.50
	Copula (Q,D)	6.06	7.71	-0.71	16.12
	Copula (V,D)	6.36	9.28	0.36	24.51

Gray color indicates the methods with smallest RRMSE or MRB for each variable.

Table 7: Comparison of RRMSE and MRB for the Magpie and Romaine stations using Gaussian, Clayton, Gumbel and Frank copulas

Magpie

	Q	
Copula	**RRMSE**	**MRB**
Gaussian	6.11	3.98
Clayton	12.62	9.28
Gumbel	8.93	4.26
Frank	8.19	10.16
	V	
	RRMSE	**MRB**
Gaussian	15.89	33.84
Clayton	9.86	15.40
Gumbel	10.67	12.88
Frank	10.19	30.10
	D	
	RRMSE	**MRB**
Gaussian	6.20	6.51
Clayton	6.06	-0.71
Gumbel	6.36	0.36
Frank	6.28	3.96

Romaine

	Q	
Copula	**RRMSE**	**MRB**
Gaussian	8.89	11.69
Clayton	8.54	15.18
Gumbel	8.34	14.36
Frank	9.69	15.05
	V	
	RRMSE	**MRB**
Gaussian	9.31	4.44
Clayton	9.52	12.77
Gumbel	7.91	4.38
Frank	12.60	14.66
	D	
	RRMSE	**MRB**
Gaussian	10.33	26.78
Clayton	7.71	16.12
Gumbel	9.28	24.51
Frank	8.24	17.13

Figures

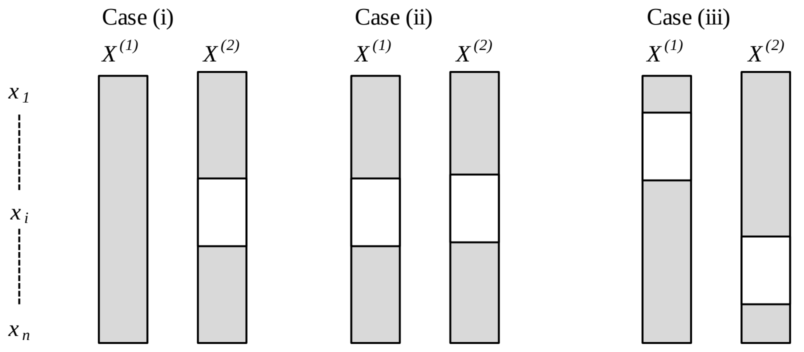

Figure 1: Examples of missing data patterns. Gray color indicates available observed data.

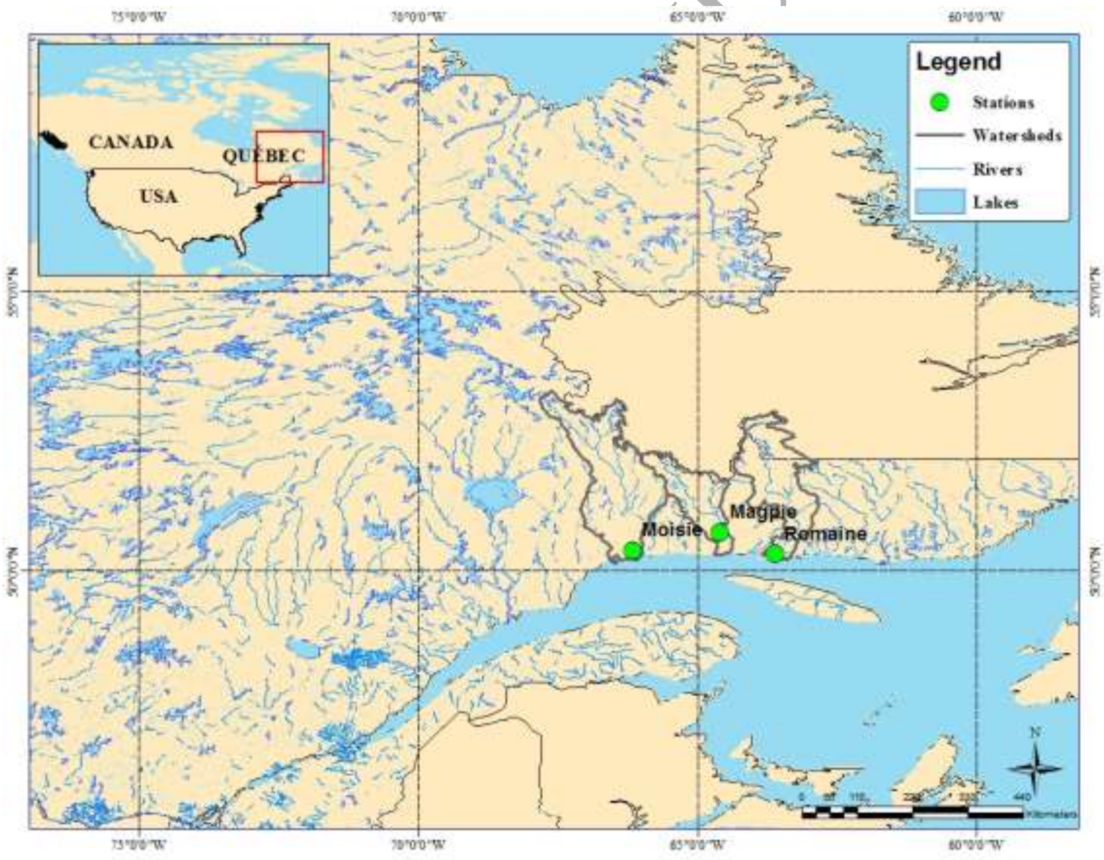

Figure 2: Geographical locations of *Moisie, Magpie* and *Romaine* stations.

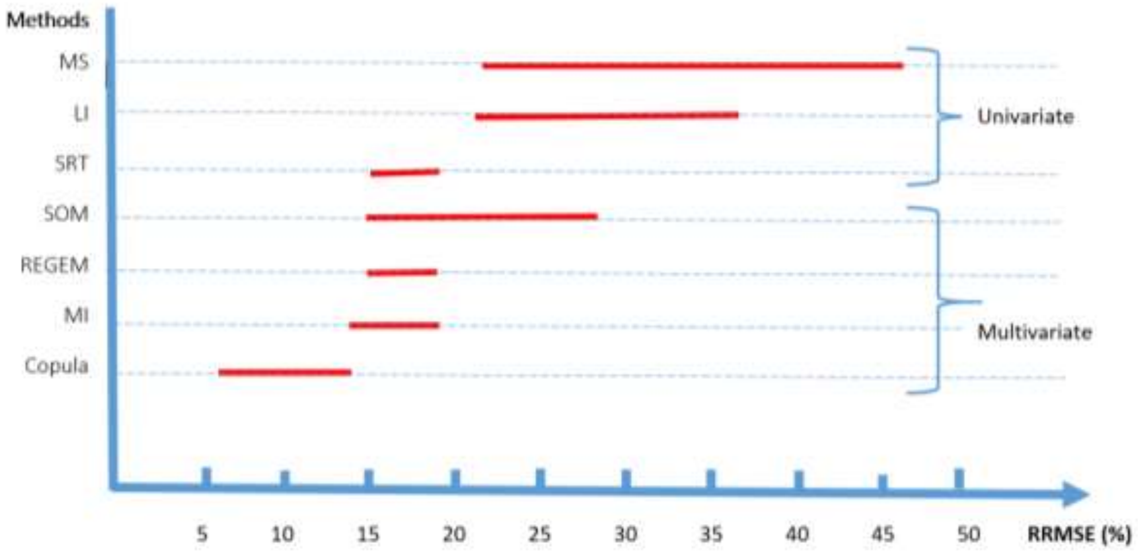

Figure 3: Summary of the RRMSE results of the considered methods.

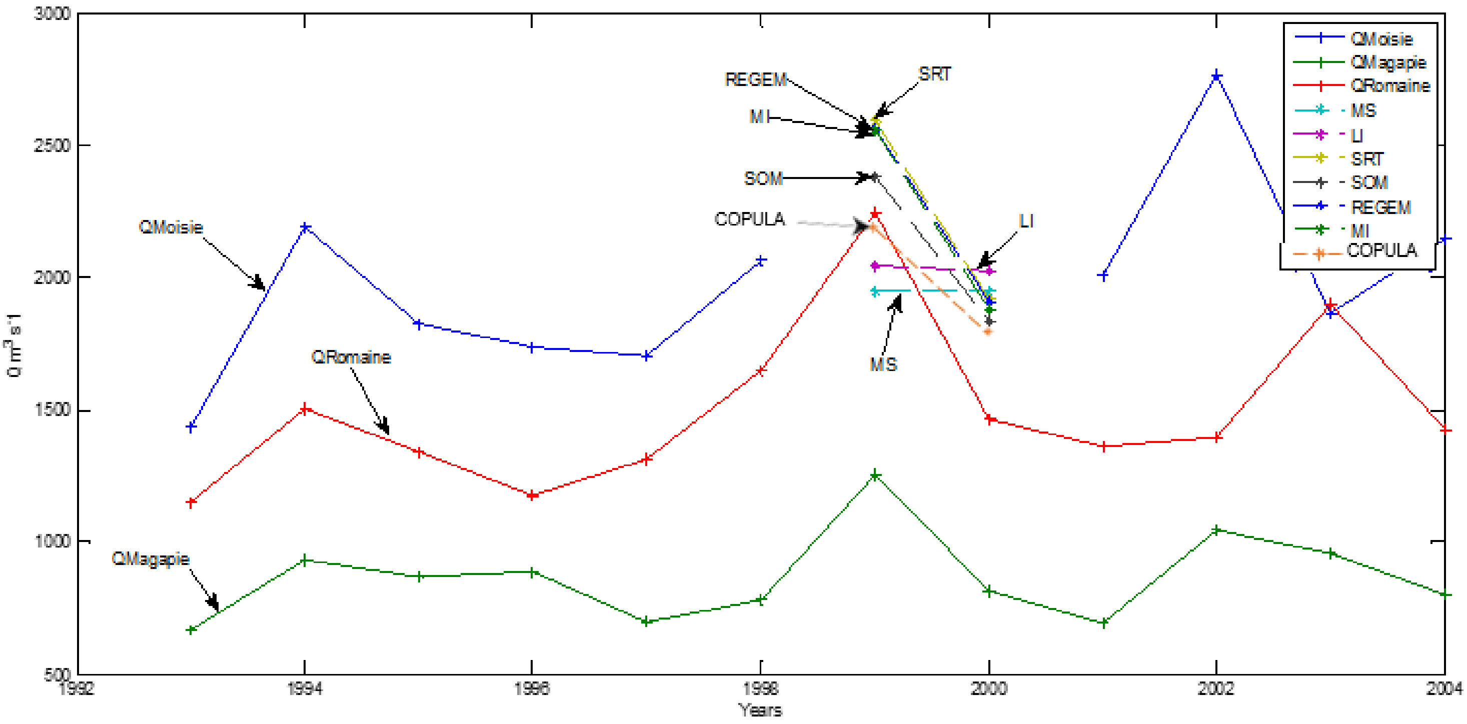

Figure 4: The Q series of the three studied stations and the estimation of MD in Moisie station.