Imputation of missing values in a precipitation–runoff process database

Aman Mohammad Kalteh and Peder Hjorth

ABSTRACT

Hydrologists are often faced with the problem of missing values in a precipitation–runoff process database to construct runoff prediction models. They tend to use simple and naive methods to deal with the problem of missing data. Thus far, the common practice has been to discard observations with missing values. In this paper, we present some statistically principled methods for gap filling and discuss the pros and cons of these methods. We employ and discuss imputations of missing values by means of self-organizing map (SOM), multilayer perceptron (MLP), multivariate nearest-neighbor (MNN), regularized expectation–maximization algorithm (REGEM) and multiple imputation (MI) in the context of a precipitation–runoff process database in northern Iran in order to construct a serially complete database for analyses such as runoff prediction. In our case, the SOM and MNN tend to give similar and robust results. REGEM and MI build on the assumption of multivariate normal data, which we don't seem to have in one of our cases. MLP tends to produce inferior results because it fragments the data into 68 different models. Therefore, we conclude that it makes most sense to use either the computationally simple MNN method or the more demanding SOM.

Key words | data fill in, imputation methods: SOM, MLP, MNN, REGEM, MI, missing values, serially complete data

Aman Mohammad Kalteh (corresponding author)
Peder Hjorth
Department of Water Resources Engineering,
LTH, Lund University,
PO Box 118,
Lund S-22100,
Sweden
Tel.: +46 46 222 8981
Fax: +46 46 222 4435
E-mail: *aman_mohammad.kalteh@tvrl.lth.se*

Aman Mohammad Kalteh (corresponding author)
Department of Range and Watershed Management,
Faculty of Natural Resources,
University of Guilan,
PO Box 1144,
Sowmehe Sara,
Guilan,
Iran
E-mail: *kalteh@guilan.ac.ir*

INTRODUCTION

In general, the problem of missing values is a common obstacle in time series analysis and specifically in the context of precipitation–runoff process modelling where it is essential to have serially complete data.

There may be various reasons for missing values, for instance equipment failure, errors in measurements or faults in data acquisition, and natural hazards such as landslides. Whatever the reasons, missing values produce a significant problem for water resources applications, which generally require a continuous database (e.g. Zoppou *et al.* 2000; Junninen *et al.* 2004; Ramirez *et al.* 2005). Consequently, finding efficient and principled methods to deal with the problem of missing values is an important issue in most hydrological analyses. However, hydrological modellers commonly discard the observations with missing values and only use the observations with complete information, which means that a lot of information contained in the dataset is lost. Furthermore, the method is inadequate for analyses that require serially complete data. As an alternative to this listwise deletion procedure ("or complete-case analysis"), modellers sometimes impute ("or fill in") a value for the missing values by using, for example, the mean of the observed variables. Such a procedure will, however, seriously distort statistical properties like standard variation, correlations or percentiles.

During the last 20 years, statisticians have tried to introduce model-based methods such as maximum likelihood using the expectation–maximization (EM) algorithm and multiple imputation (MI), arguing that they provide sounder and more promising solutions for a wider range of situations. However, these methods carry assumptions about the missing data mechanisms. According to Little &

doi: 10.2166/nh.2009.001

Rubin (2002) missing data mechanisms can be divided into three classes that consist of (1) missing completely at random (MCAR), (2) missing at random (MAR) and (3) not missing at random (NMAR). MCAR occurs when the probability of having a missing value for a component does not depend on either the available values or the missing value itself. In other words, complete observations are random samples of the full dataset. Complete-case analysis requires this assumption. MAR occurs when the probability of having a missing value for a component may depend on the available values, but not on the missing value itself. MCAR and MAR are also called ignorable response mechanisms because the reasons for missing data can be ignored during the analysis. Model-based methods require this assumption, and it is most reasonable to assume that missing hydrological data are MCAR or MAR. As one cannot obtain explicit empirical information about the missing data mechanism, one can only resort to the strategy of sensitivity analysis to assess the effect of the MCAR and MAR assumptions (Pigott 2001; Little & Rubin 2002). And finally NMAR occurs when the probability of having a missing value for a component could depend on the value of the missing value itself. NMAR is also termed nonignorable.

According to Little & Rubin (2002), different methods for dealing with missing values are available and they can be divided into three categories.

Listwise/pairwise deletion

In the listwise deletion strategy, which is also known as complete-case analysis, all cases (observations) with one or more missing values are discarded from the analysis. This strategy is easy to carry out and is the most commonly practiced one. In most statistical software packages it is the default option. This strategy may be satisfactory when only a small fraction of the database is missing and the missing data mechanism is MCAR (Little & Rubin 2002). With a larger fraction missing, the analysis would be less precise since a smaller number of observations would be available for analysis. In our case, only 45% of the data would be available for analysis. Pairwise deletion, which is also known as available-case analysis, uses different sets of sample observations for each statistic. This strategy preserves more information compared to listwise deletion. However, the reliability of its estimates depends not only on the MCAR assumption but also on existing correlations among the variables in the database (Pigott 2001; Little & Rubin 2002; Tsikriktsis 2005).

Imputation-based procedures

Here missing values are imputed with plausible values rather than being deleted entirely. Imputation has several advantages such as efficiency and precision because no observations are discarded. However, it suffers from implementation difficulties, especially in a multivariate database. Moreover, some techniques can falsify data relationships and distributions (Schafer & Graham 2002). Many different missing data imputation procedures have been applied, and among them the mean, regression and hot deck methods are the most commonly used ones. In the mean imputation procedure, the mean of the available values is used to fill in the missing data. The regression imputation procedure can be summarized as a two-step approach (Frane 1976): first, the relationships among variables are estimated and second the regression coefficients are used to estimate missing values. It requires MAR data. Finally, the hot deck imputation procedure uses values from similar observations from a selected dataset to impute.

Model-based procedures

Two model-based procedures briefly illustrated in this subsection are maximum likelihood using the expectation–maximization (EM) algorithm and multiple imputation (MI). Maximum likelihood procedures are used to estimate the parameters of a model defined for the complete data. A general method for maximum likelihood in missing data problems was described by Dempster *et al.* (1977) in their seminal paper on the EM algorithm. Maximum likelihood procedures for missing multivariate normal data can be parameterized based on the mean vector and covariance matrix. In every iteration of the EM algorithm, estimates of the mean and covariance matrix are adjusted in three stages. First, regression coefficients for the variables with missing values are computed based on the variables with available values. Second, missing values are filled in with their conditional expectation values, which is the product of the

available values and the estimated model parameters such as regression coefficients. Third, the parameters of the mean and covariance matrix are re-estimated. Then, the EM algorithm process cycles back and forth until the imputed values and the estimates of the mean and covariance matrix parameters do not change substantially.

In MI the analyst specifies an appropriate imputation model, imputes several complete databases (usually 3–5 times) and performs the desired statistical analysis on each database separately by using standard complete-data methods, and thereafter combines the results (Allison 1998; Patrician 2002). Both maximum likelihood using the EM algorithm and MI require the assumptions of multivariate normality and MAR.

As mentioned previously, both of these model-based procedures provide statistically principled solutions for a wider range of circumstances (e.g. Pigott 2001; Little & Rubin 2002; Schafer & Graham 2002; Tsikriktsis 2005).

Various methods have been applied for imputation of missing values in different disciplines: Fessant & Midenet (2002) used a self-organizing map (SOM) for imputation of survey data along with the multilayer perceptron (MLP) and hot deck methods. They concluded that the results obtained by SOM were encouraging. Musil *et al.* (2002) compared listwise deletion, mean substitution, simple regression, regression with an error term and the EM algorithm in imputing of artificially generated missing values on a single variable in nursing research. They found that regression with an error term and the EM algorithm results were encouraging while mean substitution was the least accurate one. Junninen *et al.* (2004) tested univariate (linear, spline and nearest-neighbor interpolation), multivariate (regularized expectation–maximization (REGEM) algorithm, nearest-neighbor (NN), self-organizing map (SOM), multilayer perceptron (MLP)) as well as hybrid methods (i.e. combining the best features of univariate and multivariate methods) in air quality datasets. They found that univariate methods are dependent on the gap length as well as on the variable under study. Among the multivariate methods both SOM and MLP showed slightly better imputation ability than the others. However, they also showed a slight improvement by means of hybridization. Moreover, the authors found a substantial improvement by using MI schemes.

Numerous applications of artificial neural networks in various contexts of hydrology have been documented in ASCE (2000*b*) and many studies have shown their applicability in hydrology. Kuligowski & Barros (1998) used a back-propagation neural network to estimate missing rainfall data by using concurrent data from nearby gauges. They found that a neural network approach performs better than the traditional methods of arithmetic, distance-weighted average and linear regression procedures. Abebe *et al.* (2000) used a fuzzy-rule-based model for filling in missing precipitation data using data from adjacent stations. They compared the results obtained using the fuzzy-rule-based model with those obtained using an ANN model and a traditional statistical model. The fuzzy-rule-based model was found to be slightly better. Khalil *et al.* (2001) proposed seasonal grouping of data for developing ANN models to estimate missing values in monthly runoff databases. Bhattacharya *et al.* (2003) used ANN models to fill in the missing values of wave data. More recently, Kalteh & Berndtsson (2007) showed the overall advantage of SOM over MLP in interpolating monthly precipitation in northern Iran. According to our review, there have been no investigations concerning imputation of missing values in a multivariate precipitation–runoff context and, according to our knowledge, there have, as yet, been no studies concerning the imputation ability of MI in a hydrological database.

In this study, self-organizing map (SOM), multilayer perceptron (MLP) artificial neural network models, multivariate nearest-neighbor (MNN), regularized expectation–maximization (REGEM) algorithms and multiple imputation (MI) were examined to impute the missing values in a precipitation–runoff process database where missingness may occur on any of the variables and the missing data mechanism can reasonably be assumed to be missing at random (MAR).

MATERIALS AND METHODS

Data

Our study is based on daily precipitation and runoff data from a watershed located in northern Iran. The time series span from 1969–70 to 1999–2000 and consists of five

precipitation stations, i.e. 13001, 13004, 13005, 13007 and 13013, and two runoff stations, i.e. 13005 and 13013. This database was published by the Ministry of Energy of Iran. Some details about the stations including longitude, latitude and elevation are shown in Table 1. As both precipitation and runoff have been recorded in some stations, we have used P and R in parenthesis throughout the paper to indicate precipitation or runoff data, respectively. In addition to the variables above, i.e. precipitation and runoff, two extra inputs (i.e. sine and cosine curves) were used in order to represent the annual cycle, a unique signature for each day of the year, so that the conditions of the first day follow those of the last day. To summarize, the utilized database contains nine variables including precipitation, runoff and time curves. As seen from Table 1, the maximum observed value at 13013(R) is very high compared to other variables. Prior to the analysis the data were standardized so that all data were in the range from 0 to 1, except in REGEM in which the variables were standardized to zero mean and unit variance. The amount of missing values varies among variables, as shown in Table 2 in which we have summarized the length of the missing data in each variable along with existing missing data patterns in the utilized database. The column totals of Table 2 provide the number of observations (patterns) missing for each variable along with its percentage while the row totals provide the frequencies of each missing data pattern as well as its percentage. For example, the first row total in Table 2 shows that 5,121 (45.21%) of all patterns contain all variables. In other words, over half of the observations have missing values for one or more variables. As seen from the table, our database contains 68 missing data patterns. The SOM, MLP, MNN, REGEM and MI methods were applied to impute missing values and the experimental design for each method is illustrated in the respective subsection below.

Table 1 | Description of the stations

Stations	Latitude (m)	Longitude (m)	Elevation (m)	Min	Max
13001(P)	1,311,576.1	4,092,853	1,392	0	81
13004(P)	1,275,860.5	4,087,657.2	1,038	0	96
13005(P)	1,297,143	4,085,876.5	1,092	0	110
13005(R)	1,297,143	4,085,876.5	1,092	0	44.5
13007(P)	1,371,789	4,095,050.7	3,376	0	79
13013(P)	1,244,301.2	4,086,715.7	134	0	92
13013(R)	1,244,301.2	4,086,715.7	134	0	1523.80

Precipitation and runoff units are mm and m^3/s, respectively.

Self-organizing map (SOM)

The self-organizing map or SOM (Kohonen 1982*a*,*b*, 2001) is a type of artificial neural network (ANN) designed for unsupervised pattern recognition applications. The SOM learns to map, in a nonlinear fashion, from a high-dimensional input layer to a low-dimensional, mostly two-dimensional, discrete lattice of neurons (units) in the output layer. Figure 1 shows a typical structure of a two-dimensional SOM consisting of an input layer and a Kohonen or output layer. The neurons in the output layer are connected to all inputs via weight vectors. There are two strategies for training of an SOM: a batch and a sequential approach, respectively. In this study a batch learning algorithm was used for constructing the SOM, because the batch SOM is faster and eliminates the need for an *a priori* specification of the learning rate, which eliminates convergence problems (Kohonen 2001). The batch training of the SOM is similar to a sequential training algorithm, which is the most commonly used, in the sense that both are iterative. In sequential training, the weights are updated on a vector-by-vector basis while in the batch algorithm the whole database is presented to the map before any updates are made. Thus, in each training iteration, the batch algorithm lists input vectors one by one under the best matching units and updates the neurons according to the whole database. The procedure can be summarized as follows. At the outset of the training, weight vectors w_j must be initialized to each neuron and, thanks to the unsupervised or self-organization procedure, the input vectors (x_i) are compared with the SOM neurons to find the closest matches which are called best matching units (BMUs). The most commonly used criterion for comparison is the Euclidean distance. Then, new weight vectors are calculated using the following rule:

$$w_j(t+1) = \frac{\sum_{i=1}^{n} N_{j*}(t) x_i}{\sum_{i=1}^{n} N_{j*}(t)} \quad (1)$$

where $N_{j*}(t)$ is the neighborhood function (Gaussian in this study) of the best matching neurons $j*$ at iteration t and n

Table 2 | Missing data patterns

13001(P)	13004(P)	13005(P)	13005(R)	13007(P)	13013(P)	13013(R)	#	%
O	O	O	O	O	O	O	5,121	45.21
O	O	O	O	M	O	O	694	6.12
M	O	O	O	O	O	O	642	5.66
O	M	O	O	O	O	M	588	5.19
O	O	O	O	O	O	M	516	4.55
O	O	O	O	O	M	O	420	3.70
M	M	O	O	M	O	O	372	3.28
O	O	O	M	O	O	O	278	2.45
M	O	O	O	M	O	O	263	2.32
O	O	M	O	O	O	O	205	1.81
O	M	O	O	O	O	O	198	1.74
O	O	O	O	M	O	M	192	1.69
M	O	M	O	O	O	O	142	1.25
M	O	O	O	O	O	M	137	1.20
O	O	O	M	O	O	M	108	0.95
M	M	O	O	O	O	O	96	0.84
O	O	O	M	O	M	M	95	0.83
M	O	O	O	O	M	O	84	0.74
M	M	O	O	O	O	M	72	0.63
M	O	M	O	O	M	O	68	0.60
M	M	O	O	M	M	O	65	0.57
M	O	O	O	M	M	O	60	0.52
O	O	O	O	M	M	M	48	0.42
O	O	M	M	O	M	M	48	0.42
O	M	M	O	O	O	M	48	0.42
M	M	M	O	O	O	M	48	0.42
O	O	M	O	O	O	M	46	0.40
O	M	O	M	O	O	O	36	0.31
O	M	O	O	M	O	M	36	0.31
O	O	M	O	O	M	O	35	0.30
O	O	O	O	O	M	M	24	0.21
O	O	O	M	O	M	O	24	0.21
O	M	O	O	M	O	O	24	0.21
O	M	O	O	M	M	M	24	0.21
O	M	M	O	M	M	M	24	0.21
M	O	O	M	O	M	M	24	0.21
M	O	M	O	O	O	M	24	0.21
M	O	M	M	O	M	M	24	0.21
M	O	M	O	M	O	O	24	0.21
O	M	M	O	M	O	M	23	0.20
O	O	O	O	M	M	O	19	0.16
O	M	M	O	O	M	O	18	0.15

(continued)

Table 2 | (*continued*)

13001(P)	13004(P)	13005(P)	13005(R)	13007(P)	13013(P)	13013(R)	#	%
M	O	M	O	M	M	O	18	0.15
M	M	M	O	M	M	O	18	0.15
O	O	M	O	M	O	O	13	0.11
O	M	M	M	O	O	O	13	0.11
O	O	M	M	O	O	O	12	0.10
O	O	M	M	O	O	M	12	0.10
O	O	M	O	M	O	M	12	0.10
O	O	M	O	M	M	O	12	0.10
O	O	M	O	M	M	M	12	0.10
O	M	O	M	O	O	M	12	0.10
O	M	O	O	O	M	M	12	0.10
O	M	O	O	M	M	O	12	0.10
O	M	M	M	M	M	M	12	0.10
M	M	O	M	O	O	M	12	0.10
M	M	O	O	O	M	O	12	0.10
M	M	O	O	O	M	M	12	0.10
M	M	O	O	M	M	M	12	0.10
M	M	M	O	M	O	O	12	0.10
M	M	M	M	M	M	M	12	0.10
O	M	M	O	O	O	O	11	0.09
M	O	M	M	M	O	O	6	0.05
M	M	M	O	O	O	O	6	0.05
M	M	M	M	O	M	M	6	0.05
M	M	M	O	M	O	M	6	0.05
M	M	M	O	M	M	M	6	0.05
M	O	M	O	O	M	M	4	0.03
M	O	M	M	O	O	O	1	0.008
# 2,288	# 1,858	# 981	# 735	# 2,031	# 1,264	# 2,291	11,325	
% 20.20	% 16.40	% 8.66	% 6.49	% 17.93	% 11.16	% 20.22		

M, O, #, and % stand for missing, observed, number, and percentage respectively.

is the total number of input vectors (x_i). The Gaussian neighborhood function, which is the most commonly used one, is defined as follows:

$$N_{j*}(t) = \exp\left(-\frac{\|r_{j*} - r_j\|^2}{2\delta^2(t)}\right) \quad (2)$$

where $\|r_{j*} - r_j\|$ is the distance between neurons $j*$ and j and $\delta(t)$ is the neighborhood radius at iteration t. These procedures must be iterated several times until the results can be considered as steady.

In this study, the missing values were ignored during training. Once the SOM was trained, the BMUs for the incomplete data vectors were found (using the Euclidean distance as explained before) and consequently their missing values were filled in by copying the corresponding values of the BMUs (weight vectors). The optimum number of neurons in the output layer was determined by a trial-and-error procedure with map sizes ranging from 10 × 10 to 30 × 30. A map size of 25 × 25 was selected as the optimum grid size based on the average quantification error,

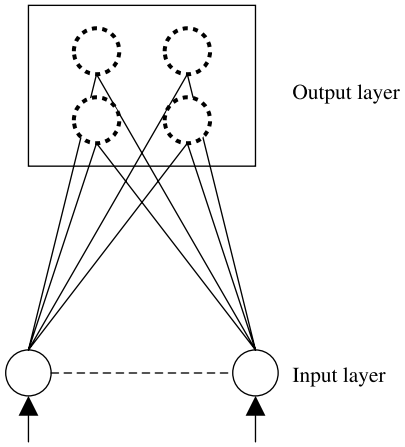

Figure 1 | A two-dimensional self-organizing map (SOM) with 2 × 2 neurons in the output layer.

which is defined as the average distance from each data vector to its BMU. This SOM model was utilized for filling in the missing data in the original data matrix.

Feed-forward multilayer perceptron (MLP)

Artificial neural networks (ANNs) have become quite popular within hydrological analysis, mostly in simulation applications. The most widely used ANN in water resources and hydrological applications is the feed-forward MLP. Figure 2 shows a three-layer feed-forward MLP that consists of an input layer, a hidden layer and an output layer. Each neuron in a layer is connected to all the neurons of the next layer via connection weights. In a feed-forward MLP, connections between neurons flow in one direction: from the input layer, through one or more hidden layers, to the output layer. These networks are generally trained by means of an error back-propagation algorithm, which is the most popular algorithm for training. As there exists an extensive literature about applications of ANNs, in particular feed-forward MLP, in water resources and hydrology (e.g. ASCE 2000*a*,*b*; Maier & Dandy 2000; Dawson & Wilby 2001), we will not go through that in detail and interested readers are referred to those references.

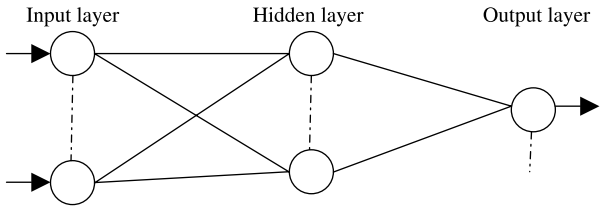

Figure 2 | Structure of a three-layer feed-forward multilayer perceptron (MLP).

In this study, several feed-forward MLP models were separately trained, one per missing variables combination. The number of input neurons is equal to the number of available values and the number of output neurons is equal to the number of missing values for each missing variables combination. The number of hidden neurons was determined by using the following formula, which was adopted by Junninen *et al.* (2004) in their application on air quality datasets:

$$H_{\text{in}} = (2 \times N_{\text{in}}) + 1$$
$$H_{\text{out}} = (2 \times N_{\text{out}}) + 1$$
$$\text{if } H_{\text{out}} < H_{\text{in}}, \quad \text{then } H_{\text{MLP}} = H_{\text{in}}, \quad \text{else } H_{\text{MLP}} = H_{\text{out}} \tag{3}$$

where H_{in} and H_{out} are the number of hidden neurons determined by the number of input and output neurons (N_{in} and N_{out}), respectively. In this study, we have adopted the above formula to determine the number of neurons in the hidden layer of each feed-forward MLP model per missing variables combination. Logistic sigmoid transfer functions were used both in the hidden layer and the output layer. To avoid instabilities, the feed-forward MLP was trained 10 times where each run starts with different initial conditions, as recommended by Hsieh & Tang (1998) among others, and then the final output was obtained by averaging the output from all of the runs. In total, 68 three-layer feed-forward MLP models were separately trained, one for each missing variables combination.

Multivariate nearest-neighbor (MNN)

Dixon (1979) described the multivariate nearest-neighbor imputation technique for dealing with missing values. In this approach, the missing values of a vector are filled in by determining which vector is the most similar vector to the vector of interest and replacing the missing values by corresponding values from that vector. This similarity can be computed by a distance function, which is in our case was Euclidean. The procedure can be summarized as follows:

- Divide the dataset into two parts: one containing vectors in which at least one of the components is missing while the remaining part will consist of the complete vectors, i.e. those without any missing values.
- For each vector in the data matrix with missing values, find the nearest neighbor from the complete data matrix. In the distance calculation, use only components that are not missing.

Regularized expectation–maximization (REGEM) algorithm

These methods are based on estimated regression models between missing and available data. In this study, the regularized expectation–maximization (EM) algorithm based on iterated linear regression analyses (hereafter REGEM) (Schneider 2001) was utilized. Detailed descriptions of the EM algorithm can be found elsewhere (e.g. Dempster *et al.* 1977).

Missing values were filled in with the regularized EM algorithm. And in any iteration *t*of the REGEM algorithm, estimates of the mean vector $\hat{\mu}^{(t)}$ and the covariance matrix $\hat{\Sigma}^{(t)}$ were revised in three steps. In the first iteration, $\hat{\mu}^{(t)}$ (the sample mean of the available values) and $\hat{\Sigma}^{(t)}$ was obtained from the completed database after substituting the missing values with the mean values.

- First, for each observation $x = x_i$ $(i = 1, \ldots, n)$, we can divide the observation into two subsets $x{:}x_o{,}x_m$, where the vector x_o consists of the variables for which the values are available and the vector x_m consists of the remaining variables for which the values are missing. Then, the regression parameters $\hat{B}$ of the variables can be computed from the estimates of the mean $\hat{\mu}^{(t)} : \hat{\mu}_o, \hat{\mu}_m$ and the covariance matrix $\hat{\Sigma}^{(t)} : \hat{\Sigma}_{oo}, \hat{\Sigma}_{mm}, \hat{\Sigma}_{om}, \hat{\Sigma}_{mo}$ as follows:

$$\hat{B} = (\hat{\Sigma}_{oo} + h^2 \hat{A})^{-1} \hat{\Sigma}_{om} \tag{4}$$

where $\hat{B}$ is the matrix of estimated regression coefficients; $\hat{\Sigma}_{oo}$ is the submatrix from the estimated covariance matrix $\hat{\Sigma}^{(t)}$ that consists of the estimated variances and covariances of the variables for which, in the given observation, the values are available; the scalar h is a regularization (ridge) parameter which is used for a rank-deficient database in estimating regression parameters; A is the diagonal matrix with the diagonal elements of the covariance matrix $\hat{\Sigma}_{oo}$; $\hat{\Sigma}_{om}$ and $\hat{\Sigma}_{mo}$ are submatrices that consist of the estimated cross-covariances of the variables for which, in the given observation, the values are available with the variables for which, in the given observation, the values are missing; $\hat{\Sigma}_{mm}$ is the submatrix from the estimated covariance matrix $\hat{\Sigma}^{(t)}$ that consists of the estimated variances and covariances of the variables for which, in the given observation, the values are missing; the vector $\hat{\mu}_o$ is that partition of the estimated mean $\hat{\mu}^{(t)}$ that belongs to the variables for which, in the given observation, the values are available; and the vector $\hat{\mu}_m$ is that partition of the estimated mean $\hat{\mu}^{(t)}$ that belongs to the variables for which, in the given observation, the values are missing.
- Second, the missing values in the given observation are filled in with conditional expectation values given the available values and mean and covariance matrix estimates $\hat{x}_m \equiv E(x_m | x_o; \hat{\mu}^{(t)}, \hat{\Sigma}^{(t)})$ as

$$\hat{x}_m = \hat{\mu}_m + (x_o - \hat{\mu}_o)\hat{B} \tag{5}$$

where the parameters are similar to those explained before.
- Third, once the missing values in all observations $x = \mathrm{x}_i$ $(i = 1, \ldots, n)$ were filled in with $\hat{x}_m$, re-estimate the mean $\hat{\mu}^{(t+1)}$ as

$$\hat{\mu}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6}$$

and covariance matrix $\hat{\Sigma}^{(t+1)}$ from the conditional expectation of the cross-products $\hat{C}_i^{(t)} \equiv E\,[x_i^{\mathrm{T}}\, x_i \mid x_o; \hat{\mu}^{(t)}, \hat{\Sigma}^{(t)}]$ as

$$\hat{\Sigma}^{(t+1)} = \frac{1}{\tilde{n}}\sum_{i=1}^{n}\left\{\hat{C}_i^t - \left[\hat{\mu}^{(t+1)}\right]^{\mathrm{T}}\hat{\mu}^{(t+1)}\right\} \tag{7}$$

where $\tilde{n}$ is the number of degrees of freedom. The conditional expectation $\hat{C}_i^{(t)}$ of the cross-products, for each observation $x = x_i$ $(i = 1, \ldots, n)$, consist of three parts; two that involve the available values in the

observation

$$E\ (x_o^{\mathrm{T}}x_o \mid x_o; \hat{\mu}^{(t)}, \hat{\Sigma}^{(t)}) = x_o^{\mathrm{T}}\ x_o \tag{8}$$

$$E\ (x_o^{\mathrm{T}}x_m \mid x_o; \hat{\mu}^{(t)}, \hat{\Sigma}^{(t)}) = x_o^{\mathrm{T}}\ x_m \tag{9}$$

and one that only considers the filled-in values in the observation

$$E\ (x_m^{\mathrm{T}}x_m \mid x_o; \hat{\mu}^{(t)}, \hat{\Sigma}^{(t)}) = \hat{x}_m^{\mathrm{T}}\hat{x}_m + \hat{R} \tag{10}$$

Here, $\hat{R} = \mathrm{Cov}(x_m, x_m \mid x_o; \hat{\mu}^{(t)}, \hat{\Sigma}^{(t)})$ is the residual covariance matrix.

The next iteration of the REGEM algorithm is conducted with the new estimates $\hat{\mu}^{(t+1)}$ and $\hat{\sum}^{(t+1)}$. The iterations are continued until convergence, that is, the filled-in values and the estimates of the mean and covariance matrix parameters do not change substantially.

It is worth mentioning that REGEM provided negative values, but very small ones, for some missing observations. However, we substituted negative values with zero.

Multiple imputation (MI)

MI is a fairly straightforward procedure. It involves, firstly, the generation of several possible values for each missing observation in order to generate several completed databases and, secondly, analyzing each database separately. The number of completed databases to generate depends on the extent of the missing data. However, Schafer (1997) suggests that five completed databases typically provide unbiased estimates.

In this study, we used Schafer's (1999) NORM program to generate five possible values for each missing observation using the data augmentation algorithm. NORM refers to the multivariate normal distribution that the model uses to generate imputations. The data augmentation algorithm treats parameters and missing data as random variables and simulates random values of parameters and missing data from their conditional distribution. The procedure of simulating parameters and missing data generates a chain that, for a sufficient number of iterations, converges to the Bayesian posterior distribution. We ran the EM algorithm prior to the run of the data augmentation algorithm, in order to obtain initial values for the data augmentation algorithm and to assess the number of iterations needed to create statistically independent imputations. By specifying 50 iterations as sufficient to ensure statistically independent values, the augmented databases at iterations 50, 100, 150, 200 and 250 were saved, hence generating five completed, independent databases. As indicated already, each completed database can be analyzed separately by means of standard statistical methods and thereafter the five estimates can be combined by averaging in order to produce a single estimate. As stated by Rubin (1996) the main objective of MI is to get valid statistical inference about the database with missing values rather than optimal point predictions.

As in REGEM, NORM provided negative values, but very small ones, for some missing observations on each completed database. However, we substituted negative values with zero.

Software

The SOM and MNN analyses were carried out by utilizing the available functions in the SOM Toolbox, which is freely available and can be downloaded from http://www.cis.hut.fi/projects/somtoolbox/. All the MLP models were performed using available functions in MATLAB. The MATLAB code for the implementation of the REGEM can be downloaded from http://www.gps.caltech.edu/~tapio/imputation/. And, finally, the NORM program can be downloaded from http://www.stat.psu.edu/~jls/misoftwa.html.

RESULTS AND DISCUSSION

This study uses the SOM, MLP, MNN, REGEM and MI methods to impute the missing values in a multivariate precipitation–runoff database from northern Iran. As stated previously, we deal with real missing values, hence it is not possible to compute the imputation errors on real missing observations to make a fair comparison of the methods. However, the standard deviation and mean of the variables under investigation before and after imputation along with the change ('%) of these parameters after imputation were used to evaluate each method. The univariate statistical analyses of the results are shown in Table 3.

Table 3 | Statistical analysis of the database. Std*—standard deviation after imputation, Std†—standard deviation before imputation, Mean*—mean after imputation, Mean†—mean before imputation, ΔStd—change (%) in standard deviation after imputation, ΔMean—change (%) in mean after imputation

	SOM					MLP					MNN					REGEM					MI				
Variables	Std*	Std†	ΔStd	Mean*	Mean†	ΔMean	Std*	ΔStd	Mean*	ΔMean	Std*	ΔStd	Mean*	ΔMean		Std*	ΔStd	Mean*	ΔMean		Std*	ΔStd	Mean*	ΔMean	
13001(P)	4.72	4.91	3.86	1.44	1.53	5.88	4.99	1.62	2.50	63.39	4.69	4.48	1.41	7.84		4.49	8.55	1.44	5.88		4.62	5.90	1.67	9.15	
13004(P)	4.52	4.53	0.22	1.51	1.49	1.34	4.92	8.60	2.44	63.75	4.42	2.42	1.47	1.34		4.23	6.62	1.48	0.67		4.35	3.97	1.64	10.06	
13005(P)	3.83	3.92	2.29	1.21	1.25	3.20	4.35	10.96	1.80	44.00	3.81	2.80	1.19	4.80		3.78	3.57	1.20	4.00		3.80	3.06	1.27	1.60	
13005(R)	1.62	1.64	1.21	1.25	1.25	0.00	1.71	4.26	1.40	12.00	1.63	0.60	1.25	0.00		1.88	14.63	1.26	0.80		1.96	19.51	1.28	2.40	
13007(P)	5.90	5.82	1.37	2.22	2.17	2.30	5.66	2.74	2.90	33.64	5.80	0.34	2.18	0.46		5.65	2.92	2.21	1.84		5.75	1.20	2.40	10.59	
13013(P)	6.20	6.34	2.20	2.02	2.09	3.34	6.46	1.89	2.81	34.44	6.17	2.68	2.03	2.87		6.06	4.41	2.04	2.39		6.14	3.15	2.20	5.26	
13013(R)	18.03	18.01	0.11	5.30	5.17	2.51	64.43	257.74	31.10	501.54	16.35	9.21	5.15	0.38		16.49	8.43	5.48	5.99		17.02	5.49	6.07	17.40	

*After.
†Before.

As seen from the table, all statistics were severely violated on the runoff variable at station 13013. This finding indicates that the quality of data is more important than the amount of missing values for the imputation result. Junninen *et al.* (2004) concluded that univariate methods are dependent on the gap length as well as on the variable under study. The authors showed that, among the multivariate methods, both SOM and MLP exhibited slightly better imputation ability than the others. However, our results show that multivariate methods are also dependent on the variable under investigation. Moreover, our study indicates that there is a substantial difference between SOM and MLP in terms of handling missing value problems. This difference is particularly significant with regard to the data quality of the variables under investigation. Other research applications have also shown the overall advantages of SOM over MLP in imputation of missing values (e.g. Fessant & Midenet 2002; Kalteh & Berndtsson 2007), hence our finding is consistent with the findings of the aforementioned applications. In addition, one single SOM can handle all missing data patterns altogether and there is no need to develop separate models for each missing data pattern. Moreover, the SOM model can provide more detailed information of the process under investigation via, for example, the component plane visualization technique. This technique is used to visualize variables under investigation for further analysis such as regionalization. Each plane shows the values of one variable in each map neuron. By looking at the planes, similarities among variables can be visualized, to reveal where there are similar patterns in identical locations on the component planes. Kalteh & Berndtsson (2007) used this technique for correlation hunting among the variables and consequently to regionalize precipitation stations in northern Iran.

The MNN performs almost as well as the other methods, which may indicate that such a simple method is very important. In particular, it has some advantages such as being computationally inexpensive as well as using existing values in the database to impute missing values, hence it does not generate new values.

As stated previously, the most important aspect of MI is that, with the statistical analyses of the imputed databases and thereafter combining the results, it makes it possible to obtain statistically valid inference about the database with

Table 4 | Multiple imputation results

Variables	13001(P)		13004(P)		13005(P)		13005(R)		13007(P)		13013(P)		13013(R)	
Dataset	Std*	Mean*	Std*	Mean*	Std*	Mean*	Std*	Mean*	Std*	Mean*	Std*	Mean*	Std*	Mean*
1	4.61	1.67	4.36	1.65	3.80	1.27	1.97	1.29	5.74	2.40	6.16	2.21	17.07	6.11
2	4.63	1.70	4.35	1.65	3.81	1.28	1.94	1.28	5.74	2.38	6.15	2.21	16.98	6.03
3	4.62	1.66	4.36	1.64	3.81	1.27	1.95	1.28	5.76	2.40	6.14	2.21	17.07	6.09
4	4.61	1.67	4.35	1.65	3.81	1.28	1.97	1.28	5.77	2.39	6.15	2.20	17.00	6.08
5	4.63	1.69	4.36	1.65	3.81	1.28	1.97	1.28	5.77	2.43	6.13	2.18	17.01	6.06

*After.

missing values. Thus, to make point predictions of individual missing values is not the main goal. Rather, it is to prepare the complete database for the ultimate user's needs in order to obtain statistically valid inference of particular importance. In Table 4, we show the univariate statistical inferences of the five imputed databases, similar to the one shown in Table 3 for the other imputation methods. By using the MI technique, we have created five versions from the original incomplete precipitation–runoff process database; thereafter it is possible to conduct the analyses of choice on each database separately and then to combine the results by averaging over the five analyses in order to obtain a single inference from the original incomplete precipitation–runoff database. This will reflect the uncertainty attached to the missing data. The resulting final single inferences, obtained by averaging of the MI results in Table 4, are shown in the MI results in Table 3.

The bivariate correlation coefficients between the runoff variable at station 13013 and other variables are shown in Table 5. This station is located at the outlet of the watershed, hence we calculated the correlation coefficients of this variable with the others. As seen from the table, the correlation coefficients have decreased for all the methods, except at 13005(R) for REGEM and MI, compared to the complete-case (CC) analysis. Increasing correlation coefficients at 13005(R) for REGEM and MI can be explained by the fact that these methods carry the assumption of multivariate normality. As seen in Table 3, the changes of the standard deviation after imputation for this variable with these methods are very high compared to the other methods. Unsurprisingly, the MLP method shows the poorest performance compared to the others, which is consistent with the results obtained from the univariate statistics in the former analysis. Even though the correlation coefficients are not very high, the point that the different methods, except MLP, provide approximately similar results gives confidence about the applicability of the methods in the context of precipitation–runoff missing data imputation. As previously indicated, the missing data problem in our database is very complicated because the fractions of missing values in most of the variables are large, which certainly affects the performance of the methods.

To summarize, a stringent comparison of the methods was not the main objective of this study. Rather, our effort was focused on illustrating the applicability of various imputation methods on real missing values in a precipitation–runoff process database, which is required for further studies such as runoff prediction. As is evident from the above discussion, each of the methods have advantages and disadvantages, hence our objective is not to introduce a single method to impute missing values in hydrological applications. However, in our case, the SOM and MNN methods seem to perform better compared to the others. Another important aspect of this study is that, according to the authors' knowledge, it is the first time that

Table 5 | Comparison of correlations between 13001(P), 13004(P), 13005(P), 13005(R), 13007(P), 13013(P) and 13013(R)

	13001(P)	13004(P)	13005(P)	13005(R)	13007(P)	13013(P)
CC	0.35	0.32	0.43	0.37	0.42	0.41
SOM	0.23	0.29	0.30	0.21	0.25	0.14
MLP	0.04	0.22	0.22	0.17	0.09	0.09
MNN	0.20	0.24	0.29	0.28	0.24	0.14
REGEM	0.20	0.27	0.30	0.69	0.26	0.14
MI	0.18	0.24	0.27	0.69	0.23	0.12

the MI method was used to impute missing values for a hydrological application. The SOM, MNN, REGEM and MI methods provided good results; however, imputing missing data is not without danger. As stated by Dempster & Rubin (1983) (quoted in Little & Rubin (2002)).

"The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases" (p 59).

CONCLUSIONS

The main objective of this study was to illustrate methods to impute (or "fill in") real missing values in a hydrological process database in order to allow the data to be used for rainfall–runoff modelling, which requires serially complete data. We explored different methods, i.e. the SOM, MLP, MNN, REGEM and MI, and made comparative analyses of the results. The results show that most of the methods yield similar results. However, there is quite a significant difference between the imputation power of the SOM neural network and the MLP, in particular for the variable which contains extreme values. The SOM is less sensitive to the location of missing values if compared to the MLP, which needs to train separate models for different combinations of variables which may lead to incoherence between imputed values. The MNN is a simple method to implement and computationally cheap. It performs as well as the other methods even for the variable with extreme values. Although most studies tend to result in the conclusion that there is no method that is superior in all applications, there seems to be an agreement that one should use statistically principled methods. This speaks in favour of REGEM and MI. MI has the great advantage of giving an estimate of the uncertainties involved in the imputation.

However this comes at a cost. Both REGEM and MI build on the assumption that the data are multivariate normal. Our results indicate that there may be significant bias if this assumption is violated. Thus, it seems safer to rely on either SOM or MNN.

MNN requires no assumptions about the probability density function. It seems to produce reliable estimates, and has the great advantage of keeping the imputed values within the range defined by the values that have actually been observed. Although SOM is considered to be a black-box methodology, it has the advantage of providing more detailed information of the process under investigation via, for example, the component plane visualization technique.

Although most studies conclude that there is no imputation methodology that can be characterized as superior to others, we suggest that for a rainfall–runoff database, SOM and MNN provide the most robust and reliable results.

REFERENCES

Abebe, A. J., Solomatine, D. P. & Venneker, R. G. W. 2000 Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. *Hydrol. Sci. J.* **45**(3), 425–436.

Allison, P. D. 1998 Multiple imputation for missing data: a cautionary tale. Available from: http://www.ssc.upenn.edu/~allison/ Accessed July 15 2006.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000*a* Artificial neural networks in hydrology. I: preliminary concepts. *J. Hydrol. Eng.* **5**(2), 115–123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000*b* Artificial neural networks in hydrology. II: hydrologic applications. *J. Hydrol. Eng.* **5**(2), 124–137.

Bhattacharya, B., Shrestha, D. L. & Solomatine, D. P. 2003 Neural networks in reconstructing missing wave data in sedimentation modelling. In *Proceedings of 30th IAHR Congress, Thessaloniki, Greece*.

Dawson, C. W. & Wilby, R. L. 2001 Hydrological modelling using artificial neural networks. *Prog. Phys. Geogr.* **25**(1), 80–108.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **39**, 1–38.

Dempster, A. P. & Rubin, D. B. 1983 Introduction. In: Madow, W. G., Olkin, I. & Rubin, D. B. (eds) *Incomplete Data in Sample Surveys. Theory and Bibliography*, (**Vol. 2**). Academic Press, New York, pp. 3–10.

Dixon, J. K. 1979 Pattern recognition with partly missing data. *IEEE Trans. Syst. Man Cybern. SMC-9* **10**, 617–621.

Fessant, F. & Midenet, S. 2002 Self-organizing map for data imputation and correction in surveys. *Neural Comput. Appl.* **10**, 300–310.

Frane, J. W. 1976 Some simple procedures for handling missing data in multivariate analysis. *Psychometrika* **41**, 409–415.

Hsieh, W. W. & Tang, B. 1998 Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Am. Meteor. Soc.* **79**, 1855–1870.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. & Kolehmainen, M. 2004 Methods for imputation of missing values in air quality data sets. *Atmos. Environ.* **38**, 2895–2907.

Kalteh, A. M. & Berndtsson, R. 2007 Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). *Hydrol. Sci. J.* **52**(2), 305–317.

Khalil, M., Panu, U. S. & Lennox, W. C. 2001 Groups and neural networks based streamflow data infilling procedures. *J. Hydrol.* **241**, 153–176.

Kohonen, T. 1982*a* Analysis of a simple self-organizing process. *Biol. Cybern.* **44**, 135–140.

Kohonen, T. 1982*b* Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69.

Kohonen, T. 2001 *Self-Organizing Maps*. Springer-Verlag, Berlin.

Kuligowski, R. J. & Barros, A. P. 1998 Using artificial neural networks to estimate missing rainfall data. *J. AWRA* **34**(6), 1437–1447.

Little, R. J. A. & Rubin, D. B. 2002 *Statistical Analysis with Missing Data*. Wiley, New York.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Modell. Softw.* **15**, 101–124.

Musil, C. M., Warner, C. B., Yobas, P. K. & Jones, S. L. 2002 A comparison of imputation techniques for handling missing data. *West. J. Nurs. Res.* **24**(7), 815–829.

Patrician, P. A. 2002 Multiple imputation for missing data. *Res. Nurs. Health* **25**, 76–84.

Pigott, T. D. 2001 A review of methods for missing data. *Edu. Res. Eval.* **7**(4), 353–383.

Ramirez, M. C. V., Velho, H. F. D. C. & Ferreira, N. J. 2005 Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region. *J. Hydrol.* **301**, 146–162.

Rubin, D. B. 1996 Multiple imputation after 18 + years. *J. Am. Stat. Assoc.* **91**, 473–489.

Schafer, J. L. 1997 *Analysis of Incomplete Multivariate Data*. Chapman & Hall, New York.

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. & Graham, J. W. 2002 Missing data: our view of the state of the art. *Psychol. Methods* **7**(2), 147–177.

Schneider, T. 2001 Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.* **14**, 853–871.

Tsikriktsis, N. 2005 A review of techniques for treating missing data in OM survey research. *J. Oper. Manage.* **24**, 53–62.

Zoppou, C., Roberts, S. & Hegland, M. 2000 Spatial and temporal rainfall approximation using additive models. *ANZIAM J.* **42**, C1599–C1611.

First received 5 September 2006; accepted in revised form 14 November 2008