Article:

Nkiaka, E, Nawaz, NR and Lovett, JC (2016) Using Self-Organizing Maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. Environmental Monitoring and Assessment, 188. 400. ISSN 0167-6369

https://doi.org/10.1007/s10661-016-5385-1

Nkiaka, E., Nawaz, N.R. & Lovett, J.C. 2016. Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. Environmental Monitoring and Assessment.
DOI 10.1007/s10661-016-5385-1

Using Self-Organizing Maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin

E. Nkiaka[1], N.R. Nawaz[1] and J.C. Lovett[1]
[1]School of Geography, University of Leeds:
Correspondence to: E. Nkiaka (gyenan@leeds.ac.uk)

Abstract

Hydro-meteorological data is an important asset that can enhance management of water resources. But existing data often contains gaps, leading to uncertainties and so compromising their use. Although many methods exist for infilling data gaps in hydro-meteorological time series, many of these methods require inputs from neighbouring stations, which are often not available, while other methods are computationally demanding. Computing techniques such Artificial Intelligence can be used to address this challenge. Self-Organizing Maps (SOMs), which are a type of Artificial Neural Network, was used for infilling gaps in a hydro-meteorological time series in a Sudano-Sahel catchment. The coefficients of determination obtained were all above 0.75 and 0.65 while the average topographic error was 0.008 and 0.02 for rainfall and river discharge time series respectively. These results further indicate that SOMs are a robust and efficient method for infilling missing gaps in hydro-meteorological time series.

Keywords: Artificial Neural Networks, hydro-meteorological data, infilling missing data, Logone catchment, Self-Organizing Maps,

1) Introduction

Economic progress, rising standard of living, growing populations and expansion of commercial agriculture in developing countries is putting increasing pressure on fresh water resources (WWAP, 2015). At the same time climate extremes such as droughts and floods are becoming more frequent (Coumou & Rahmstorf, 2012). Better informed water resource management is needed to respond to demand and climate variability. A major requirement for planning is the availability of good quality and long term hydro-meteorological data. This data provides indicators of past hydro-climatic behaviour of a region/catchment and is fundamental to the development of models for prediction of system behaviour (Harvey et al., 2012).

Existing hydro-meteorological time series used for planning and management decisions often contains missing observations, particularly in developing countries. The gaps are caused by many reasons, including equipment failure, destruction of equipment by natural catastrophes such as floods, war and civil unrest, mishandling of observed records by personnel or loss of files containing the data in a computer system (Elshorbagy et al., 2000). The presence of gaps, even if there are very short, in a hydro-meteorological time series can hinder calculation of important statistical parameters as data patterns maybe hidden. This can compromise their use for water resources planning as it increases the level of uncertainty in the datasets (Ng and Panu, 2010; Campozano et al., 2014). This problem is particularly acute in the Sudano-Sahel region where rainfall is highly variable in both space and time, meteorological and flow gauging stations are scarce and the available datasets are riddled with gaps.

Several methods exist for infilling gaps in hydro-meteorological time series. However, the application of each method depends on a range of factors including the information available for that station; additional datasets from neighbouring stations; the percentage of gaps present within the time series to be infilled; the season within which the gaps are present; the length of the existing data series; and the type of application that the infilled series will be used for (Mwale et al., 2012). These infilling methods range from simple techniques such as linear interpolation, Inverse Distance Weighting (IDW) and Thiessen polygons; to more complicated advanced techniques such as time series models, Markov models, Global Imputation, Multiple Regression models, Artificial Intelligence (Kalteh & Hjorth, 2009; Presti et al., 2010; Ismail et al., 2012; Mwale et al., 2012; Campozano et al., 2014).

Most of the methods, require additional input data from neighbouring stations in order to produce reliable results and these additional inputs are often not available. Furthermore, some of the methods are time consuming and demand subtantial computer power for simulation because of the complicated algorithms involved (Presti et al., 2010). Some methods also require that the time series be split into different seasons to obtain reliable results. Although these challenges could be overcome by using numerical models (e.g. hydrological models); models also demand high data inputs and cannot be applied to many stations at the same time due to parameter calibration requirements which are site specific and consequently results cannot be transferred to other stations even within the same catchment (Harvey et al., 2012).

Some of these challenges can be overcome by using computing techniques such as Artificial Intelligence (AI) (Daniel et al., 2011). In this class of technique, the most promising approaches include Artificial Neural Networks (ANN), Fuzzy Logic (FL) and Genetic Algorithms (GA). The application of Artificial Intelligence in hydrology and water resources management is well established (ASCE 2000; Kingston et al., 2008a, 2008b; Daniel et al., 2011). Among the AI class of models, ANNs are probably the most popular as these use available data to learn about the behaviour of a time series. In addition, they possess capabilities for modelling complex nonlinear systems; do not require prior knowledge of the system process(s) under study and are robust even in the presence of missing observations in the time series (Mwale et al., 2012). The main advantage of ANNs over conventional methods is their ability to model physical processes without the need for detailed information of the system (Daniel et al., 2011); and they have often been used for infilling gaps in hydro-meteorological time series (Kalteh & Hjorth, 2009; Dastorani et al., 2010; Adeloye et al., 2012; Ismail et al., 2012; Mwale et al., 2012; Mwale et al., 2014; Kim et al., 2015).

Within the ANN family, the Multilayer Perceptron (MLP) is one of the most widely used for infilling gaps in hydro-meteorological time series (Kalteh & Hjorth, 2009; Dastorani et al., 2010; Mwale et al., 2012; Mwale et al., 2014; Kim et al., 2015). Although MLP is robust for performing this task, it usually demands a long time series for training; and if part of the data to be used for training is missing, additional pre-processing of the time series will have to be carried out to provide estimates in the input space before the training can begin (Rustum & Adeloye, 2007; Mwale et al., 2012). This therefore limits application in situations where significant portions of the time series to be used for training have incomplete data; or for short time series as the data may not be sufficient for training. It is also computationally intensive and needs additional storage memory (Kalteh et al., 2008).

Another member of the class of ANNs known as Self-Organizing Maps (SOMs), which is a competitive and unsupervised ANN, is becoming popular for infilling gaps in hydro-meteorological times series and has been shown to outperform ANNs-MLP (Kalteh & Hjorth, 2009; Mwale et al., 2014; Kim et al., 2015). Many studies have successfully applied SOMs for infilling gaps in hydro-meteorological time series with satisfactory results (Kalteh & Hjorth, 2009; Rustum & Adeloye, 2011; Adeloye et al., 2012; Mwale et al., 2012; Mwale et al., 2014; Kim et al., 2015).

Self-Organizing Maps (SOMs) were first introduced by Kohonen, (1995, 1997). The success of their application in other research disciplines led to their wide application in water resources processes and systems research especially for data mining, infilling of missing data, estimation and flow forecasting, clustering etc. (Kalteh et al., 2008). This is due to their ability to convert nonlinear statistical relationships between high dimensional data onto a low dimensional display (Ismail et al., 2012). Data points that show similar characteristics are placed closed to each other or clustered together in the output space. This mapping approach does a quasi-preservation of the most important topological and metric relationship of the original data (Rustum & Adeloye, 2007). Adeloye et al. (2012) asserted that, the ability of SOMs to cluster data together makes them robust for data mining and infilling datasets with gaps and outliers as the gaps/outliers are replaced by their features in the map. The SOMs algorithm generally executes assigned tasks using an unsupervised and competitive learning approach to discover patterns in the data (Kalteh & Berndtsson, 2007) thus, the whole process in entirely data driven. A SOM is made up of two layers: a multi-dimensional input layer and an output layer. Both layers are fully connected by adjustable weights and the output layer is made up of neurons arranged in a two dimensional grid of nodes (Figure 1). Each neuron in the output layer of the SOM contains exactly the same set of variables contained in the input vectors. Despite its wide application for infilling missing data in many studies around the world, it has rarely been used Africa in general and the Sudano-Sahel region in particular.

A Self-Organizing Maps approach was applied to infill missing data in monthly rainfall and daily river discharge time series from January 1950 to December 2007 in the Logone river catchment covering Cameroon, the Central Africa Republic and Chad. Infilling of missing gaps in hydro-meteorological time series usually precedes most hydro-climatic studies (Kashani & Dinpashoh, 2012), and this work is part of an on-going research project to assess the vulnerability of this catchment to drought and flood events under anticipated increased climate variability.

The paper is structured as follows: Section 2 describes the data and methodology used in the study. In Section 3 the results obtained are presented and discussed. Section 4 gives a general summary and conclusion of the study.

2) Methodology

2.1) Study area

The Logone catchment is part of the greater Lake Chad basin. It lies between latitude 6˚-12˚N and longitude 13˚-16˚E and is a transboundary catchment in the Sudano-Sahel transitional zone in Central Africa with an estimated catchment area of 86,500 km^2 (Figure 1).

The Logone River has its source in Cameroon through the Mbere and Vina Rivers, which flow from the northeastern slopes of the Adamawa plateau. It is joined in Lai by the Pende River from the Central Africa Republic and flows from south to north to join the Chari River in Ndjamena (Chad) and continue flowing in a northward direction before finally emptying into Lake Chad. The climate in the catchment is characterized by high spatial variability and is dominated by seasonal changes in the tropical continental air mass (the Harmattan) and the marine equatorial air mass (monsoon) (Candela et al., 2014).

2.2) Data Sources

Monthly gauge rainfall was obtained from SIEREM (Boyer et al., 2006) available for 18 stations covering the period 1950-2000 while daily river discharge data was obtained from the Lake Chad Basin Commission (LCBC). Discharge time series are available for the stations of Lai, Bongor, Katoa and Logone Gana covering the period 1973-1998 for Lai and 1983-2007 for the rest of the stations.

2.3) Implementation of the SOM Algorithm

A SOM algorithm is implemented in a series of steps.

The multi-dimensional input data is first standardized to make sure that very high or low value variables do not dominate the map. Since SOMs use Euclidian metrics to measure distances between vectors, standardization gives equal weight to all the input variables (Vesanto et al., 2000). In this analysis, data was not standardized because rainfall and river discharge time series were trained separately.

The input vector is then chosen at random and presented to each of the individual neurons for comparison with their weight vectors in order to identify the weight vector most similar to the presented input vector. The identification uses the Euclidian distance defined as:

$$D_i = \sqrt{\sum_{j=1}^{n} m_j (x_j - w_{ij})^2}; \quad i = 1, 2, 3 \ldots.\ M \qquad (1)$$

Where:

D_i = Euclidian distance between the input vector and the weight vector i; x_j = j element of the current vector; w_{ij} = j element of the weight vector I; n = the dimension of the input vector; m_j = "mask".

When the input vector contains missing elements, the mask is set to zero for such elements and because of this, the SOM algorithm can conveniently handle missing elements in the input vector. The neuron whose vector closely matches the input vector (i.e. with D_i minimum) is chosen as the winning node or best matching unit (BMU).

After finding the BMU, the weight vector of the winner neuron is adjusted so that the BMU and its adjacent neurons move closer to the input vectors in the input space, thereby increasing the agreement between the input vector and the weight vector. This adjustment is carried out using the following equation:

$$w_t(t+1) = w_t(t) + \alpha(t) h_{ci}[x(t) - w_t(t)] \qquad (2)$$

Where: w_t = element of the weight vector; t = time; α(t) = learning rate at time t; h_{ci}(t) = neighbourhood function centred in the winner unit c at time t.

From here, each node in the map develops the ability to recognize input vectors that are similar to itself. This ability is referred to as self-organizing as no external information is added for this process to take place. The learning procedure continues until the SOM algorithm converges. Generally, the learning rate decreases monotonically as the number of iterations increase as shown by the following equation:

$$\alpha(t) = \alpha_0 (\frac{0.005}{\alpha_0})^{\frac{t}{T}} \qquad (3)$$

Where: α(t)= learning rate; α_0 = initial learning rate; T = training length

The neighbourhood function used is in this analysis is Gaussian centred in the winner unit c, calculated as:

$$h_{ci}(t) = \exp\left\{-\frac{\left||r_c - r_i|\right|^2}{[2\sigma^2(t)]}\right\} \qquad (4)$$

Where:

h_{ci}(t) = neighbourhood function centred in the winner unit c at time t; r_c and r_i = positions of nodes c and i on the SOM grid; σ(t) = neighbourhood radius which also decreases monotonically as the number of iterations increases.

The quality of the trained SOM is measured by the total average quantization error and total topographic error. The average quantization error is a measure of how good the map fits the input data (it measures the average distance between each data vector and its Best Matching Unit (BMU)). The smaller the quantization error, the smaller the average of the distance from the vector data to the prototypes, meaning that the data vectors are closer to its prototypes; it is a positive real number with a value close to zero indicating a good fit between the input and the map. The quantization error is calculated as:

$$q_e = \frac{1}{N}\sum_{i=1}^{N} \|X_i - W_{ic}\| \qquad (5)$$

Where: q_e = quantization error; N = number of input vectors used to train the map; X_i = ith data sample or vector; W_c =prototype vector of the best matching unit for X; ||.|| = denotes the Euclidian distance.

Topographic error measures how well the topology of the data is preserved by the map by considering the map structure. The lower the topographic error, the better the SOM preserves the topology of the data. It is a positive real number between 0 and 1 with a value close to 0 indicating good quality. It is calculated as:

$$t_e = \frac{1}{N}\sum_{i=1}^{N} u(X_i) \qquad (6)$$

Where: te = topographic error; N = number of input vectors used to train the map;

ui = binary integer such that it is equal to 1 if the first and second BMU for Xi are not adjacent units; otherwise it is zero.

Since there is always a trade-off between which of the two can be minimized at the expense of the other, in this study, effort was focused on reducing the topographic error to ensure that the infilled values reflect the seasonal trend of the different time series. The coefficient of determination (R^2) was used to check the quality of the newly generated time series. R^2 gives the proportion of the variance of one variable that is predictable from the other variable and varies between 0 and 1. R^2 is calculated as:

$$R^2 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}\ [(x_i - \bar{x})^2]\sum_{i=1}^{n}[(y_i - \bar{x})^2]} \tag{7}$$

Where: x_i = the ith observed value; y_i = the ith trained value; $\bar{x}$= the mean of observed value; $\bar{y}$ = the mean of the trained value; n = the number of observations.

2.4) Setting of SOM algorithm parameters

According to Gabrielsson & Gabrielsson, (2006), the radius of the SOM should be chosen wide enough at the beginning of the learning process so that the map can be ordered globally as the radius decreases monotonically with time. To determine the optimum number of neurons, if M is the total number of input elements, Garcia and Gonzalez (2004), propose that the number of neurons in the output can be calculated as:

$$N = 5\sqrt{M} \tag{8}$$

Where: M = total number of samples and N = the number of neurons.

Once N is known, Garcia and Gonzalez, (2004) further propose that the number of rows and columns of N can be calculated by:

$$\frac{l_1}{l_2} = \sqrt{\frac{e1}{e2}} \tag{9}$$

Where l_1 and l_2 are the number of rows and columns respectively, e1 is the biggest eigenvalue of the training data set and e2 is the second biggest eigenvalue.

In the initialization phase of the algorithm, since the learning process involve in the computation of a feature map is a stochastic process, according to Gabrielsson & Gabrielsson, (2006) the accuracy of the map depends on the number of iterations executed by the SOM algorithm. These authors recommend that for good statistical accuracy, the number of iterations should be at least 500 times the number of network nodes. In this study, the random initialization option was used as it is recommended for hydrological applications e.g. (Kalteh et al., 2008), while the default parameters set by the SOM software for map size and lattice (rows and columns) were adopted that were exactly the same as using equations (8) and (9).

The basic steps required to complete the infilling process consists of the following:

1) Data gathering and normalization: The data to be infilled (e.g. rainfall and discharge time series) is assembled together and standardized; these are the depleted input vectors;
2) Training: The depleted input vector (data matrix) is introduced to the iterative training procedure to form the SOM. At the beginning of the training, weight vectors must be initialized by using either a random or a linear initialization method. The process of

comparison and adjustment continues until the optimal number of iteration is reached or the specified error criteria are attained.

3) Extracting information from the trained SOM: Check all the minimum Euclidian distances and isolate the SOM's BMU for the depleted input vector (i.e. with missing values). The BMU identified in this step is a node of trained SOM and thus has the full complement of the missing values;
4) Replacement of missing values: Replace the missing values of the input depleted vector by their corresponding values in BMU identified in step 3 above.

2.5) Application of SOM

For the application of the SOM algorithm for infilling of missing data in this analysis, a SOM toolbox developed at Helsinki University of Technology Finland (www.cis.hut.fi/projects/somtoolbox/) was used in the Matlab® 2014b environment and a batch training algorithm was adopted. Due to the fact both datasets (rainfall and river discharge) had different time-steps, each of the datasets were trained separately. The data was presented in columns with each column representing measurements from each station. The entries without data were recorded as NaN (Not a Number) to meet Matlab® data entry requirements. To train all the data together in a single simulation, the data entries should overlap such that there is no single day/month for all the stations with no data entry.

The stations with the longest period of continuous missing observation were Katoa with 1418 consecutive days (01/04/1997-18/025/2001) approximately 4 years and Lai with 1200 consecutive days (31/01/1979-15/05/1982), approximately 3 years. Donomanga had the longest period of missing monthly rainfall observations.

3) Results and Discussion

Initial simulation results using discharge time series produced an average topographic error of 0.04 and a visual inspection of the time series was carried out to check the seasonal trends. Sporadic cases of numerical instability were noticed especially in portions of the time series with extensive gaps where infilling was done. In some cases, high flow values were observed in the dry season and low flow values observed in the rainy season. This was not logical as periods of high flows could not be followed by a single day of abrupt low flow and vice versa. These values were manually deleted for all the stations and a second simulation was performed using the same initial parameters. After this second simulation, these abnormalities disappeared and the average topographic error reduced to 0.02. Results of the overall performance of the model are shown in Table 1.

The results indicate that after the second simulation, the model was able to replicate with high accuracy the trends and flow magnitudes (high and low) in the respective seasons as shown in Figures 3 to 6. This justifies the low value of average topographic error 0.02 and the high values of R^2. From these results, the newly trained time series were used to infill missing gaps in the different time series in the Logone catchment. The preservation of topology, especially for discharge time series is important because seasonal variation causes high and low flows. The results obtained indicate that this seasonal variation was well preserved across all the gauging stations during the infilling process. In this research more emphasis was put on reducing the topographic error to ensure that the infilled values reflect the seasonal variation of the time series.

However, a visual observation of flow hydrographs (Figures 3-6) indicate that, the possibility of errors in the original river discharge time series may not be discounted especially for the Bongor station, and this may have a negative impact on the overall performance of the SOM algorithm in this study.

The results obtained for rainfall observations were similar to those obtained for discharge with the lowest R^2 value of 0.76 and average topographic error of 0.008. Although some authors (Kalteh & Berndtsson, 2007; Mwale et al., 2012) have proposed that to the rainfall time series should be trained together according to spatial location to improve the results, this method was not applied in this study because results obtained were judged to be satisfactory. Of the 18 rainfall stations, 10 had R^2 values of 0.90 and above while 7 stations had R^2 values of 0.80 and above with only one station (Donomanga) which has the highest percentage of missing observations having a value of 0.76. However, it was noticed that the performance of the model reflected the spatial location of the stations. For example, apart from Bongor CF with a R^2 of 0.80, all stations located above 10°N had R^2 values above 0.90 while most stations located below this latitude had R^2 values between 0.80-0.90. Since the graphs of the all the 18 rain gauge stations cannot be shown, (Figures 7 & 8) are used for illustration. Furthermore, it was observed that the SOM algorithm was able to preserve seasonal variation when infilling missing data in rainfall time series just as it did for discharge.

The results also indicate that, although this method is quite robust for infilling gaps in hydro-meteorological time series, it cannot be used for infilling gaps in time series with extended periods of missing observations as model performance starts diminishing. This is logical as in such situations the model does not have sufficient data to learn from, thus cannot correctly replicate the pattern in the data. For example time series of measured discharge at Katoa had 1200 consecutive days of missing observations, which represent 13% of the total data entries, produced an R^2 of 0.65 compared to Logone Gana with 97 consecutive days of missing observations with an R^2 of 0.91. This implies that time series with extended periods of missing observations should not be used as the model may infill the missing observations but still fail to replicate the pattern in the data. Although, as shown by Kalteh et al. (2007) and Mwale et al. (2012) this issue can be resolved for rainfall time series by training such time series with data from the same spatial zone, this cannot apply for discharge time series as it is influenced by other catchment characteristics and the river morphology which vary along the river channel.

Nevertheless results obtained suggest that SOMs are suitable for infilling gaps in hydro-meteorological time series in Sudano-Sahel catchments. Results obtained from this study are comparable to those obtained by Mwale et al. (2012, 2014) in the Lower Shire Floodplain in Malawi, Kang & Yusuf (2012) in the Kelantan and Damansara river basins in Malaysia and Kim et al. (2015) in the Taehwa watershed in Korea

The relationship between discharges measured at various stations along the Logone River is shown in Figure 9. The Unified distance matrix (U-matrix) is a graphical display used to illustrate the clustering of the reference vectors in the SOM, it shows the distance between neighbouring map units. The U-matrix can be seen as several component planes which are stacked together one on top of the other. Component planes can either be coloured or grey shaded in a two dimensional lattice. Light colours indicate areas in which the variables are close to each other in the input space, while dark colours illustrate large distances between variables in the input space. Dark colours can be seen as cluster separators while light colours are clusters themselves. Component planes are therefore, mostly used for visualizing the correlation between the various variables in the SOM since they can give information concerning the spread of values in each component (Gabrielsson & Gabrielsson, 2006).

From Figure 9, the relationship between the discharges measured at Bongor, Katoa and Logone Gana is not very discernible. To illustrate that there is no relationship between the discharge time series, Figure 10 shows that the discharges measured at Katoa and Logone Gana gauging stations, which are located downstream of Bongor, are paradoxically lower than discharge measured at Bongor station upstream. This can partly be explained by the fact that during the rainy season when the river overflows its banks, immediately after Bongor station, part of the flow is diverted to fill the Maga dam and part is lost to the floodplains. During the dry season, water is withdrawn from the river without control for various purposes by the inhabitants thus reducing the quantity that eventually reaches Logone Gana station located downstream. This can also be attributed to transmission losses as a result of infiltration to the aquifer through channel bed. Seeber (2013) observed that the discharge recorded at Ndjamena flow gauging station located downstream was lower than that recorded upstream at the Logone Gana station. Candela et al. (2014) reported that a significant proportion of groundwater in the Lake Chad aquifer system was from the Logone River through river and aquifer interactions.

4) Conclusion

The main objective of this study was to use Self-Organizing Maps (SOMs) to infill missing gaps in hydro-meteorological time series in the Logone catchment using data from four river discharge and 18 rain gauge stations riddled with gaps

The combination of artificial intelligence and human intelligence (to be able to distinguish the seasonal discharge trends, patterns and magnitudes) greatly improved the overall performance of the SOM algorithm in handling missing data. Other advantages of SOMs include: (i) it does not require input data from neighbouring stations; (ii) unlike other ANN methodologies it does not require extra datasets to train the time series; (iii) it is not computationally intensive; and (iv) it does not require extra storage capacity.

Results obtained from this study indicate that, the SOMs algorithm is quite robust for infilling gaps in hydro-meteorological time series, though it is not suitable for infilling gaps in time series with extended periods of missing observations as model performance starts diminishing. This methodology can be used by practitioners to enhance the planning and management of water resources in areas where available records are infested with missing observations. Preservation of topology through a good replication of trends and discharge magnitudes in the time series obtained in this study will reduce the data input uncertainty in our future modelling studies in the catchment.

Acknowledgements

This research was supported by a Commonwealth Scholarship award to the first author. We are grateful to SIEREM and the Lake Chad Basin Commission for providing the data used in this research.

References

Adeloye, A. J., Rustum, R., & Kariyama, I. D. (2012). Neural computing modelling of the reference crop evapotranspiration. Environmental Modelling & Software, 29, 61-63.

Alhoniemi, E., Himberg, J., Parhankangas, J. & Vesanto, J. (2002). SOM Toolbox - Online Documentation.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000). Artificial Neural Networks in Hydrology. II: Hydrologic Applications. Journal of Hydrologic Engineering, 5:2(124), 124-137.

Boyer, J., Dieulin C., Rouche, N., Cres, A., Servat, E, & Paturel, J. (2006). SIEREM: An Environmental Information System for Water Resources, Vol. IAHS Publication 308. IAHS Press: Wallingford, United Kingdom.

Campozano, L., Sanchez, E., Aviles, A., & Samaniego, E. (2012). Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the Ecuadorian Andes. Maskana, 5(1), 99-115.

Candela, L., Elorza, F. J., Tamoh, K., Jiménez-Martínez, J., & Aureli, A. (2014). Groundwater modelling with limited data sets: the Chari– Logone area (Lake Chad Basin, Chad). Hydrological Processes, 28, 3714-3727.

Coumou, D. & Rahmstorf, S. (2012). A decade of weather extremes. Nature Climate Change, 2, 491-496.

Daniel, E. B., Camp, J. V., LeBoeuf, E. J., Penrod, J. R., Dobbins, J. P. & Abkowitz, M. D. (2011). Watershed modelling and its applications: A state-of-the-art review. The Open Hydrology Journal, 5, 26-50.

Dastorani, M. T., Moghadamnia, A., Piri, J. & Ramirez, M. R. (2010). Application of ANN and ANFIS models for reconstructing missing flow data. Environmental Monitoring and Assessment, 166(1-4), 421-34.

Elshorbagy, A. A., Panu, U. S. & Simonovic, S. P. (2000). Group-based estimation of missing hydrological data: Approach and general methodology. Hydrological Sciences Journal, 45(6), 849-866.

Gabrielsson, S., & Gabrielsson, S. (2006). The use of Self-Organizing Maps in Recommender Systems: A survey of the Recommender Systems field and a presentation of a State of the Art Highly Interactive Visual Movie Recommender System. Master Thesis, Uppsala University.

Garcia, H., & Gonzalez, L. (2004). Self-organizing map and clustering for wastewater treatment monitoring. Engineering Applications of Artificial Intelligence. 17(3), 215–225.

Harvey, C. L., Dixon, H. & Hannaford, F. (2012). An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. Hydrology Research, 43(5) 618-636. DOI: 10.2166/nh.2012.110.

Ismail, S., Shabri, A., & Samsudin, R. A. (2012). Hybrid model of self-organizing maps and least square support vector machine for river flow forecasting. Hydrology and Earth System Sciences, 16, 4417-4433.

Kagoda, P. A., Ndiritu, J., Ntuli, C., & Mwaka, B. (2010). Application of radial basis function neural networks to short-term streamflow forecasting, Physics and Chemistry of the Earth, 35, 571-581.

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

Kalteh A. M., Hjorth, P., & Berndtsson R. (2008). Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling & Software, 23, 835, – 845.

Kalteh, A. M., & Berndtsson, R. (2007). Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP), Hydrological Sciences Journal, 2(2), 305-317.

Kang, H. M., & Yusof, F. (2012). Application of Self-Organizing Map (SOM) in Missing Daily Rainfall Data in Malaysia. International Journal of Computer Applications, 48(5).

Kashani, M. H., & Dinpashoh, Y. (2012). Evaluation of efficiency of different estimation methods for missing climatological data. Stochastic Environmental Research and Risk Assessment, 26, 59–71.

Kim, M., Baek, S., Ligaray, M., Pyo, J., Park, M., & Cho, K. H. (2015). Comparative Studies of Different Imputation Methods for Recovering Streamflow Observation. Water, 7, 6847–6860.

Kingston, G. B., Dandy, G. C. & Maier, H. R. (2008a). AI Techniques for Hydrological Modelling and Water Resources Management. Part 2 - Optimization, in L. N. Robinson (editor). Water Resources Research Progress, Nova Science Publishers, pp. 67-99.

Kingston, G. B., Dandy, G. C., & Maier, H. R. (2008b). AI Techniques for Hydrological Modelling and Water Resources Management. Part 1 - Simulation, in L. N. Robinson (editor). Water Resources Research Progress, Nova Science Publishers, pp. 15-65.

Kohonen, T. Self-Organizing Maps, Springer Series in Information Sciences, vol. 30, Springer, Heidelberg, 1st ed., 1995; 2nd ed., 1997.

Mwale, F. D., Adeloye A. J., & Rustum R. (2014). Application of self-organising maps and multi-layer perceptron-artificial neural networks for streamflow and water level forecasting in data-poor catchments: the case of the Lower Shire floodplain, Malawi. Hydrology Research, 45(6), 838-854.

Mwale, F. D., Adeloye, A. J., & Rustum R. (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, 50-52, 34-43.

Ng, W. W., & Panu, U. S. (2010). Infilling missing daily precipitation data at multiple sites using the multivariate truncated normal distribution model for weather generation. Water, 8 pp.

Presti, R. L., Barca, E. & Passarella, G. A. (2010). Methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Environmental Monitoring and Assessment, 160, 1-22.

Rustum, R., & Adeloye, A. J. (2007). Replacing Outliers and Missing Values from Activated Sludge Data Using Kohonen Self-Organizing Map. Journal of Environmental Engineering, 133(9), 909-916.

Seeber, K. (2013). Consultation of the Lake Chad Basin Commission on Groundwater Management. Project: Sustainable Management of the Lake Chad Basin, BGR No:05-2355.

Vesantu, J., Himberg, J., Alhoniemi, E & Parhankangas, J. (2000). SOM Toolbox for Matlab 5. Report A57. Helsinki University of Technology, Helsinki, Finland.

WWAP (United Nations World Water Assessment Programme). 2015. The United Nations World Water Development Report 2015: Water for a Sustainable World. Paris, UNESCO

Table 1: Station location, percentage of missing data, results of statistical evaluation and average topographic error.

Flow gauging	Latitude	Longitude	Time interval	Proportion of missing data (%)	R^2	Average topographic error
Lai	11.55	15.15	1973-1997	17.5	0.85	0.02
Bongor	10.83	15.08	1983-2007	19.2	0.8	
Katoa	10.27	15.42	1983-2007	26.8	0.65	
Logone Gana	9.40	16.30	1983-2007	6.45	0.91	
Rain gauge stations						
Ngaoundere	7.35	13.56	1950-2000	7.52	0.86	0.008
Baibokoum	7.73	15.68	1950-2000	8.82	0.88	
Bekao	7.92	16.07	1950-2000	5.88	0.90	
Pandzangue	8.10	15.82	1950-2000	14.2	0.81	
Donia	8.30	16.42	1950-2000	12.9	0.84	
Moundou	8.57	16.08	1950-2000	5.39	0.94	
Doba	8.65	16.85	1950-2000	4.08	0.94	
Delli	8.72	15.87	1950-2000	5.88	0.91	
Donomanga	9.23	16.92	1950-2000	16.2	0.76	
Guidari CF	9.27	16.67	1950-2000	12.3	0.85	
Goundi	9.37	17.37	1950-2000	6.05	0.91	
Kello	9.32	15.80	1950-2000	8.99	0.88	
Lai	9.40	16.30	1950-2000	5.23	0.92	
Bongor	10.27	15.40	1950-2000	10.8	0.80	
Yagoua	10.35	15.25	1950-2000	8.17	0.92	
Bousso	10.48	16.72	1950-2000	6.37	0.93	
Bailli	10.52	16.44	1950-2000	5.23	0.95	
Massenya	11.40	16.17	1950-2000	5.72	0.95	

Latitude and Longitude in degrees

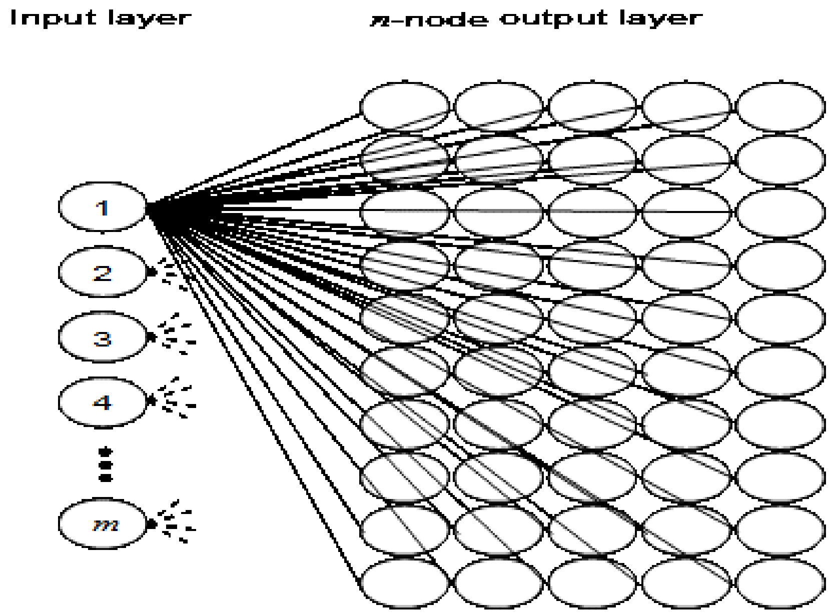

Figure 1: Architecture of an SOM (Adapted from Kagoda et al., 2010)

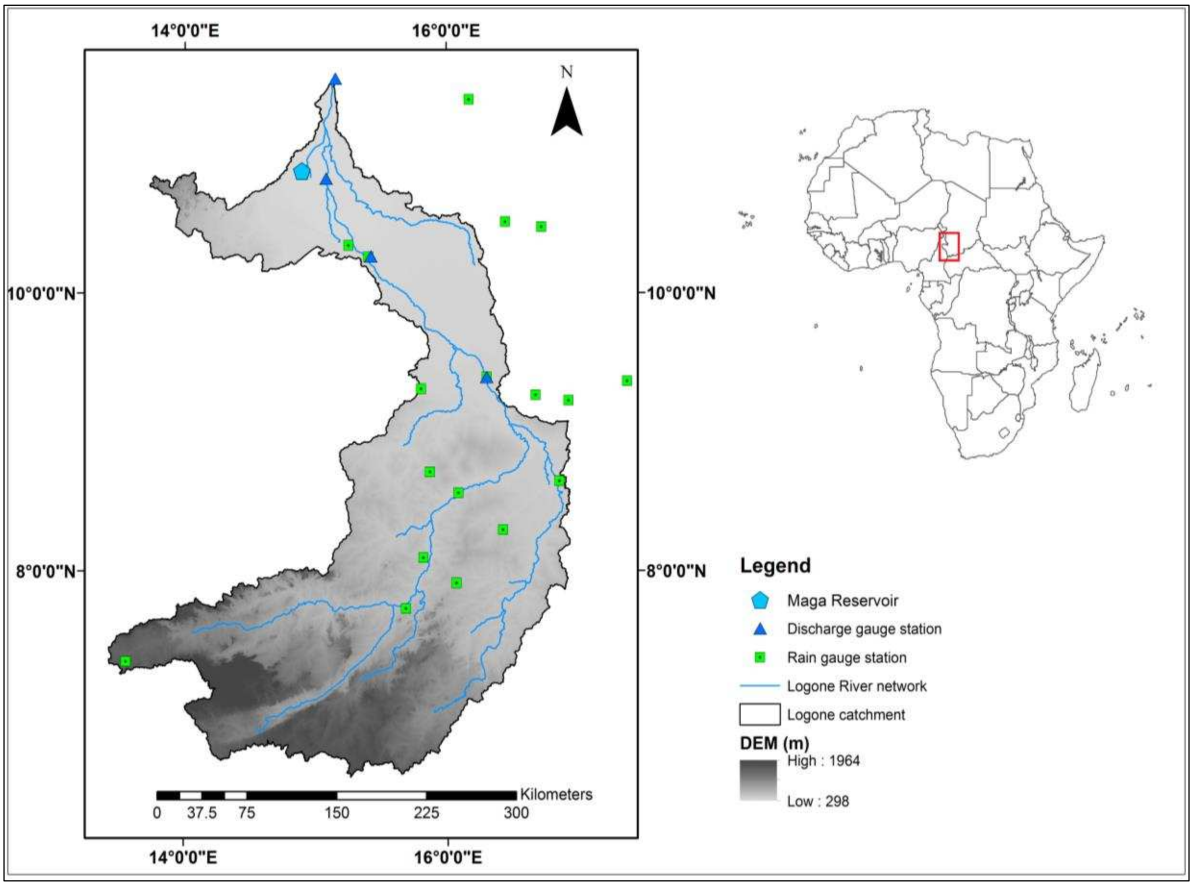

Figure 2: Map of study area showing rain and flow gauging stations

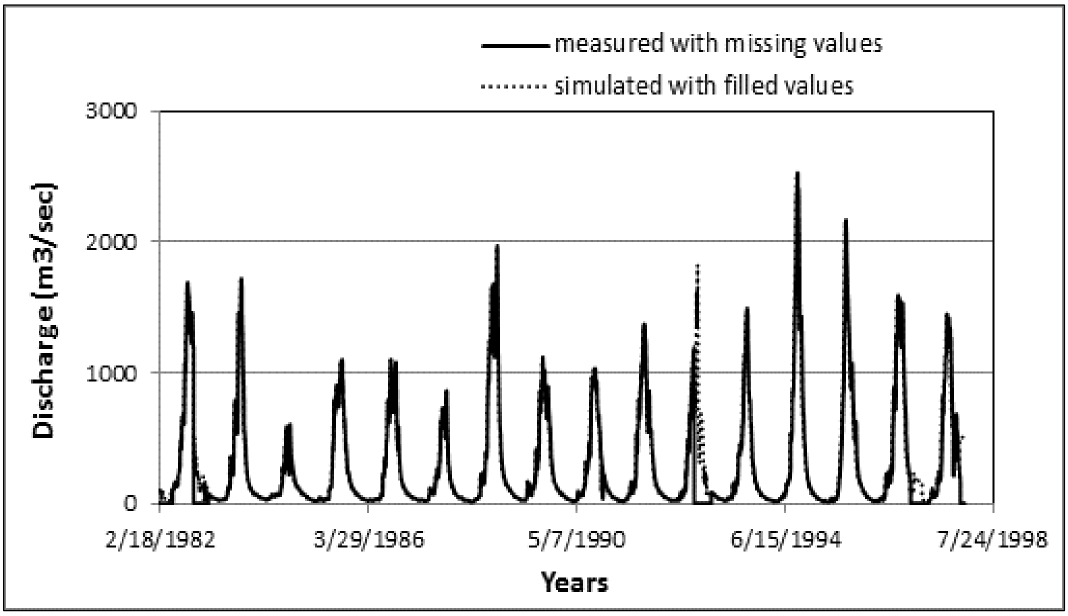

Figure 3: Observed and simulated discharge for Lai station 1973-1997

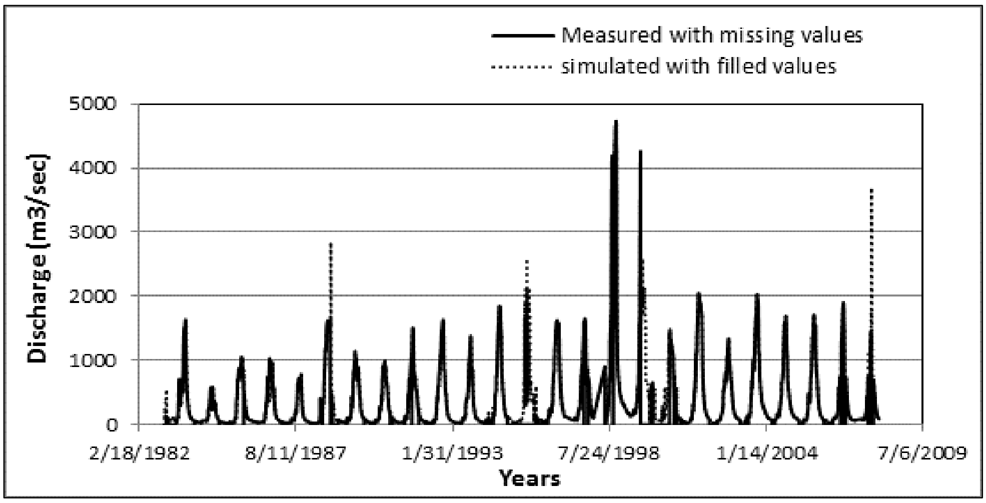

Figure 4: Observed and simulated discharge for Bongor station 1983-2007

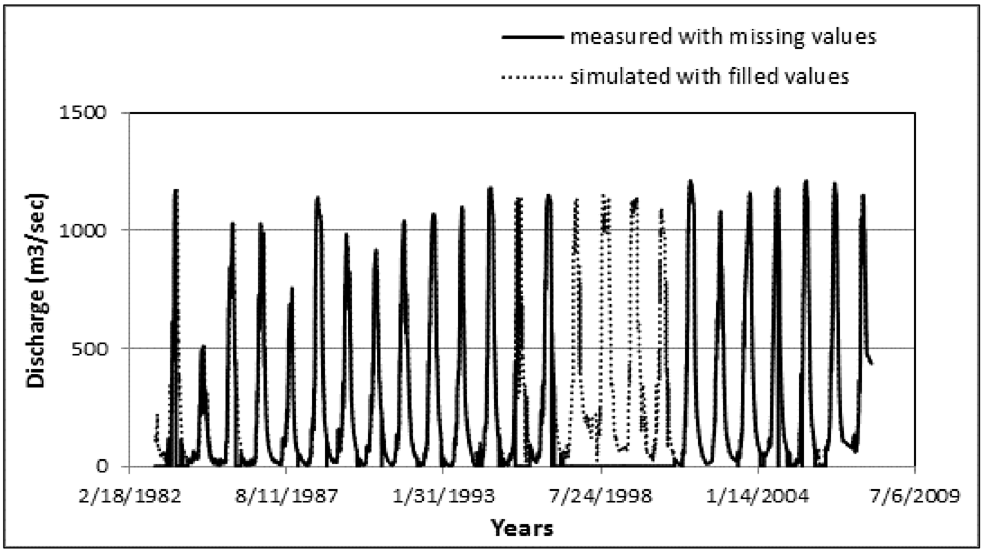

Figure 5: Observed and simulated discharge for Katoa station 1983-2007

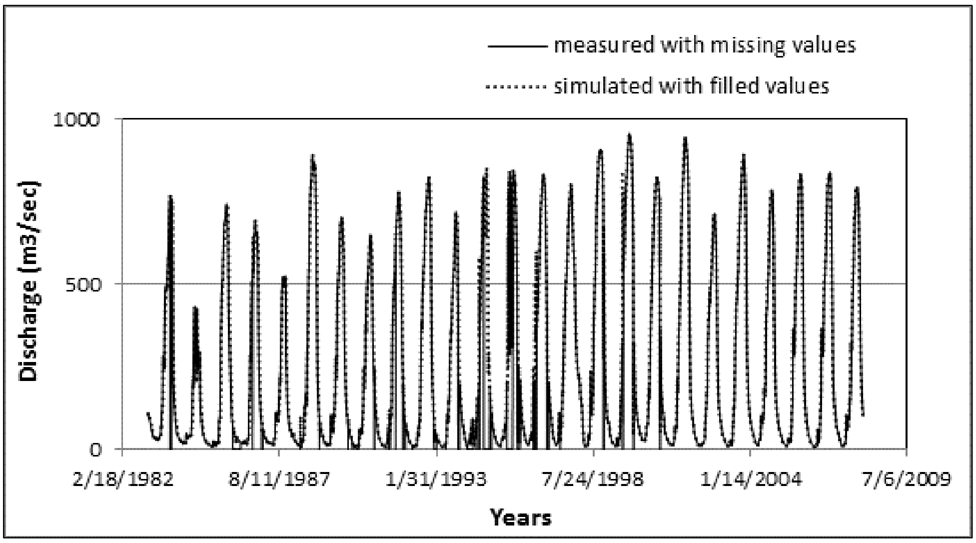

Figure 6: Observed and simulated discharge for Logone Gana station 1983-2007

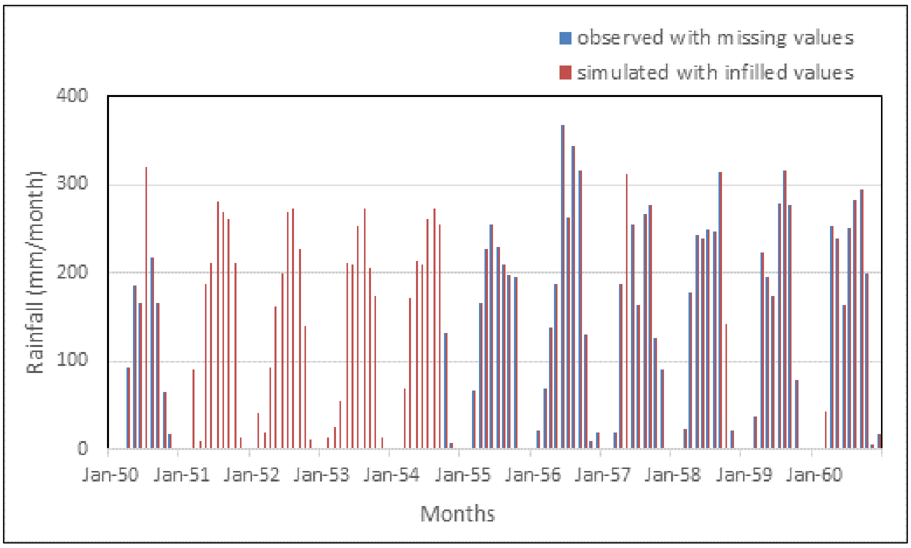

Figure 7: Observed and simulated rainfall for Ngaoundere (1950-1960)

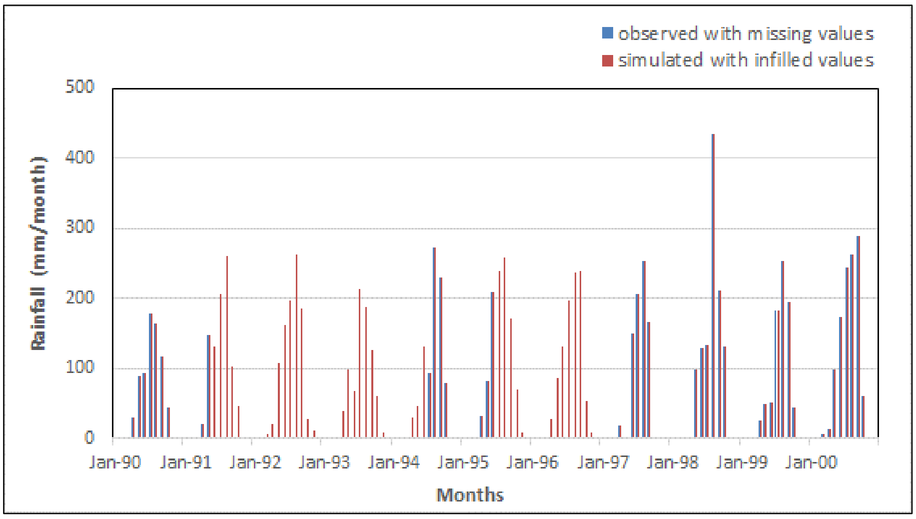

Figure 8: Observed and simulated rainfall for Kello (1990-2000)

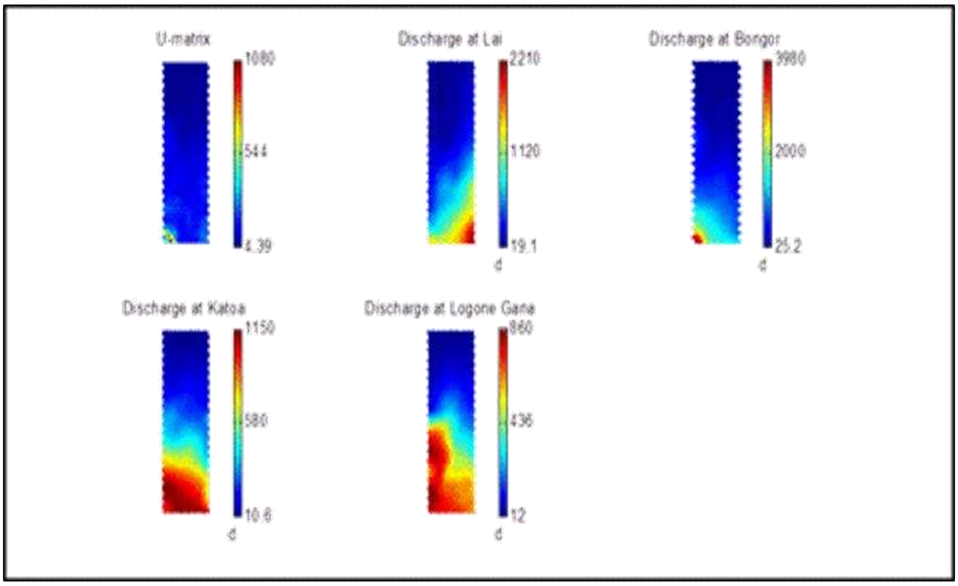

Figure 9: Component planes for discharge at all the stations

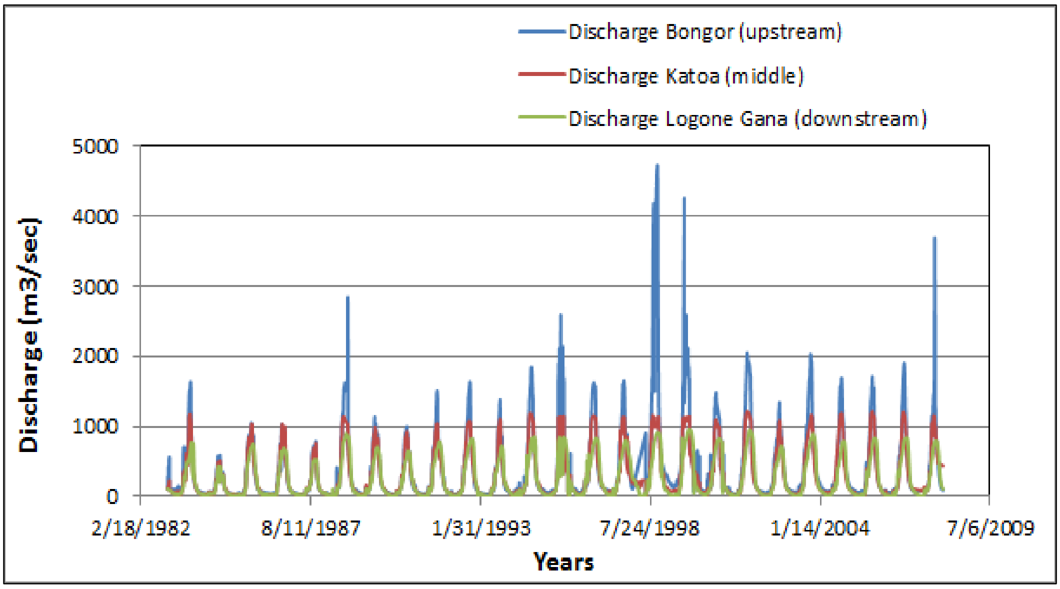

Figure 10: Discharge at Bongor, Katoa and Logone Gana 1983-2007