International Conference "Water, Environment, Energy and Society" (WEES–2009)
New Delhi, 12–16 January 2009

Uncertainty Quantification for Hydrologic Models Using Copula

Rajib Maity[1]
Department of Civil Engineering, Indian Institute of Technology
Powai, Mumbai - 400 076, INDIA
E-mail: rajib@civil.iitb.ac.in

D. Nagesh Kumar
Department of Civil Engineering, Indian Institute of Science
Bangalore - 560 012, INDIA
E-mail: nagesh@civil.iisc.ernet.in

ABSTRACT: A copula is a function that joins or 'couples' multivariate distribution function to their one dimensional marginal distribution functions. The strong theoretical background of copula draws the attention of researchers in recent years in many application fields, including hydrology and water resources. In this paper, a brief introduction of copula is presented. A new methodology is proposed for uncertainty quantification based on the theory of copula, which is applicable to any distributional form of hydrologic time series. Thus, basic requirement of normality, as in the case of many hydrologic models, can be relaxed which is very important in its own right. The proposed methodology is explained in the context of rainfall-runoff modeling. The proposed methodology is shown to capture and provide the information of uncertainty associated with prediction. The proposed methodology is shown to be promising and, being general, can be applied to any other modeling approach.

Keywords: Copula, Hydrologic Models, Uncertainty Quantification, Rainfall, Streamflow.

INTRODUCTION

In the recent years, copula is proven to be a very useful to many fields of application. It helps to analyze the multivariate events, which is of great interest to many applied statistical fields including hydrology and water resources. Application of copula in the field of water resources is still is in its nascent stage and, mostly limited to frequency analysis (Favre *et al.,* 2004; Salvadori and Michele, 2004; Grimaldi and Serinaldi, 2005; Zhang and Singh, 2006) and few others (Wang, 2001; Salvadori and Michele, 2006).

Assessment of dependence between two random variables is the key issues of many modeling approaches. Standard correlation is the widely used measure of dependence, which is applicable in the context of multivariate normal distribution and linear dependence. However, the nonlinear dependence, which is very common in many applications, and nonnormal behavior of time series create difficulties in the multivariate analysis. For example, distributional form of streamflow and many other hydrologic variables are far from Normal distribution. Simulation of a pair of random variables, with non-normal distribution, preserving their dependence structure is very cumbersome, if not impossible. However, copulas are able to couple the marginal densities of any form to produce their joint distribution, preserving their dependence structure.

In this paper, after presenting a brief overview of the theory of copula, a new copula-based methodology for uncertainty quantification is proposed. The methodology is elucidated with an example of typical conceptual rainfall-runoff model.

Rest of the paper is organized as follows. The theory of copula is briefly presented in section 2. Methodology for uncertainty quantification using the theory of copula is elaborated in section 3. The methodology is elucidated with an example of typical conceptual rainfall-runoff model in section 4 along with relevant discussion. Finally summary and conclusions are presented in section 5.

A BRIEF THEORY OF COPULA

A copula is a function that joins or couples multiple distribution functions to their one-dimensional marginal distribution functions (Nelsen, 2006). Let X and Y be a pair of Random Variables (RVs) with Cumulative Distribution Functions (CDF) as $F_X(x)$ and $G_Y(y)$ respectively. Also let their joint CDF be $H_{X,\,Y}(x,\,y)$. Each pair $(x,\,y)$ leads to a point in the unit square

[1] Conference speaker

[0, 1] × [0, 1] and this ordered pair in turn corresponds to a number, $H_{X,Y}(x, y)$, in [0, 1]. This correspondence is indeed a function. This function is known as Copula (Figure 1).

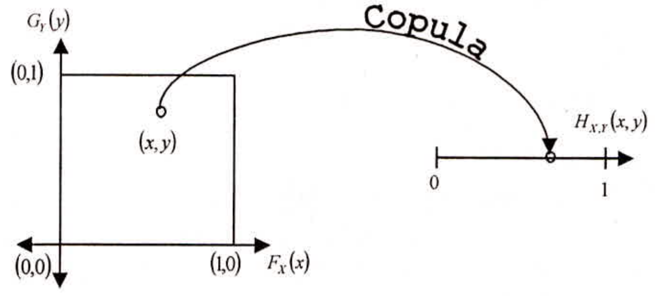

Fig. 1: Graphical representation of Copula

Application of copula to probability and statistics is achieved through the popular Sklar Theorem (Sklar, 1959), which states that, if $H_{X,Y}(x, y)$ is a joint distribution function with marginal distribution of X and Y as $F_X(x)$ and $G_Y(y)$ respectively, then there exists a copula $C(u, v)$ such that for all x, y in $\overline{R} \in (-\infty, \infty)$, $H_{X,Y}(x,y) = C[F_X(x),\ F_Y(y)]$. The most important point to be noted here is that the relationship is independent of the form of the marginal distribution, which is the reason behind the popularity of copula theory in many research areas.

Copula can be used in the study of dependence or association between random variables in terms of 'scale-invariant' or 'scale-free' measures. The widely used scale-free 'measures of association' for dependence structure are Kendall's tau (τ) and Spearman's rho (ρ_s). It may be noted that, Pearson's product moment correlation (ρ) is a 'measure of linear association' between random variables. It is obvious that, the estimate of ρ changes under nonlinear transformation of random variables. However, τ and ρ_s are scale-invariant and copulas are able to capture these scale-invariant properties of joint distribution which are invariant under strictly increasing transformation (Schweizer and Wolff, 1981).

If X and Y are two continuous RVs with copula C, population version of Kendall's tau (τ) can be expressed as,

$$\tau = 4\int C(u,v)\,dC(u,v) - 1 \qquad \ldots (1)$$

A complete discussion on copula can be found elsewhere (Genest and MacKay, 1986a; Genest and Rivest, 1993; Nelsen, 2006).

Archimedean Copula

Archimedean Copula, a particular class of copulas, is most popular to researchers due to its nice mathematical properties. A copula that can be expressed in terms of $C(u,v) = \varphi_\theta^{[-1]}(\varphi_\theta(u) + \varphi_\theta(v))$, is known as 'Archimedean copula' (Genest and MacKay, 1986a). $\varphi_\theta(\bullet)$ is known as generator of the copula and θ is the associated parameter. $\varphi_\theta^{[-1]}(\bullet)$ is the 'pseudo inverse' of $\varphi_\theta(\bullet)$ and defined as,

$$\varphi_\theta^{[-1]}(t) = \begin{cases} \varphi_\theta^{-1}(t), & 0 \le t \le \varphi(0) \\ 0, & \varphi(0) \le t \le \infty \end{cases} \qquad \ldots (2)$$

Basic properties of this class of copulas make them suitable for most of the research applications. It can be shown that if X and Y are two random variables whose joint distribution function is an Archimedean copula, C, generated by φ, equation 1 for Kendall's tau (τ) gets reduced to,

$$\tau = 1 + 4\int_0^1 \frac{\varphi_\theta(u)}{\varphi_\theta'(u)}\,du \qquad \ldots (3)$$

Example of Few Archimedean Copulas

Frank (*Frank*, 1979), Clayton (*Clayton*, 1978), Ali-Mikhail-Haq (AMH) and Gumble-Hougaard (GH) (*Gumbel*, 1960, *Hougaard*, 1986) are few examples of copulas belonging to the class of Archimedean copulas. Functional forms of these copulas are as follows:

Frank

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\left(1 + \frac{\left(e^{-\theta u} - 1\right)\left(e^{-\theta v} - 1\right)}{e^{-\theta} - 1}\right),$$

where $\varphi_\theta(t) = -\ln\left(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right)$ and $\theta \in (-\infty, \infty)$ excluding 0.

Clayton

$$C_\theta(u,v) = \left[\max\left(u^{-\theta} + v^{-\theta} - 1\,,\,0\right)\right]^{-1/\theta}$$

where $\varphi_\theta(t) = \frac{1}{\theta}\left(t^{-\theta} - 1\right)$ and $\theta \in [1, \infty)$ excluding 0.

Ali-Mikhail-Haq

$$C_\theta(u,v) = \frac{uv}{1 - \theta(1-u)(1-v)}$$

where $\varphi_\theta(t) = \ln\frac{1-\theta(1-t)}{t}$ and $\theta \in [-1,1)$

Gumbel-Hougaard

$$C_\theta(u,v) = \exp\left(-\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{1/\theta}\right)$$

where $\varphi_\theta(t) = (-\ln t)^\theta$ and $\theta \in [-1,\infty)$

UNCERTAINTY QUANTIFICATION USING COPULA

A methodology is proposed to quantify the uncertainty associated with predicted values of a hydrological model. However, the uncertainty is associated with both the observed and predicted values of hydrologic variables, which are considered to be random variables.

The methodology is explained in the context of rainfall-runoff model. Steps involved to quantify the uncertainty associated with predicted streamflows are as follows:

1. Distributional form of streamflow is identified by standard statistical methods. This can be achieved by fitting different probability distributions to streamflow data and selecting the most appropriate one.
2. Observed and predicted streamflows are pair wise transformed through their Cumulative Distribution Function (CDF), which is identified in step 1. Thus a pair of uniformly distributed random variables over [0, 1] is obtained.
3. Association between the observed and predicted streamflows is estimated in terms of Kendall's tao. Let (x_1,y_1), (x_2,y_2), … , (x_n,y_n) be the paired samples of two random variables, X and Y. Two pairs (x_i,y_i) and (x_j,y_j) are known to be concordant if $(x_i-x_j)(y_i-y_j)>0$ and discordant if $(x_i-x_j)(y_i-y_j)<0$. Sample estimate of Kendall's tau is obtained as the difference between the probability of concordance and the probability of discordance. Out of n paired samples, there are ${}^nC_2$ different ways of selecting two pairs. If there are c number of concordant pairs and d number of discordant pairs, sample estimate of Kendall's tau is expressed as,

$$\begin{aligned}\hat{\tau} &= P\left[(x_i-x_j)(y_i-y_j)>0\right] \\ &\quad - P\left[(x_i-x_j)(y_i-y_j)<0\right] \qquad \ldots (4) \\ &= \frac{c}{{}^nC_2} - \frac{d}{{}^nC_2} = \frac{c-d}{{}^nC_2}\end{aligned}$$

4. Parameter θ, associate with an Archimedean copula, is estimated by replacing the population version of Kendall's tao with its sample estimate ($\hat{\tau}$), in equation 3. Thus, transformed streamflows can be simulated through this Archimedean copula preserving their dependence structure. Such simulation can be achieved by the algorithm as explained below (Genest and MacKay, 1986b).
   (a) Functional form of $\varphi^{[-1]}(\bullet)$, $\varphi'(\bullet)$ and $\varphi'^{(-1)}(\bullet)$ are obtained where $\varphi(\bullet)$ is the generator function of the Archimedean copula after replacing the value of θ and hence, subscript θ is omitted hereafter. Equation 2 is used to obtain $\varphi^{[-1]}(\bullet)$. Same can be used for calculating $\varphi'^{(-1)}(\bullet)$ after obtaining $\varphi'(\bullet)$ which is the derivative of $\varphi(\bullet)$ with respect to $\bullet$.
   (b) Two independent uniformly distributed $[\sim Un(0,1)]$ random variates U and T are generated.
   (c) Two new variables, S and W, are obtained as $S = \varphi'(U)/T$ and $W = \varphi'^{(-1)}(S)$.
   (d) Another variable, V, is obtained as $V = \varphi^{[-1]}\left[\varphi(W) - \varphi(U)\right]$. The pair U and V are the simulated pair preserving the dependence structure.
   (e) U and V are then back transformed by their inverse cumulative distribution function to generate simulated observed and predicted streamflows in original scale.
5. Steps (4a) through (4b) are repeated for different Archimedean copulas. Genest and Rivest (1993) described a procedure to select the most appropriate copula, which is also applied in hydrology (Zhang and Singh, 2006) and is followed in this study also. Steps involved to select the most appropriate copula are as follows:
   (a) For a particular Archimedean copula C with generator function φ, a parametric function $K(z)$ is defined as $K(z) = z - \frac{\varphi(z)}{\varphi'(z^+)}$. $K(z)$ is the distribution function of random variable, $C(U, V)$ where u and v are the uniformly (0, 1) distributed RVs (Nelsen, 2006).
   (b) A nonparametric estimate of above function, $K_n(z)$ is obtained as the proportion of $z_i < z$, , where z_i is,

$$z_i = \frac{\text{Number of } (x_j, y_j) \text{ such that } x_j < x_i \text{ and } y_j < y_i}{(N-1)}$$

(c) A scatter plot between $K(z)$ and $K_n(z)$ is prepared.

(d) Steps (a) to (c) are repeated for all the copulas considered. The better the fit, the closer will be the corresponding scatter to a 45° line through origin.

(e) Most appropriate copula may also be found out by calculating the Sum of Square Errors (SSE) from the 45° line through origin for all copulas. The copula with the smallest SSE is the most appropriate one.

6. Joint distribution function is obtained analytically after selecting the best copula. However, in many cases of hydrologic applications, analytical solution may not always be feasible. In such cases, large number of numerically simulated jointly distributed values are used, which is carried out in this study also.
7. Keeping any predicted streamflow value at the centre, a sufficiently 'small' window around it, is selected. Statistical property of the observed streamflow values, lying within this window, is expressed through box plot. The median of these values is used as a prediction. The interquartile range (75th percentile – 25th percentile) of these values indicates the associated uncertainty. This is equivalent to the conditional distribution in an analytical sense.

It may be noted that the association between the observed and predicted time series is simulated preserving their scale free measure of association with each other. Since the predicted time series is associated with all possible sources of uncertainty, the copula-based methodology for uncertainty quantification integrates the uncertainties from all the possible sources. This is a valuable addition towards the uncertainty quantification of hydrologic models.

Another important point is that, the choice of window size, mentioned above, is a subjective choice. A "too large" window will conceal the existing nonlinearities whereas a "too small" window may suffer from insufficient data points to represent the statistical properties. An acceptable remedy is to simulate a sufficiently 'large' number of data points and minimize the window size as much as possible.

RESULTS AND DISCUSSIONS OF A CASE STUDY

The methodology explained above is applied to a subbasin of Mahanadi River located in the state of Chattisgarh in India (Figure 2). The watershed is having an area of approximately 46000 sq. km. with an undulating topography. Mahanadi River is an ephemeral

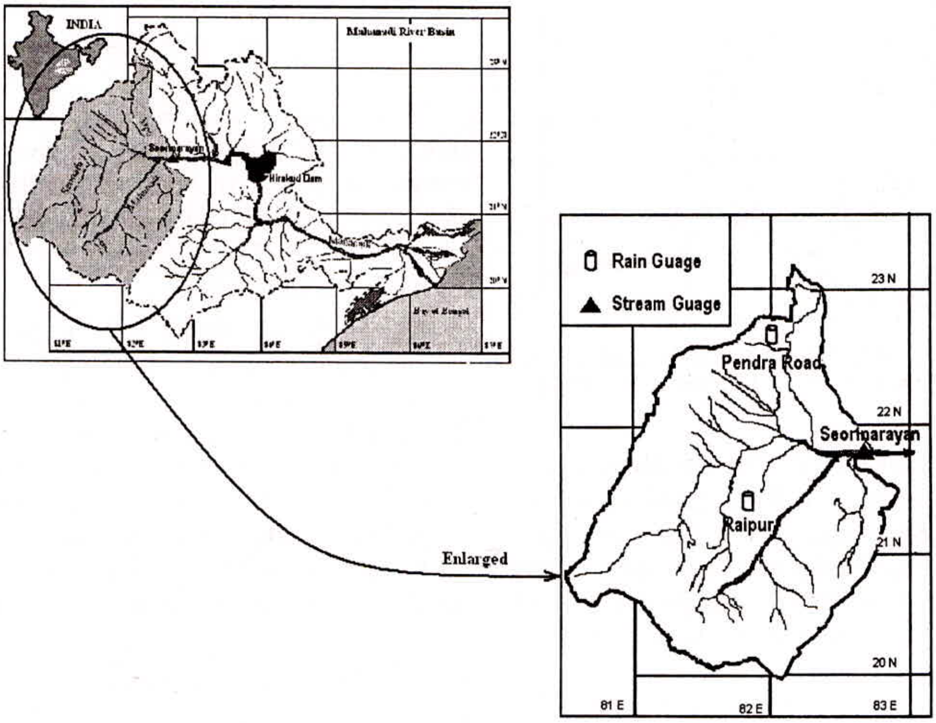

Fig. 2: Location map of Watershed

river and rainfall during the monsoon season (June through September) mostly causes the streamflows. As a consequence major floods occur during monsoon season only. For the rest of the year, streamflows are mostly zero (Patri, 1993). So, only the monsoon period is considered in this study.

Daily streamflow at the outlet of the watershed (Seorinarayan) and daily rainfall data from two rain gauge stations (Raipur and Pendraroad) in the upstream catchment (refer Figure 2) are obtained from the office of Executive Engineer, Mahanadi Division, Central Water Commission (CWC), India.

Considering the streamflow at Seorinarayan and weighted average rainfall at Raipur and Pendraroad, a conceptual rainfall-runoff model is applied. The conceptual rainfall-runoff model is not explained in this paper. However, the results during calibration and testing period are shown in Figures 3 and 4 respectively. It is observed that, during testing period, predicted values are highly associated with the observed values having a correlation coefficient (ρ) of 0.90 (linear dependence) and Kendall's tao (τ) of 0.76 (scale free measure of association). Moreover, the model successfully captures the low flows as well as high flows.

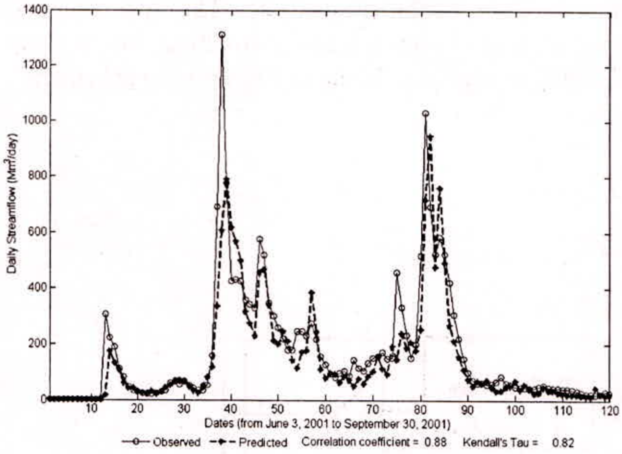

Fig. 3: Results of a conceptual rainfall-runoff model during calibration period (June 3, 2001 to September 30, 2001)

In a view to get at uncertainty quantification, distributional form of streamflows is investigated. Observed streamflows are plotted against corresponding non-exceedence probabilities and cumulative distribution functions for different probability densities are fitted (Figure 5). Daily streamflow is found to be best fitted with gamma distribution as $f_X(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$, with parameters $\alpha = 0.545$ and $\beta = 314$.

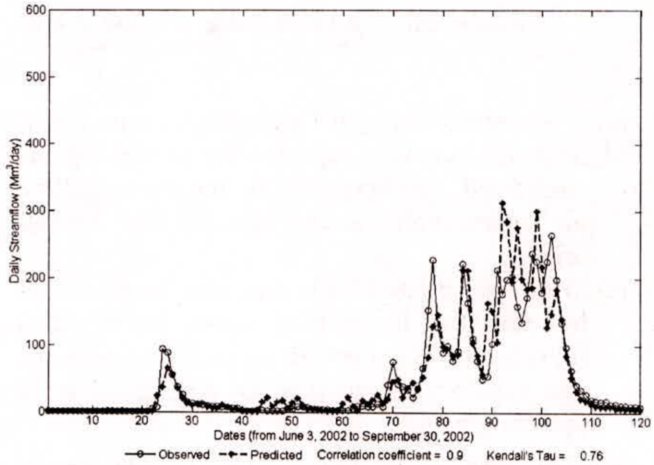

Fig. 4: Results of a conceptual rainfall-runoff model during testing period (June 3, 2002 to September 30, 2002)

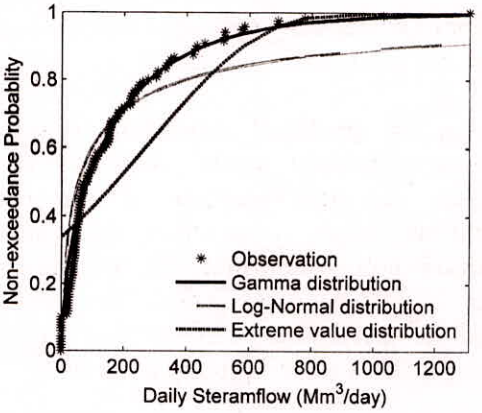

Fig. 5: Daily streamflow fitted with different probability densities

Association between observed and predicted streamflows is simulated with four different Archi-medean copulas namely: (i) Frank, (ii) Clayton, (iii) Ali-Mikhail-Haq (AMH) and (iv) Gumbel-Hougaard (GH) copulas (Figure 6). Following specific obser-vations are made form Figure 6.

(a) AMH copula is found to be worst case to capture the association between observed and predicted streamflows as compared to other cases.

(b) Though the GH copula seems to be well fitted, a critical observation is that the errors are more or less same for low as well as high streamflows which is quite unlikely in reality. The error should be less for low streamflows and high for high streamflows.

(c) Both Frank and Clayton copulas are found to be equally well.

However, to select the better among Frank and Clayton copulas, the procedure explained in step 5 of methodology is followed. Scatter plots between $K(z)$ and $K_n(z)$ are prepared for Frank and Clayton copulas

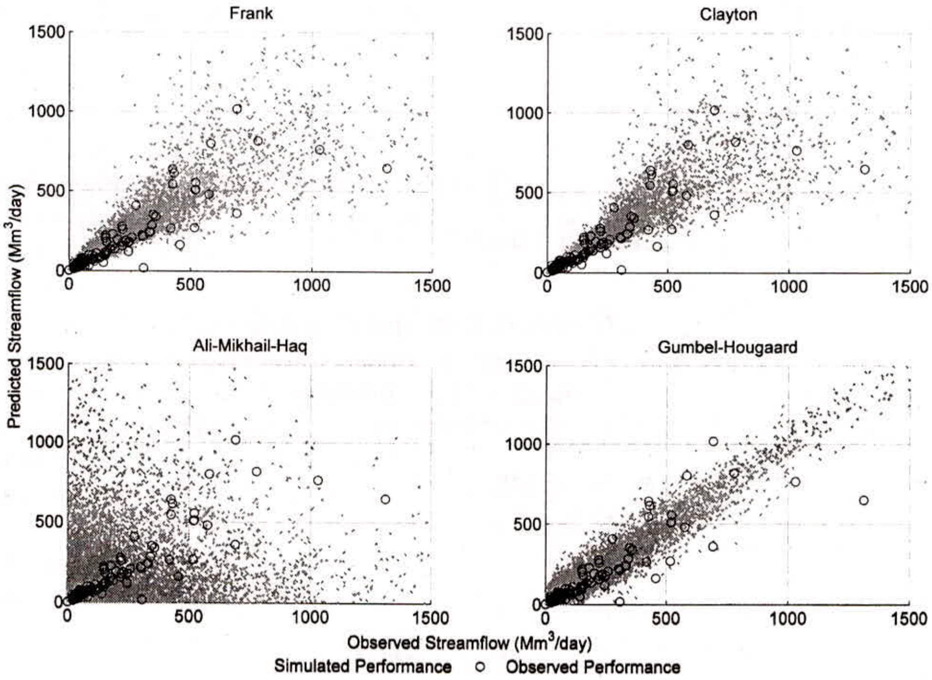

Fig. 6: Simulation of association between observed and predicted streamflows using different copulas

(Figure 7) and the corresponding SSE are displayed within parentheses. It is found that Frank copula is performing better than Clayton and thus selected. Numerically simulated values with Frank copula are used to prepare the box plots showing the uncertainty associated with the predicted values as described in step 7 of the methodology.

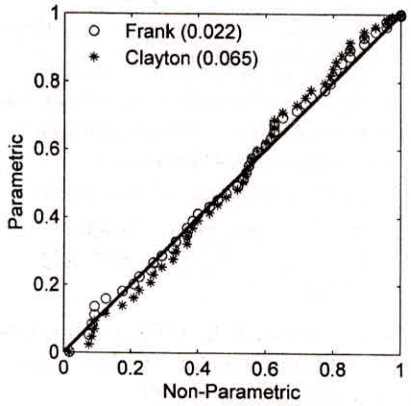

Fig. 7: Scatter plot between parametric, $K(z)$ and non-parametric, $K_n(z)$ for Frank and Clayton copulas with corresponding SSE displayed within parentheses

Model Performance

The methodology is tested for the monsoon season of 2002, i.e., June, 2002 to September, 2002. A comparison plot between observed and predicted streamflows is shown along with box plots, showing the information regarding uncertainty associated with the predicted values (Figure 8). It is observed that observed streamflows are mostly lying within the interquartile range of the predicted uncertainty, except for few cases, which indicates the successful capture of uncertainty. It is also observed that, the higher the streamflows the higher is the associated uncertainty, which is obvious and reflects the factual position.

Thus, the proposed methodology for uncertainty quantification is an useful technique that aggregates all the possible sources of uncertainty, associated with the predicted values as discussed earlier. This methodology, being general, can be applied to any other modeling approach too.

SUMMARY AND CONCLUSIONS

In this paper a new methodology is proposed for uncertainty quantification associated with the prediction of hydrologic variables using copulas. The methodology is elucidated in the context of a conceptual rainfall-runoff model even though the conceptual rainfall-runoff model is not in this paper. Rather the performance of the model is explained for streamflow prediction at daily time-scale and is shown to be able to capture a wide range of possible streamflows, including very low flows as well as very high flows. It is observed that the predicted daily streamflows are highly associated with the observed streamflows with correlation coefficient (ρ) of 0.90 (linear dependence) and Kendall's tao (τ) of 0.75 (scale free measure of association).

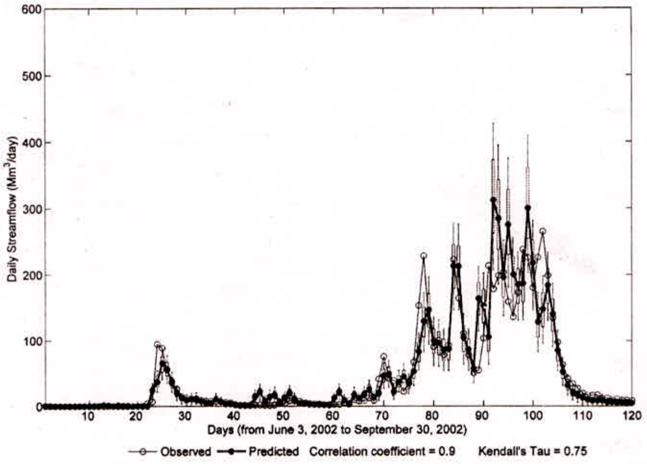

Fig. 8: Observed and predicted streamflows along with box plots, showing the information of uncertainty associated with the predicted values

Proposed methodology for uncertainty quantification using copulas is a valuable addition to the field of hydrology and water resources. Apart from its simplicity, the approach, being general, can be applied to any prediction model. Two major advantages of the proposed methodology, apart from the usual benefits of copula, are as follows:

1. The predicted time series is associated with all possible sources of uncertainty. Being the fact that the association between observed and predicted time series is simulated, preserving the association between them, the proposed copula-based methodology integrates the uncertainties from all the possible sources. This is a valuable addition towards the uncertainty quantification in hydrologic models.
2. Distributional form of hydrologic variables may vary over a wide range of distributions. Exploiting the capability of copula, the methodology can be applied to any hydrologic variables irrespective of its distributional form.

Association, in terms of correlation coefficient (linear dependence) and Kendall's tao (scale free measure of association), between observed and predicted hydrologic variables for other hydrologic models may vary over a wide range. Copula being capable of simulating a wide range of dependence (Favre *et al.,* 2004), the proposed methodology is applicable irrespective of the strength of association. However, lesser strength of association will lead to vaguer information of uncertainty.

REFERENCES

Clayton, D.G. (1978). "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence", *Biometrika*, 65, 141– 151.

Favre, A.-C., El Adlouni, S., Perreault, L., Thie´monge, N. and Bobe´e, B. (2004). "Multivariate hydrological frequency analysis using copulas", *Water Resour. Res.*, 40, W01101, doi:10.1029/2003WR002456.

Frank, M.J. (1979). "On the simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$", *Aequationes Math.*, 19, 194–226.

Genest, C. and MacKay, J. (1986a). "The joy of copulas: Bivariate distributions with uniform marginals', *The American Statistician*, 40(4), 280–283.

Genest, C. and MacKay, J. (1986b). "Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données Copules", *The Canadian Journal of Statistics*, 14(2), 145–159.

Genest, C. and Rivest, L.-P. (1993). "Statistical Inference Procedures for Bivariate Archimedean Copulas", *Journal of the American Statistical Association*, 88(423), 1034–1043.

Grimaldi, S. and Serinaldi, F. (2005). "Asymmetric copula in multivariate flood frequency analysis", *Advances in Water Resources*, In press.

Gumbel, E.J. (1960). "Bivariate exponential distributions", *J. Am. Stat. Assoc.*, 55, 698–707.

Hougaard, P. (1986). "A class of multivariate failure time distributions", *Biometrika*, 73, 671–678.

Nelsen, R.B. (2006). *An introduction to copulas*, Lecture Notes in Statistics, Springer, New York.

Patri, S. (1993). *Data on flood control operation of Hirakud dam*, Report, Dept. of Irrigation, Gov. of Orissa, India.

Salvadori, G. and De Michele, C. (2004). "Frequency analysis via copulas: Theoretical aspects and applications to hydrological events", *Water Resour. Res.*, 40, W12511, doi:10.1029/2004WR003133.

Salvadori, G. and De Michele, C. (2006). "Statistical characterization of temporal structure of storms", *Advances in Water Resources*, 29, 827–842, doi:10.1016/j.advwatres.2005.07.013.

Schweizer, B. and Wolff, E.E. (1981). "On nonparametric measures of dependence for random variables", *The Annals of Statistics*, 9(4), 879–885.

Sklar, A. (1959). "Fonctions de répartition à n dimensions et leurs marges", *Publ. Inst. Stat. Univ. Paris*, 8, 229–231.

Wang Q.J. (2001). "A Bayesian joint probability approach for flood record augmentation", *Water Resour. Res.*, 37(6), 1707–1712.

Zhang, L. and Singh, V.P. (2006). "Bivariate flood frequency analysis using the copula method", *Journal of Hydrologic Engineering*, 11(2), 150–164, doi: 10.1061/(ASCE)1084–0699(2006)11:2(150).