

# Towards improved post-processing of hydrologic forecast ensembles

Shahrbanou Madadgar,<sup>1</sup> Hamid Moradkhani<sup>1\*</sup> and David Garen<sup>2</sup>

<sup>1</sup> Department of Civil and Environmental Engineering, Portland State University, Portland, Oregon, USA

<sup>2</sup> USDA, Natural Resources Conservation Service, NWCC, Portland, Oregon, USA

## Abstract:

Forecast ensembles of hydrological and hydrometeorological variables are prone to various uncertainties arising from climatology, model structure and parameters, and initial conditions at the forecast date. Post-processing methods are usually applied to adjust the mean and variance of the ensemble without any knowledge about the uncertainty sources. This study initially addresses the drawbacks of a commonly used statistical technique, quantile mapping (QM), in bias correction of hydrologic forecasts. Then, an auxiliary variable, the failure index ( $\gamma$ ), is proposed to estimate the ineffectiveness of the post-processing method based on the agreement of adjusted forecasts with corresponding observations during an analysis period prior to the forecast date. An alternative post-processor based on *copula* functions is then introduced such that marginal distributions of observations and model simulations are combined to create a multivariate joint distribution. A set of 2500 hypothetical forecast ensembles with parametric marginal distributions of simulated and observed variables are post-processed with both QM and the proposed multivariate post-processor. Deterministic forecast skills show that the proposed copula-based post-processing is more effective than the QM method in improving the forecasts. It is found that the performance of QM is highly correlated with the failure index, unlike the multivariate post-processor. In probabilistic metrics, the proposed multivariate post-processor generally outperforms QM. Further evaluation of techniques is conducted for river flow forecast of Sprague River basin in southern Oregon. Results show that the multivariate post-processor performs better than the QM technique; it reduces the ensemble spread and is a more reliable approach for improving the forecast. Copyright © 2012 John Wiley & Sons, Ltd.

KEY WORDS ensemble streamflow prediction; copula; quantile mapping; bias correction

Received 22 February 2012; Accepted 29 August 2012

## INTRODUCTION

### Background

Seasonal water supply outlooks, or volume of total seasonal runoff, are routinely used by water managers and decision makers for making commitments for water deliveries, determining industrial and agricultural water allocations, and operating reservoirs for multiple uses such as hydropower and flood control. These forecasts can either be seasonal volumes based on statistical regression equations or ensembles of hydrographs produced by hydrologic models, the so-called ensemble streamflow prediction (ESP). In ESP (Twedt *et al.*, 1977; and Day, 1985), the hydrologic model is driven for historical time period until the initial condition (IC) of the basin is obtained. The hydrologic model is then run with the resampled historical meteorological data to generate an ensemble of possible streamflow forecasts for the future time periods. In ESP, future meteorological condition is assumed to mimic the historical meteorology but with an uncertainty of not knowing exactly what historical condition may happen in the future; therefore, creating an ensemble of

hydrologic forecasts reflect the uncertainty in the forcing data. Moreover, hydrologic models and consequently ensemble forecasts are prone to several additional sources of uncertainty making the ensemble forecasts probably biased or over/under dispersed.

Several techniques have been employed to incorporate all sources of uncertainty in ESP. Recent studies have examined or developed techniques that have potential to provide skillful predictions of seasonal runoff volume using statistical methods by means of optimal predictor selection (e.g. Moradkhani and Meier, 2010), using ESP combined with data assimilation (Dechant and Moradkhani, 2011) or using weighted ESP traces according to climate signals (Najafi *et al.*, 2012). An ensemble of forecast trajectories is generally generated to capture total forecast uncertainty due to several sources of uncertainty including atmospheric forcing, IC, model structure, and parameters (Moradkhani and Sorooshian, 2008; Olsson and Lindström, 2008; Wood and Lettenmaier, 2008; DeChant and Moradkhani, 2011; Moradkhani *et al.*, in review; Parrish *et al.*, 2012). In generating the ensemble of forecasts, different methodologies may be employed. In hydrologic applications with the lack of knowledge about future climate conditions, the sampling of historical meteorological data can provide a range of possible future climate condition used for generating the ensemble hydrologic forecasts (McEnery *et al.*, 2005; Wood and Lettenmaier, 2008). To

\*Correspondence to: Hamid Moradkhani, Associate Professor, Department of Civil and Environmental Engineering, Portland State University, Portland, Oregon, USA.  
E-mail: hamidm@cecs.pdx.edu

improve forecast skill, some studies generate ESPs from meteorological forecast ensembles made by numerical weather prediction models (Clark and Hay, 2004; Roulin and Vannitsem, 2005; Thirel *et al.*, 2008; Li *et al.*, 2009; Najafi *et al.*, 2012). To address the uncertainty of ICs at the forecast date, Wood and Lettenmaier (2008) proposed reverse-ESP approach, while resampled historical climatology was used as the forcing of forecast horizon. Their results indicated that the impact of uncertain ICs on the forecast quality is related to the forecast date, lead time, and the area of study. In a recent study, DeChant and Moradkhani (2011) employed the data assimilation method as a flexible and statistically defensible procedure to quantify the IC uncertainty by obtaining the probability distribution function (PDF) of state variables at the time of forecast and then used those for generating ESPs. In meteorological forecasts, however, some recent studies suggest generating the ensemble of precipitation or temperature forecasts from the single-valued forecasts (Schaaake *et al.*, 2007). Using bivariate meta-Gaussian distribution (Kelly and Krzysztofowicz, 1997) to join the model output (forecast;  $Y$ ) and observations ( $X$ ), the forecast ensemble is generated from the conditional distribution of observed values given the single-valued forecast ( $f_{Y|X}(y|x)$ ) at each time step within the forecast horizon. Assigning bivariate Gaussian distribution to the forecast and observed variables necessitates the transformation of non-normal variables ( $X$  and  $Y$ ) to the standard normal variables ( $U$  and  $V$ ) using their marginal distributions ( $F_X(x)$ ,  $F_Y(y)$ ). Brown and Seo (2010), however, argued that back and forth transformation from the Gaussian space can invalidate the optimality of estimated parameters of the conditional probability distribution.

### Post-processing

In spite of the efforts to incorporate all sources of uncertainty into the forecast, and regardless of the methodologies applied to generate the forecast ensembles, they are still subject to errors and systematic biases. This means that post-processing is still necessary to ensure that ensemble forecasts are unbiased and have the proper dispersion. Several techniques have been tried to accomplish this, which are reviewed below. In an initial study to remove the forecast errors, Smith *et al.* (1992) assumed constant errors multiplied by the monthly simulations generated from a particular forcing regardless of the ICs at the forecast date. The multiplied error was estimated by historical simulations and observations. Another post-processing method applied to forecast ensembles is the quantile mapping (QM) technique (Hashino *et al.*, 2006; Wood and Lettenmaier, 2006; Biagorria *et al.*, 2007; Piani *et al.*, 2010; among others). With this method, a mapping between observation and simulation cumulative distribution functions (CDFs) is used. The observation and simulation CDFs may be estimated by either empirical CDFs or parametric distributions fitted to historical data (Ines and Hansen, 2006; Piani *et al.*, 2010). A major drawback of this

method, however, is that it does not maintain the pairing of corresponding simulated and observed flows. To restrict the shortcoming of QM technique, Madadgar and Moradkhani (2011) generated several ESPs for various analysis periods prior to the forecast period. Several simulation CDFs were produced for the simulations associated with each historical forcing implemented on the analysis periods, which are then used for bias correction of the forecast trajectory corresponding to that particular forcing. Bias correction of forecasts with particular CDFs produced specifically for each forcing data reduces the forcing uncertainty of QM method. In a recent study, Candille *et al.* (2010) applied a bias correction method with the so called 'on the fly' scheme (Cui *et al.*, 2008) updating and correcting the ensemble bias over time. In their study, the multi ensemble, from the so-called North American Ensemble Forecast System comprising National Centres for Environmental Prediction and Meteorological Service of Canada ensembles, are bias corrected through individual on-the-fly analysis scheme for each model of ESP. Only the variables with normally distributed errors like temperature and wind vector components were bias corrected, and the results indicated that the reliability of forecasts was improved when on-the-fly method was in use. In another study, Djalalova *et al.* (2010) used the Kalman-Filter (KF) method (Brookner, 1998) to estimate the bias from air quality forecasts. They used the data from the 7 past days to implement the KF bias correction, and their results confirmed the advantage of employed bias correction methods in improving forecast skill.

Bias correction of forecasts is mathematically equivalent to estimating the conditional probability distribution of the observed variable given the real-time forecast. To approximate the conditional probability of the observation given a forecast, their joint probability should first be estimated. The bivariate normal distribution is usually applied to join the forecast and observed variables during an analysis period (Schaaake *et al.*, 2007; Zhao *et al.*, 2011). Zhao *et al.* (2011) assumed Gaussian marginal distributions of both simulations and observations, which further required transformation of raw variables to the standard normal variates. They used the normal quantile transformation proposed by Krzysztofowicz (1997) for variable transformation. Todini (2008), introduced the Model Conditional Processor where the multivariate normal distribution was used to estimate the predictive uncertainty of forecast variable conditioned on the predictions of several models. Similar to Zhao *et al.* (2011), the uncertainty processor proposed by Todini (2008) required the marginal distributions to be normal distributions. However, Brown and Seo (2010, 2012) argued the drawbacks of fitting parametric distributions to the observations and simulations and proposed a non-parametric post-processor analogous to indicator co-Kriging in geostatistics (Isaaks and Srivastava, 1989). They also discussed that, according to the aggregate effect of various physical processes on meteorological and hydrological variables, the joint behaviour of their observations and

simulations is not usually well-fitted to any parametric distributions. Instead, they proposed a non-parametric method based on Bayesian optimal linear estimation of indicator variables as described by Schweppe (1973). The proposed conditional probability is estimated as the non-exceedance probability of a discrete threshold of the observed variable ( $x \leq c_a$ ; e.g.  $c_a = \text{flood stage}$ ) given the forecast of the  $j^{\text{th}}$  ensemble member ( $z_j$ ). To capture the accurate shape of conditional probability, a large number of thresholds should be defined for the observed variable. A shortcoming of this technique, however, is its inability to specify the conditional probability of a certain observed value given the forecast. In fact, using the non-parametric probability does not allow the conditional probability to be estimated at a particular threshold but rather enables the approximation of the conditional probability of either exceeding or non-exceeding the thresholds. Furthermore, the size of the forecast ensemble is an effective factor in the accurate estimation of the non-parametric conditional probability. Thus, for an accurate estimation of the expectation operator, a relatively large number of forecast members is required.

The present study proposes an alternative approach by applying a group of multivariable probability functions, the so-called copula functions, which do not make any restriction on the type of marginal distributions. Using copula functions, the conditional probability of the observed variable can be estimated at any particular forecast value. Furthermore, as discussed later, the copula functions bind the marginal CDFs. Hence, unknown and complicated relationships in hydrological processes do not hinder fitting the multivariable joint distribution to the observed and forecast variables.

Copula functions, supported by Sklar's Theorem (Sklar, 1959), are multivariate joint distributions of univariate marginal distributions being uniform on the interval [0, 1] (Joe, 1997; Nelsen, 1999). They are capable of modeling the joint behaviour of variables with any level of correlation and dependency; therefore, they seem to offer a potential procedure in post-processing of hydrological and hydro-meteorological forecasts. Copula applications were initially reported in the area of finance and econometrics (Embrechts *et al.*, 2003; Cherubini *et al.*, 2004), while its application in hydrology has received increasing attention over the past few years. As a multivariable joint distribution, they have been frequently employed in flood analyses (e.g., Favre

*et al.*, 2004; De Michele *et al.*, 2005; Zhang and Singh, 2007a; Salvadori and De Michele, 2010), rainfall event analyses (Salvadori and De Michele, 2006; Zhang and Singh, 2007b; Kao and Govindaraju, 2008), and low flow events and drought analyses (Shiau, 2006; Dupuis, 2007; Kao and Govindaraju, 2010; Wong *et al.*, 2010; Madadgar and Moradkhani, 2012).

The aim of this study is first to evaluate and better understand the characteristics of the QM method in the bias correction of hydrological forecasts and then to introduce copula functions as an alternative post-processor for ESPs. An auxiliary variable, the so-called failure index, is also proposed to estimate the performance of the QM technique from the observations and model simulations during an analysis period. The paper is organized as follows. Section 2 explains the QM technique in bias correction of forecasts and addresses the limitations of this method. Thereafter, the failure index reflecting the efficiency of QM performance is proposed. Section 3 describes the mathematics of copula functions and modifies the conditional probability of the bias correction procedure to make it compatible for copulas application. Evaluation of both bias correction methods in Section 4 includes a set of hypothetical case studies and a real case study of river flow forecasts for the Sprague River basin in southern Oregon. Finally, Section 5 provides the conclusion.

## A NEW INDEX FOR ANALYZING THE POST-PROCESSING METHODS

QM is a statistical technique and the most popular post-processing method in hydrologic forecasting that adjusts model forecasts based upon the CDFs of historical observations and model simulations. The approach was primarily designed to remove bias from forecasts; however, its outcome is not always appropriate and under some circumstances, discussed later in this section, may degrade rather than improve the results. In the QM approach, as shown in Figure 1, the forecast quantile at a given time is found from the simulation CDF, and the corresponding observed quantile is taken from the observation CDF to adjust the forecast. A major drawback of this approach, however, is that the pairing between individual simulated and observed values is not preserved, the two CDFs being constructed independently

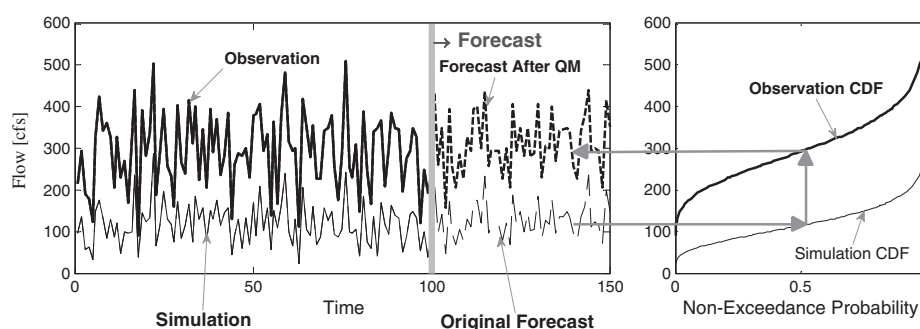


Figure 1. Schematic of quantile mapping technique in post-processing (bias correction) of original forecasts

from each other, so this connection is not represented in any way. Therefore, QM may be also called a ‘blind-matching’ approach that sometimes degrades the results; and in some circumstances, as shown in Figure 2, the adjusted simulated values may deviate even further from the observations than the unadjusted simulated values. As seen in Figure 2, at  $t=3$ , the bias-corrected simulation after QM does not get closer to the corresponding observation but rather moves further away from the observation, creating an even larger error. In other words, the direction of the desired move (towards the observation) is opposite from the adjustment move (by QM application). However, unlike the improper adjustment at  $t=3$ , the original forecast at  $t=7$  moves towards the observed value, and then QM at this point has a positive effect. A large number of points with adjustments in the opposite direction of what is desired may lead to the overall deficiency of the QM method.

Using the historical observations and model simulations, this study proposes a measure ( $\gamma$ ) to predict the overall performance of the post-processing methods like QM technique:

$$\gamma = \frac{1}{T} \sum_{t=1}^T I\{(\beta_t < 0) \text{ or } (\beta_t > 2)\} \quad (1)$$

$$\beta_t = \frac{x_t - y_t}{o_t - y_t} \quad (2)$$

$$x_t = F_O^{-1}\{F_Y(y_t)\} \quad (3)$$

where,  $o_t$  and  $y_t$  are the observation and simulation, respectively, at time  $t$ ;  $x_t$  is the QM-adjusted simulation at time  $t$ ;  $F_O$  and  $F_Y$  are the CDF of observations and simulations, respectively;  $T$  is the number of time steps in the analysis (historical) period; and  $I(\cdot)$  is the Indicator function defined as follows:

$$I\{(\beta_t < 0) \text{ or } (\beta_t > 2)\} = \begin{cases} 1 & \text{if } (\beta_t < 0) \text{ or } (\beta_t > 2) \quad \forall t \in [1, T] \\ 0 & \text{Otherwise} \end{cases} \quad (4)$$

The proposed index,  $\gamma$ , hereinafter called failure ratio of QM, is the fraction of time steps during the analysis period when  $\beta_t$  is negative or greater than 2. Indeed,  $\gamma$  represents the

frequency of simulated values being degraded after QM application, varying between 0 and 1. The term  $\beta$  calculates the ratio of the difference between the simulated and adjusted values to the difference between the simulated and observed values (elaborated later), and it can take any real number in  $\mathbb{R}$ . Since observations are not available for the forecast time period, the QM technique is employed for the analysis period to adjust the simulations and derive  $\gamma$  to predict the performance of QM in forecast mode. It is noted that in the QM technique, the behaviour of the entire system is assumed to be similar in both the analysis and forecast periods, which is equivalent to having identical CDFs in these two periods.

In case of a river flow forecast,  $\beta$  maps the non-negative values of  $y_t$ ,  $o_t$ , and  $x_t$  to a real number  $\mathbb{R}$ ;  $\beta: [0, \infty)^3 \rightarrow (-\infty, \infty)$ . In perfect adjustments,  $\beta_t$  is equal to 1, meaning that the adjusted forecast exactly equals the observation. Any time that the simulation change is not towards the observation, i.e. the movements are not in the same direction,  $\beta_t$  would be negative (Figure 3). Additionally, if both changes have the same direction whereas  $(x_t - y_t) >> (o_t - y_t)$ ,  $\beta_t$  may become greater than 2. Data point  $b$  in Figure 3 shows the situation where both moves are in the same direction but the  $(x_t - y_t)$  is more than twice the  $(o_t - y_t)$ . As can be seen, the absolute error after

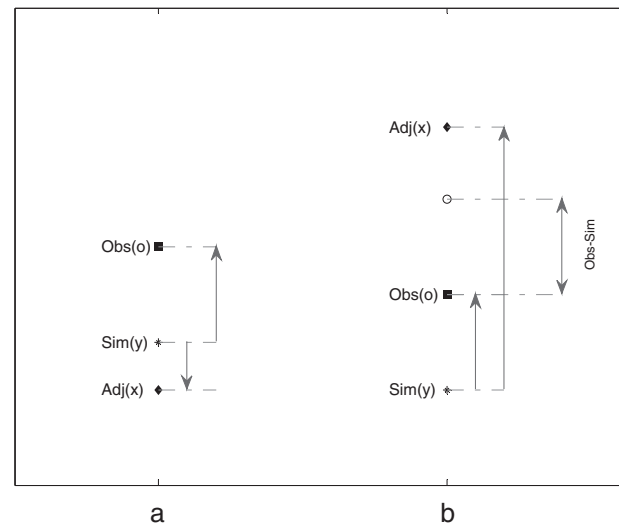


Figure 3. Schematic of data points with a)  $\beta < 0$ , and b)  $\beta > 2$

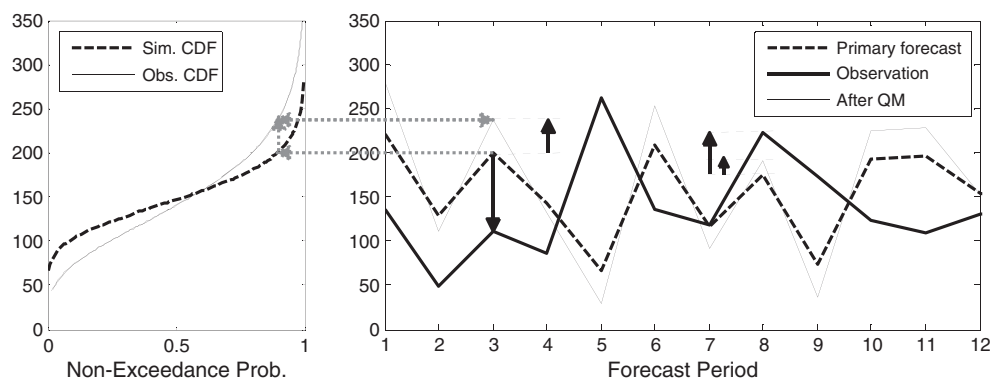


Figure 2. Failure or success of QM method caused by blind-matching procedure

such an adjustment would be greater than the absolute error before the adjustment. Furthermore, as seen in data point  $a$ , the opposite direction of movement causes a larger error regardless of the amount of move. Therefore,  $\beta$ , values smaller than zero or greater than 2 are associated with the data points where the QM method does not perform effectively. And, according to Equation (1),  $\gamma$  (failure index) reflects the frequency of such data points in the analysis period in which the QM technique would have a negative impact on them.

Hence, small values of  $\gamma$  state that the QM technique has been ineffective at only a small number of data points, and as the value of  $\gamma$  increases, more and more data points are negatively affected by the QM method. Therefore, efficient performance of the QM should be accompanied by a small value of  $\gamma$  in Equation (1).

For more clarification on  $\beta$  as the main component of the failure index, two different cases are shown in Figure 4. Simulation and observation time series are fitted to lognormal distributions in each case with different parameter values. The first row of the plots shows associated CDFs, and the second row shows their PDFs. Case A represents a situation where simulated values are very different from the observed values, that is, there is little to no overlap between the simulation and observation ranges as seen in the PDF plots. In such circumstances, moving from the simulated value to the adjusted value is in the same direction as moving from the simulated value to the observed value regardless of where it is located in the range of observations.  $\beta$  is therefore always positive, and QM is an effective approach unless  $\beta$  exceeds 2 in too many points.

Cases with CDFs located close to each other probably have more frequent points with  $\beta > 2$ . Case B shows a situation where an overlap of simulated and observed values occurs. As depicted in the CDF plot, depending on where a simulated value is located, the direction of movement to the adjusted value differs; it may be either towards the corresponding observation or in the opposite direction. Therefore, both positive and negative signs are possible for  $\beta$ . Moreover,  $\beta > 2$  may also occur frequently in such cases. Hence, QM usually functions effectively in cases with distant CDFs and very small or no overlapped PDFs. However, it is more likely that the QM fails where the CDFs are close or the PDFs are largely overlapped. This makes intuitive sense. Despite the deficiency in the QM technique by not accounting for the pairing between individual simulated and observed values, it can still be helpful in correcting gross differences between simulated and observed values. However, when the two distributions are relatively close, as would be the case for a well-calibrated hydrologic model, this deficiency in the QM technique becomes more significant, and the technique may fail.

#### POST-PROCESSING BY COPULA FUNCTIONS

Copulas are joint CDFs of univariate marginal distributions being uniform on the interval  $[0, 1]$ , i.e.  $C: [0, 1]^d [0, 1]$  (Joe, 1997; Nelsen, 1999). Supported by Sklar's Theorem (Sklar, 1959), copulas are able to express the joint behaviour among correlated variables through their marginal CDFs:

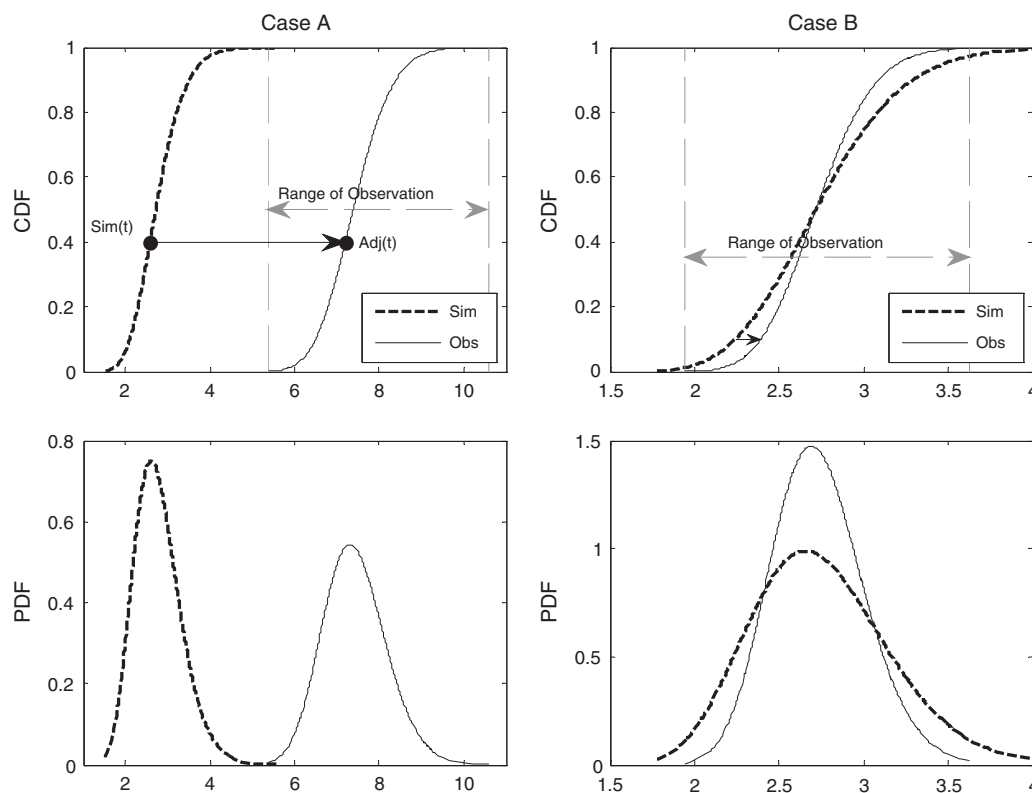


Figure 4. Impact of relative position of simulation and observation CDFs on the performance of the QM adjustments: Case A with distant CDFs is more likely to be well-adjusted by QM method comparing to Case B with close CDFs

$$H(x_1, x_2, \dots, x_n) = C[F_{X_1}(x_1), F_{X_2}(x_2), \dots, F_{X_n}(x_n)] \\ = C(u_1, u_2, \dots, u_n) \quad (5)$$

where,  $x_i$  is the  $i^{th}$  random variable;  $H(\cdot)$  is the joint distribution of random variables;  $F_{X(\cdot)}(\cdot)$  is the marginal distribution of  $X(\cdot)$  random variable;  $C$  is the copula function; and  $u_i$  corresponds to the  $i^{th}$  uniformly distributed variable transformed from  $x_i$  using its marginal distribution function  $F_{X_i}(x_i)$ . A copula satisfies the following properties:

- Boundary conditions

- 1)  $C(\mathbf{u})=0$  if  $\{u_i=0, i \notin \varphi\}$ ; i.e. there is at least one component of  $\mathbf{u}$  where  $u_i=0$ ,  $\varphi$  is the null set.
- 2)  $C(\mathbf{u})=u$  if  $\{u_i=u, u_j=1 \forall j \neq i\}$ ; i.e. all components of  $\mathbf{u}$  are equal to 1 except  $u_i$

- Increasing condition

The probability of any  $n$ -dimensional hypercube in the

unit hypercube is non-negative:  $\sum_{k_1=1}^2 \dots \sum_{k_n=1}^2 (-1)^{\sum_{i=1}^n k_i} C(u_{1k_1}, \dots, u_{1k_1}, \dots, u_{nk_n}) \geq 0$  for all  $0 \leq u_{i1} \leq u_{i2} \leq 1$  where in 2D copula, the conditions are simplified to:

- Boundary conditions

$$3) C(u_1, 0) = C(0, u_2) = 0$$

$$4) C(u_1, 1) = u_1, \quad C(1, u_2) = u_2$$

- Increasing condition

$$C(u_{12}, u_{22}) + C(u_{11}, u_{21}) \geq C(u_{12}, u_{21}) + C(u_{11}, u_{22}) \text{ for } u_{11} \leq u_{12} \text{ and } u_{21} \leq u_{22}$$

According to Equation (5), copulas return the multivariate joint probability of random variables:

$$C(u_1, \dots, u_n) = \Pr\{U_1 \leq u_1, \dots, U_n \leq u_n\} \quad (6)$$

While the copula density,  $c(u_1, \dots, u_n)$ , is defined as:

$$c(u_1, \dots, u_n) = \frac{\partial^n C(u_1, \dots, u_n)}{\partial u_1 \dots \partial u_n} \quad (7)$$

The joint density function can be also decomposed as follows:

$$f(x_1, \dots, x_n) = c(u_1, \dots, u_n) \prod_{i=1}^n f_{X_i}(x_i) \quad (8)$$

The main advantage of copula application is to use separate marginal distributions of random variables while

at the same time their inherent correlations are reflected. Random variables with some level of dependency and correlation are joined through copula functions. There are several families of copula with different properties, among which the Elliptical and Archimedean families (Nelsen, 1999; Embrechts *et al.*, 2003) are employed in this study. Later, in section 4, we summarize a list of copula functions used in this study; where the Gaussian and t-copulas are the Elliptical copula, and Gumbel, Clayton, and Frank copulas are the Archimedean copulas. Elliptical copulas are only able to model the group of variables with a positive-definite correlation matrix. They can also reflect all pair-wise correlations among the variables with any level of correlation. It is statistically proved that a covariance matrix is positive-definite matrix unless one variable is an exact linear combination of the others. Therefore, to ensure the application of the Elliptical family of copulas in real applications, correlation matrix is defined in forms of the covariance matrix. Moreover, this family of copulas does not have a closed-form expression. Unlike Elliptical copulas, the Archimedean copulas have closed-form expressions but do not preserve all pair-wise correlations for problems with more than two variables. Archimedean copulas are divided into symmetric and asymmetric functions; Gumbel and Clayton copulas are from the asymmetric group, and the Frank copula is from the symmetric group.

Since observations and model simulations are correlated variables, copulas can be applied to bind them; and by means of some mathematical efforts, copulas appear to have good potential in bias correction of prediction ensembles. Unlike the QM method, copula-based post-processing would be based upon the joint behaviour of historical observations and simulations, avoiding the drawbacks of 'blind-matching' from ignoring the joint behaviour of these variables in the QM method as discussed earlier. To benefit from copulas in post-processing the forecasts, a mathematical interpretation of bias correction should first be provided. Since bias correction aims at adjusting future forecasts towards the real-time observations, which are unknown at the time of forecast, bias correction of the forecast would be mathematically expressed as finding the mode of the conditional probability distribution of the observation given the forecast where the system behaviour in the future is assumed to be similar to that in the past. Following Equation (8), adjustment of simulations towards associated observations at time  $t$  ( $x_t$  in Equation 2) is developed below if the observations and model simulations during the analysis period are joined through a copula framework:

$$x_t = \text{Max}[f(o_t|y_t)] \\ f(o_t|y_t) = \frac{f(o_t, y_t)}{f(y_t)} = \frac{c(U_O = u_o, U_Y = u_{y_t}) f(y_t) f(o_t)}{f(y_t)} \\ = c(U_O = u_o, U_Y = u_{y_t}) f(o_t) \quad (9)$$

Where;  $f(o_t, y_t)$  is the joint PDF of the observation and simulation at time  $t$ ,  $f(\cdot)$  is the marginal PDF of desired



variable, and  $c(.,.)$  is the copula PDF of marginal distributions. As discussed later in section 4, the copula density function -  $c(.,.)$  - is obtained from an analysis period before the forecast date.

The first two terms of Equation (9) simply refer to the basic definition of conditional PDF, where the contribution of the copula is reflected in the third term. Since  $f(o_t|y_t)$  returns the conditional PDF of the observation given the simulation at time  $t$ , the mode of this PDF is correspondent to the most probable observation at that time. Therefore, the observation with highest conditional probability is taken as the adjusted simulation. To build the conditional PDF of Equation (9) and extract its mode, we suggest Monte Carlo sampling from the copula density function in Equation (9), where  $u_{y_t}$  is computed for the simulated value at time  $t$ . In  $c(U_O = u_o, U_Y = u_{y_t})$ ,  $U_Y$  is fixed at  $u_{y_t}$  at time  $t$  while  $U_O$  is to be varied and tested to return the mode of the conditional PDF. Once a sample from the copula density function is taken, the right-hand side of Equation (9) is obtained. Proceeding the Monte Carlo sampling leads to form the conditional PDF-  $f(o_t|y_t)$  whose mode is the most probable observation given the simulation at time  $t$ . Figure 5 visualizes the PDF of a copula and the marginal distribution at  $U_1 = 0.8$  ( $c(U_2, U_1 = 0.8)$ )

#### APPLICATION OF POST-PROCESSING METHODS

The post-processing methods employed in this study are evaluated by hypothetical and real case studies. In the hypothetical case study, simulations and observations are sampled from separate parametric distributions, and then each post-processing method is applied. In the real case study, the streamflow forecasts for a river basin in southern Oregon, USA are post-processed using the above methods.

##### Hypothetical cases

To evaluate the performance of each post-processing method and explore the relation of  $\gamma$  values (Equation (1)) with the effectiveness of each method, 2500 sets of simulation and observation data series are tested in this

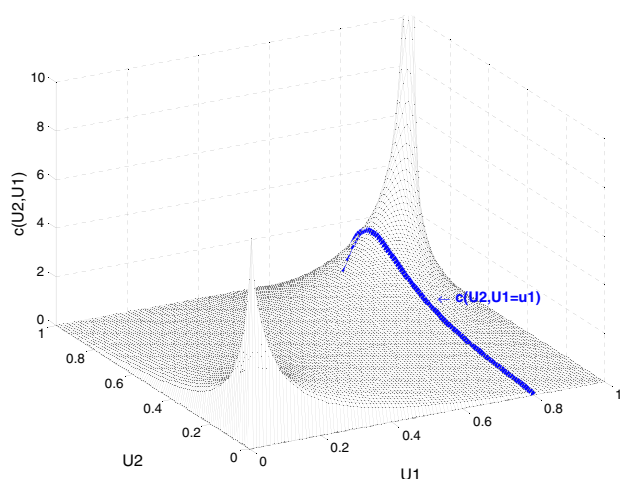


Figure 5. Schematic of a copula PDF with marginal distribution of  $U_2$  shown by the solid line

section. Test cases are generated independently from each other. Each case is to describe the dependence between the simulations and observations by an appropriate t-copula. Simulations and observations of each case have a level of dependency and correlation with each other; nevertheless, the simulations and observations in a single case are produced totally independent from those of another case. Gamma and lognormal distributions with 30 different parameter sets are used to randomly sample the simulations and observations of 2500 cases. Following steps are taken to form the hypothetical case studies:

1.  $N = 1$ , case number
2. Form the data series for the analysis period
  - a. Sample from a parametric distribution ( $D_1$ ) for 1000 times to build the simulation timeseries.  $D_1$  is either Gamma or Lognormal distribution.
  - b. Sample from either Gamma or Lognormal distribution ( $D_2$ ) for 1000 times to build the observation timeseries.
3. Find a bivariate t-copula to describe the dependence between the simulations and observations generated in steps (2-a) and (2-b).
4. Form the data series for the forecast period
  - a. Sample from a  $D_1$  (step 2-a) for 12 times to build the forecast timeseries. Forecast lead-time is set as 12.
  - b. Repeat step (4-a) for 50 times to make a forecast ensemble with 50 traces.
  - c. Sample from a  $D_2$  (step 2-b) for 12 times to build the observation timeseries.
5. Post-process the forecast ensemble obtained in step (4-a) by either copula-based or QM method.
6.  $N = N + 1$
7. If  $N \leq 2500$ , then go to 2. Else, terminate!

In step (4-b), the real-time observations in forecast mode are sampled from the observation distribution function to enable a performance evaluation of the post-processing methods. The list of performance measures used in this study is presented in Table I. Point-wise performance measures are utilized in evaluating the ensemble mean forecast while the probabilistic measures are used to assess the performance of the forecast ensembles. Figure 6 shows the results of the QM technique against the copula-based post-processor. Probability of success in Figure 6 is the probability that the employed post-processing method performs successfully with respect to the associated metric for different values of the failure index ( $\gamma$ ).  $\gamma$  is computed for the analysis period of each case, and then cases with a given value of  $\gamma$  are taken out from the pool of 2500 cases. Therefore, for each metric of interest, cases with successful performance are counted to compute the probability of success. Success is defined upon the metric value, that is, if implementation of the post-processing method improves

Table I. Performance metrics to evaluate the employed bias correction methods

Performance measure	Mathematical representation	Terminology	Description
Mean Absolute Error (MAE)	$MAE = \frac{1}{T} \sum_{t=1}^T  \bar{y}_t - o_t $	$\bar{y}_t$ : Average over ensemble predictions at time $t$ $o_t$ : Observation at time $t$	Deterministic metric, varies $[0, \infty)$ , with perfect score of 0.
Nash–Sutcliffe Efficiency <sup>a</sup> (NSE)	$NSE = 1 - \frac{1}{\sigma_o^2} \left[ \frac{1}{T} \sum_{t=1}^T (\bar{y}_t - o_t)^2 \right]$	$\sigma_o^2$ : Variance of observations	Deterministic metric, varies $(-\infty, 1]$ , with perfect score of 1.
Rank Probability Skill Score <sup>b</sup> (RPSS)	$RPSS = 1 - \frac{RPSS}{RPSS_{ref}}$ $RPSS_t = \sum_{k=1}^K [P_k(y_t) - P_k(o_t)]$ $P_k(y_t) = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} I(y_{t,i} < Q_k)$ $P_k(o_t) = I(o_t < Q_k)$ $RPSS_{ref} = \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^K [P_k(o_t) - \overline{P(o)}]$	$I$ : Indicator function $Q_k$ : $k^{th}$ threshold of $Q_{\overline{P(o)}}$ : Average probability of all observations over all thresholds. $\overline{P(o)} = \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^K P_k(o_t)$	Probabilistic metric, varies $(-\infty, 1]$ , with perfect score of 1. Negative values mean predictions perform worse than the average observation; while for $RPSS = 0$ , predictions transfer the same information as the average observation does.
Reliability <sup>c</sup> ( $\alpha$ )	$\alpha = 1 - 2 \left[ \frac{1}{T} \sum_{t=1}^T  P_t(o_t) - U(o_t)  \right]$ $P_t(o_t) = \frac{I(y_{t,i} < o_t)}{N_{ens}}$	$P_t(o_t)$ : Non-exceedance probability of $o_t$ using the ensemble predictions at time $t$ $U(o_t)$ : Non-exceedance probability of $o_t$ using the uniform distribution $U[0,1]$	Probabilistic metric, measuring the closeness of quantile plot of the observations to the corresponding uniform quantiles. It varies $[0, 1]$ , while 0 is the worst and 1 is the best score.
Reliability <sup>c</sup> ( $\varepsilon$ )	$\varepsilon = 1 - \varepsilon'$ $\varepsilon' = \frac{1}{T} \sum_{t=1}^T I[P_t(o_t) = 1 \text{ or } P_t(o_t) = 0]$	Similar to $\alpha$	Probabilistic metric, indicating the complement of the average times where the observation falls outside the ensemble predictions. It varies $[0, 1]$ , while 0 is the worst and 1 is the best score.
Resolution <sup>c</sup> ( $\pi$ )	$\pi = \frac{1}{T} \sum_{t=1}^T \frac{E[y_{t,i}]}{\sigma[y_{t,i}]}$	$E[y_{t,i}]$ : Expected value of ensemble predictions at time $t$ $\sigma[y_{t,i}]$ : Standard deviation of ensemble predictions at time $t$	Probabilistic metric, a measure of precision of ensemble predictions. Greater values indicate less precision (larger uncertainty) of forecasts.

<sup>a</sup> Nash and Sutcliffe (1970)<sup>b</sup> Wilks (2006)<sup>c</sup> Renard *et al.* (2010)



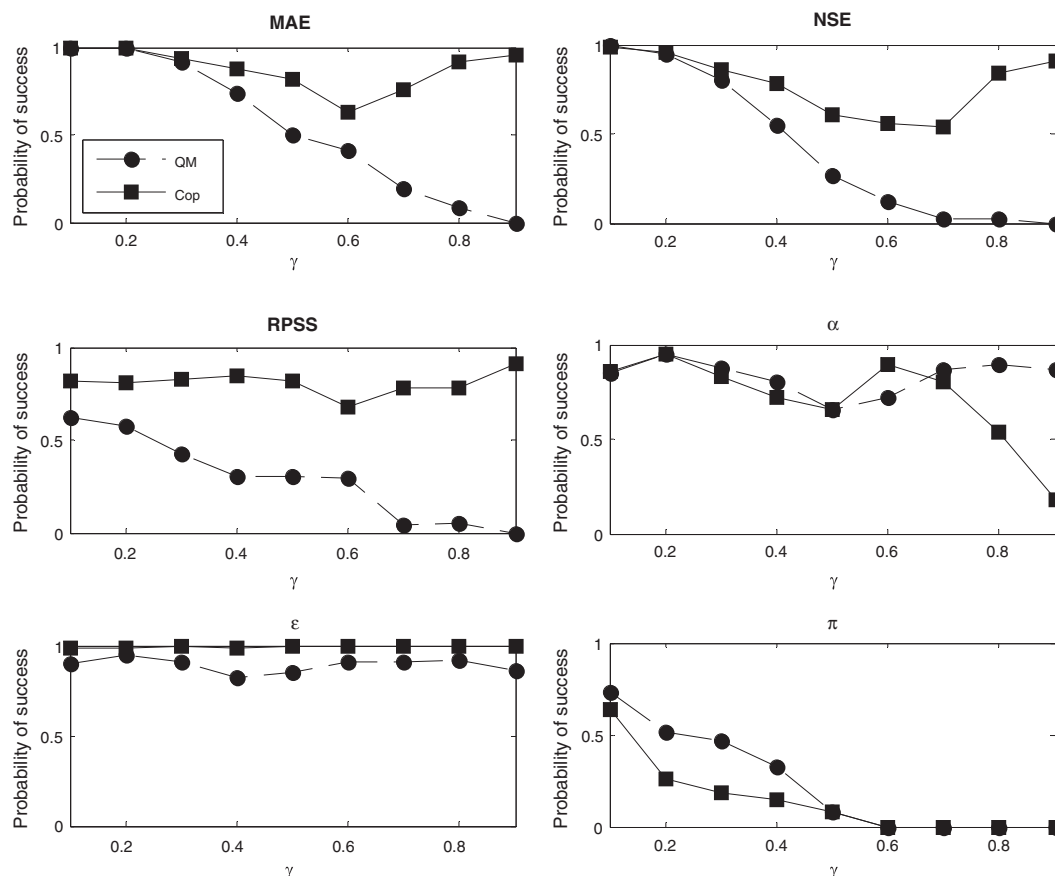


Figure 6. Probability of success against  $\gamma$  for point-wise (MAE and NSE) and probabilistic performance measures (RPSS,  $\alpha$ ,  $\varepsilon$ ,  $\pi$ ) in QM and copula-based post-processing methods. Probability of success is obtained with respect to the associated metric for different values of the failure index

the metric score towards its perfect value as noted in Table I, then the method is considered as successful for that metric. Figure 6 shows that as  $\gamma$  increases, the probability of success strictly declines in the first three metrics (Mean Absolute Error (MAE), Nash–Sutcliffe Efficiency (NSE), and Rank Probability Skill Score (RPSS)) when QM is in use. Given the definition of  $\gamma$  in Equation (1), if the post-processing method constantly degrades the simulations,  $\gamma$  becomes greater and approaches 1. In such circumstances, the QM method may not be able to improve the forecasts owing to its inherent blind-matching nature that is solely dependent on quantiles. Evidently, the probability of success in the QM method is dependent on the  $\gamma$  value, whereas this is not the case for the copula-based post-processing method. The main reason of insensitivity of copula-based method to the failure index value is its ability to model the joint behaviour of the simulations and observations unlike the QM method with inherent blind-matching approach. In other words, the copula approach is able to perform effectively even in cases with a large failure index. Generally, the copula approach is more likely to succeed than the QM method in the first three metrics. Other metrics in Figure 6 ( $\alpha$ ,  $\varepsilon$ , and  $\pi$ ) are the supportive quantitative scores derived from the predictive quantile-quantile (QQ) plot (Laio and Tamea, 2007; Thyer *et al.*, 2009), which compares the empirical CDF of the probability of observations ( $P_f(o_t)$  in Table I) using the forecast ensemble at each

time  $t$  (CDF of the probabilities) against the CDF of a uniform distribution. For a perfect forecast ensemble, the empirical CDF of the  $p$ -values is consistent with the CDF of the uniform distribution on the interval  $[0,1]$ . The metrics  $\alpha$  and  $\varepsilon$  assess the reliability of forecasts, and  $\pi$  indicates the resolution (precision, sharpness) of the forecast ensemble. According to Thyer *et al.* (2009), as the area between the empirical CDF of the observations'  $p$ -values and the CDF of the uniform distribution in the predictive QQ plot becomes larger, the value of  $\alpha$  decreases towards zero. Results indicate that for  $\gamma \leq 0.7$ , the post-processing methods perform closely, while for large  $\gamma$  values, the QM method is more successful than the copula-based method for the  $\alpha$  measure. The subplot of the  $\varepsilon$  metric illustrates that the copula method is more effective than the QM method (regardless of  $\gamma$  value) to envelop observations after post-processing of the forecasts. In other words, fewer observations fall outside the range of the forecast ensemble after post-processing by the copula approach. The resolution ( $\pi$ ), also called sharpness, states that adjustment by QM leads to greater resolution (precision). However, comparison of sharpness may not be a meaningful approach when the employed methods do not primarily perform equally in the  $\alpha$  and  $\varepsilon$  metrics. Assuming that precision has lower priority than reliability, given similar forecast reliabilities, the method with greater resolution (lower uncertainty) is preferred; otherwise, the method with higher resolution does not reveal any superiority.

As a brief summary of the hypothetical case results, the multivariate copula-based post-processor performs considerably better than the QM method in the point-wise measures. For the RPSS metric among the probabilistic measures, the copula procedure is again evaluated as a much better method than QM. The predictive uncertainty is also more reliable in encompassing observations when the multivariate copula-based post-processor is in use. Moreover, unlike the QM method, performance of the multivariate post-processor is generally insensitive to the failure index of the analysis period. Using the QM method, the predictability of the forecast ensemble is not effectively improved in cases with large  $\gamma$  values, illustrating the drawback of the blind-matching procedure that corresponds to the same quantiles of simulation and observation CDFs.

#### *Streamflow forecast in Sprague River basin*

The Sprague River basin, with a drainage area of approximately 4100 km<sup>2</sup>, is a sub-basin of the Upper Klamath River basin located in southern Oregon and northern California, USA (Figure 7). The Sprague River valley is enclosed by forested mountain ridges and includes large marshes, meadows, and irrigated pastures. A large proportion of irrigation water demand is supplied by river flow, and the rest is pumped from local wells. A major environmental concern in the Sprague River basin is the water quality, which directly impacts fish and wildlife habitat throughout the Upper Klamath basin as reported by Klamath Basin Ecosystem Foundation (2007). Some flow conditions interrupt fish passage through the Sprague River, which necessitates accurate forecast for better understanding of flow conditions in coming seasons. The Sprague River is also a major tributary to Upper Klamath Lake, an important and highly contested water body used for irrigation water supply, hydropower generation, and fish habitat.

The bias correction methods discussed in previous sections are applied to streamflow forecasts of the Sprague River basin. A distributed parameter hydrologic model – the U.S. Geological Survey Precipitation-Runoff Modeling System (PRMS; Leavesley *et al.*, 1983) – is employed to predict daily river flow of the basin. Temperature and precipitation data for PRMS are retrieved from two different sources: the NWS Cooperative Network, and the NRCS Snow Telemetry network of weather stations. Using the Shuffled Complex Evolution global search algorithm (Duan *et al.*, 1994), the model parameter sets are calibrated within the multiple-objective stepwise calibration Let Us calibrate (LUCA) (Hay and Umemoto, 2006) framework.

#### *Results*

Copula application starts with fitting appropriate marginal distributions to the variables aimed for the bias correction procedure. Monthly flow observations and model simulations of Sprague River basin outflow are each fitted to eight distributions, including Gamma, Generalized Extreme Value (GEV), Lognormal, Gaussian, Generalized Pareto (GP), Weibull, Gumbel, and Exponential distributions. Several forecast periods starting from different months (Jan, Feb, Mar) of 2001–2003 are analyzed for the post-processing application. The forecast lead-time is fixed at 6 months, and the nine forecast periods are chosen as Jan–Jun, Feb–Jul, and Mar–Aug for three years of 2001–2003. The marginal distributions are separately fitted to historical observations and model simulations in the analysis periods. The historical period from 1980 to 2000 is used to set the analysis periods associated with each forecast period. The analysis periods are then taken as Jan–Jul, Feb–Jul, and Mar–Aug of 1980–2000. The histograms of monthly averaged PRMS simulations and river flow observations for the analysis period of Feb–Jul in 1980–2000, and the fitted distributions are shown in

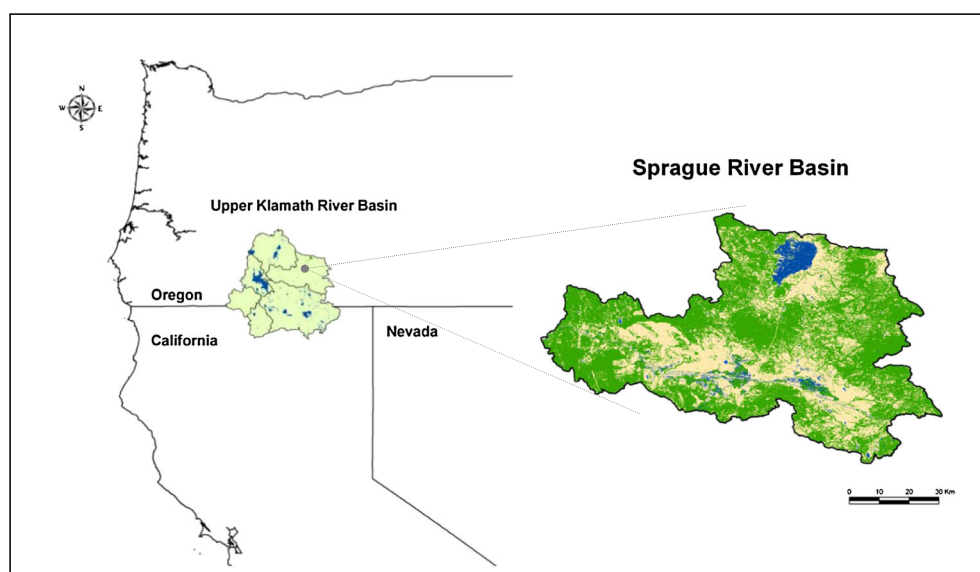


Figure 7. Sprague River basin, a sub-basin of Upper Klamath River Basin in southern OR and northern CA

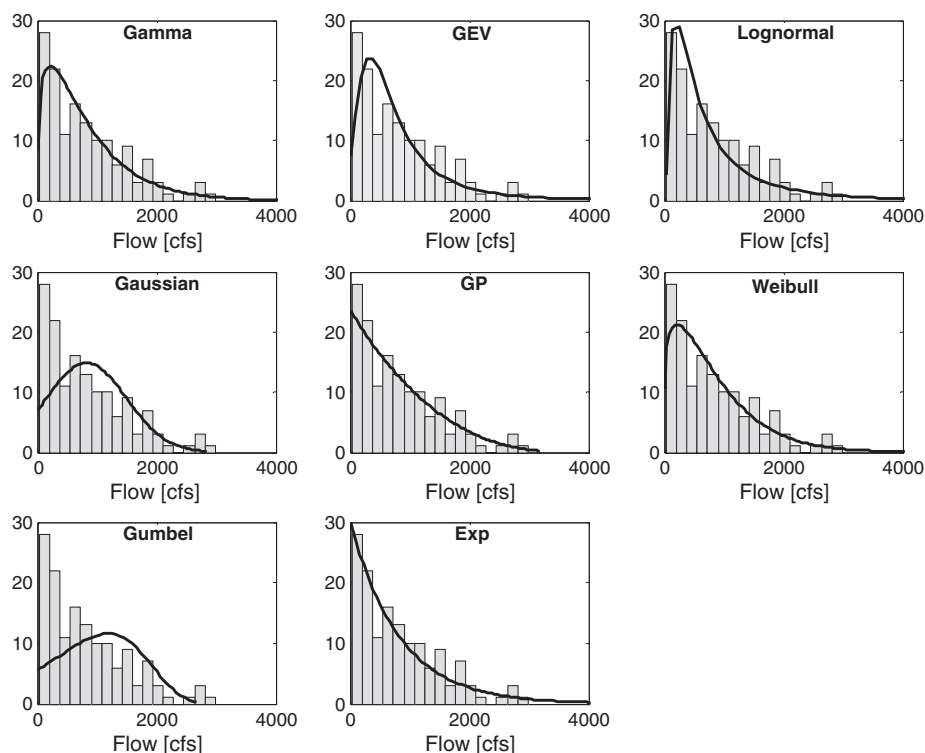


Figure 8. Flow histogram against marginal distribution fitted to monthly averaged PRMS simulations in the analysis period of Feb–Jul, 1980–2000

Figures 8 and 9. The parameters of the marginal distributions are estimated by the Maximum Likelihood Estimation method. From visual inspection, most theoretical distributions except Gaussian and Gumbel are well-fitted to PRMS simulations. It seems hard, however, to find a suitable distribution to fit flow observations properly, with only the GEV and Lognormal

distributions looking suitable. Table II lists the statistics used for evaluation of the theoretical marginal distributions. The Kolmogorov–Smirnov (K–S) test is used to evaluate the appropriateness of the distribution fitted to the data. Its test statistic ( $D$ ) measures the maximum distance of the empirical CDF to the CDF of the reference distribution:

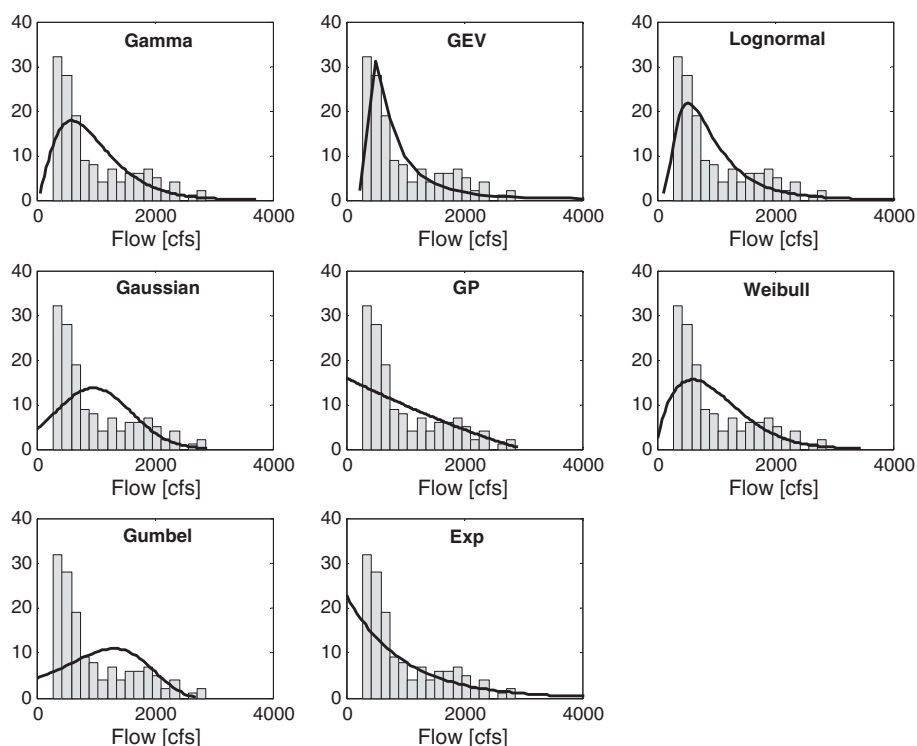


Figure 9. Flow histogram against marginal distribution fitted to monthly averaged river flow observations during the analysis period of Feb–Jul, 1980–2000

Table II. K–S test statistics of fitting different distributions to the simulated and observed flows during different analysis periods in 1980–2000

	Jan–Jun			Feb–Jul			Mar–Aug		
	<i>D</i>	<i>P</i> -value	Hypothesis test	<i>D</i>	<i>P</i> -value	Hypothesis test	<i>D</i>	<i>P</i> -value	Hypothesis test
Simulated flow									
<i>Gamma</i>	0.08	<b>0.25</b>	Accept	0.07	<b>0.39</b>	Accept	0.07	<b>0.44</b>	Accept
<i>GEV</i>	0.08	<b>0.25</b>	Accept	0.09	<b>0.22</b>	Accept	0.07	<b>0.40</b>	Accept
<i>Logn</i>	0.11	<b>0.07</b>	Accept	0.10	<b>0.10</b>	Accept	0.08	<b>0.27</b>	Accept
<i>Gaus</i>	0.13	0.01	Reject	0.18	0.03	Reject	0.12	0.03	Reject
<i>GP</i>	0.08	<b>0.22</b>	Accept	0.06	<b>0.64</b>	Accept	0.06	<b>0.64</b>	Accept
<i>Wbl</i>	0.08	<b>0.20</b>	Accept	0.08	<b>0.31</b>	Accept	0.07	<b>0.40</b>	Accept
<i>Gumbel</i>	0.22	2.2E-6	Reject	0.21	5.8E-6	Reject	0.21	3.2E-6	Reject
<i>Exp</i>	0.07	<b>0.47</b>	Accept	0.08	<b>0.29</b>	Accept	0.08	<b>0.25</b>	Accept
Observed flow									
<i>Gamma</i>	0.13	0.01	Reject	0.15	0.01	Reject	0.16	0.002	Reject
<i>GEV</i>	0.08	<b>0.32</b>	Accept	0.06	<b>0.24</b>	Accept	0.08	<b>0.25</b>	Accept
<i>Logn</i>	0.10	<b>0.13</b>	Accept	0.14	0.03	Reject	0.13	0.01	Reject
<i>Gaus</i>	0.18	1.3E-3	Reject	0.22	5.2E-5	Reject	0.21	3.4E-6	Reject
<i>GP</i>	0.18	1.6E-4	Reject	0.24	3.2E-5	Reject	0.21	7.1E-6	Reject
<i>Wbl</i>	0.14	9E-3	Reject	0.18	0.006	Reject	0.16	0.001	Reject
<i>Gumbel</i>	0.21	4.4E-6	Reject	0.29	1.4E-6	Reject	0.23	3.7E-7	Reject
<i>Exp</i>	0.24	1.3E-7	Reject	0.27	9.6E-9	Reject	0.28	3.7E-10	Reject

$$D = \text{Max} \{ |F(x) - G(x)| \} \quad (10)$$

where  $F(x)$  and  $G(x)$  are the empirical and reference CDFs, respectively. The results of K–S test is reported in Table II along with the  $p$ -value at the selected significance level of  $\alpha = 0.05$ . The null hypothesis ( $H_0$ ) of the K–S test states that the data set belongs to the reference distribution.

According to the test statistics in Table II, the Gaussian and Gumbel distributions are not suitable choices for simulated flows for any of the three analysis periods, while the other distributions fit more or less well. On the contrary, almost none of the distributions except GEV properly fit the observed flows. These results have been visually verified in Figures 8 and 9 for the analysis period of Feb–Jul, 1980–2000. Furthermore, for the Jan–Jun analysis period, the Lognormal distribution is the second-most suitable choice for the observed flow. However, the GEV distribution is the best candidate for the observations of any analysis period; hence, for the copula application, the GEV distribution is hereinafter coupled with the marginal distributions of the simulated flows.

The Elliptical and Archimedean families of copulas (Table III) are applied to join the marginal distributions of historical monthly observations and model simulations during each analysis period (Jan–Jun, Feb–Jul, and Mar–Aug of 1980–2000). To verify which copula describes the dependence between the observation and simulation better than others, various methods may be applied as the goodness-of-fit (GOF) test. The simplest method is a visual comparison between the empirical copula and the theoretical copula. The scatterplot would follow the line 1:1 if the theoretical copula perfectly fit the empirical copula. Nevertheless, to compare different copulas fitted to the same set of data, it is more reliable to use the GOF test statistics instead of a mere visual

inspection. A mathematical GOF test for copula functions may be based on the distance between the empirical copula and the parametric copula under the null hypothesis ( $H_0$ ). Genest and Rémillard (2008) implemented a bootstrapping process to obtain the Cramér-von Mises (Equation (11)) and K–S statistics as the measures of distance between the empirical and parametric copulas. There are some other test statistics analogues to the Cramér-von Mises and K–S statistics which are based on Kendall's transform (Genest *et al.*, 2006; Savu and Trede, 2008) and Rosenblatt's transform (Rosenblatt, 1952). Recently, the extended version of the Kullback–Leibler Information Criterion developed by Diks *et al.* (2010) has been applied in copula selection (Weiß, 2011), but the results showed that the criterion does not perform better than GOF test statistics in detecting the best copula fitted to the data. On the other hand, some studies show that the GOF test statistics based on the empirical copula outperform the others (Berg, 2009; Genest *et al.*, 2009). Therefore, this study proceeds with the GOF test statistic based upon the empirical process with the following definition for Cramér-von Mises statistic:

$$S_n = \int_{\mathbf{u}} \Delta C_n(\mathbf{u})^2 dC_n(\mathbf{u}) \quad (11)$$

where,  $S_n$  is Cramér-von Mises statistic, and  $\Delta C_n$  is expressed as:

$$\Delta C_n = \sqrt{n}(C_n - C_{\theta_n}) \quad (12)$$

where  $C_n$  is the empirical copula with a sample size of  $n$ , and  $C_{\theta_n}$  is the parametric copula estimated for a sample size of  $n$ . Genest *et al.* (2009) elaborated on a parametric bootstrap procedure to find the  $p$ -value of the test via Monte Carlo sampling. Since the null hypothesis of the test is that the

Table III. Summary of bivariate copula functions used for bias correction of river flow forecast

Copula	Function	Support
Gaussian	$C(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_2)} \int_{-\infty}^{\Phi^{-1}(u_1)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{2(1-\rho^2)}\right\} dx_1 dx_2$ $u_1 = \Phi(x_1), \quad u_2 = \Phi(x_2)$ $\rho: \text{Linear correlation coefficient}$ $\Phi: \text{Standard normal cumulative distribution function}$	$x_1, x_2 \in R$
T	$C(u_1, u_2) = \int_{-\infty}^{t_v^{-1}(u_2)} \int_{-\infty}^{t_v^{-1}(u_1)} \frac{1}{2\pi(1-p^2)^{1/2}} \exp\left\{1 + \frac{x_1^2 + x_2^2 - 2px_1 x_2}{v(1-p^2)}\right\}^{-(v+2)/2} dx_1 dx_2$ $u_1 = t_v(x_1), \quad u_2 = t_v(x_2)$ $\rho: \text{Linear correlation coefficient}$ $t_v: \text{Cumulative distribution function of } t \text{ distribution with } v \text{ degree of freedom.}$	$x_1, x_2 \in R$
Gumbel	$C(u_1, u_2) = \exp\left\{-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta\right]^{1/\theta}\right\}$ $\theta: \text{Measure of dependency between } u_1 \text{ and } u_2. \text{ Either Pearson's correlation coefficient or Kendall's tau correlation are usually used to estimate } \theta.$	$\theta \in [1, \infty)$
Clayton	$C(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$ $\theta: \text{Similar to Gumbel copula.}$	$\theta \in (0, \theta)$
Frank	$C(u_1, u_2) = -\frac{1}{\theta} \ln \left[ 1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1} \right]$ $\theta: \text{Similar to Gumbel copula.}$	$\theta \in R$

parametric copula fits the data ( $H_0: C_n \in C_{\theta_n}$ ),  $p$ -values greater than the significance level ( $\alpha$ ) means the null hypothesis is accepted; otherwise, it is rejected. Therefore, among a group of copulas, the one with the greatest  $p$ -value (and smallest  $S_n$ ) is preferred.

As described earlier, three analysis periods having their specific marginal distributions are tested in this study. Elliptical and Archimedean copulas are fitted to the simulations and observations of each analysis period; the  $S_n$  statistic and corresponding  $p$ -values of testing the null hypothesis ( $H_0: C_n \in C_{\theta_n}$ ) are summarized in Table IV. Results are the mean value of Cramér-von Mises statistic ( $S_n$ ) and corresponding  $p$ -value when a given copula is applied to different combinations of marginal distributions. As discussed earlier, GEV is selected for observed flows, and Gamma, GEV, Lognormal, GP, Weibull, and Exponential distributions are selected for simulated flows. The  $p$ -values are computed using a parametric bootstrap procedure with  $N=1000$  replications and a significance level of  $\alpha=0.01$ . In each case, the copula function with the smallest  $S_n$  (Equation (11)) and largest  $p$ -value is preferred; hence, among the copula functions, the Frank copula is the best choice for the Jan–Jun and Feb–Jul periods, whereas the Gumbel copula is the best for the Mar–Aug period. However, the Clayton copula function is the worst choice to describe the dependence during any analysis period.

Using Equation (9), post-processing of each forecast ensemble is implemented after fitting the appropriate copula to the univariate marginal distributions of the

associated analysis period. To evaluate the performance of the methods, several 6-month hind-cast periods within three target years (2001, 2002, and 2003) are tested. As explained earlier, three hind-cast periods are chosen in each target year: Jan–Jun, Feb–Jul, and Mar–Aug. Note that post-processing is applied to monthly averaged flows of each forecast period. Using the same performance metrics illustrated in Figure 6 and Table I, the performance of post-processing methods on the stream-flow forecasts of the Sprague River basin are shown in Figures 10–12. The QM and copula post-processor are both used to adjust the hind-casts predicted by PRMS. The initial states of the model are obtained by running the PRMS model up to the hind-cast date. To implement the multivariate copula-based post-processor, the selected copula function (see Table IV) is applied to all (six) possible combinations of marginal distributions that are best fitted to the simulated and observed flows in the analysis period (see Table II). The curves of copula application in Figures 10–12 are associated with the average metric value over all the combinations of marginal distributions. Prior to the QM application, the ability of QM to improve forecasts would be predicted from the value of  $\gamma$  estimated for the analysis period. The failure index for Jan–Jun, Feb–Jul, and Mar–Aug analysis periods through 1980–2000 is, respectively, found as 0.32, 0.29, and 0.28. Thus, according to Figure 6, it was not expected that QM would produce encouraging results for any of the analysis periods. As shown in Figures 10–12, the QM method is not effective in



Table IV. Results of GOF test for copula selection in each analysis period. Values of Cramér-von Mises statistic ( $S_n$ ) are presented along with the corresponding  $p$ -value in parentheses. Statistics of the best fitted copulas are bolded

Analysis period copula function	Jan–Jun	Feb–Jul	Mar–Aug
Gaussian	0.0232 (0.0664)	0.0292 (0.0255)	0.0550 (0.0005)
T	0.0330 (0.0055)	0.0406 (0.0025)	0.0653 (0.0005)
Gumbel	0.0242 (0.0644)	0.0299 (0.0315)	<b>0.0383</b> <b>(0.0135)</b>
Clayton	0.2173 (0.0005)	0.2316 (0.0005)	0.3239 (0.0005)
Frank	<b>0.0217</b> <b>(0.0794)</b>	<b>0.0257</b> <b>(0.0415)</b>	0.0521 (0.0025)

improving the forecast ensemble with respect to the point-wise measures; MAE and NSE show the general failure of QM in reducing the error of the mean forecast. The copula post-processor, however, performs better than the QM method, and it adjusts the forecast ensemble closer to the observations except for the Mar–Aug forecast period of 2003. For the Jan–Jun and Feb–Jul forecast periods in any of the target years, the copula function performs significantly better than the original forecast and the QM method. Regarding the RPSS metric, QM generally fails to improve forecast traces, and it even worsens the quality of original forecasts in almost all forecast periods. Failure of the QM method is also predictable according to the  $\gamma$  values and associated results in Figure 6. Copula application, on the other hand, is consistently the prominent method for the Jan–Jun and Feb–Jul forecast periods. As the forecast starting date moves towards spring, the performance of multivariate post-processing

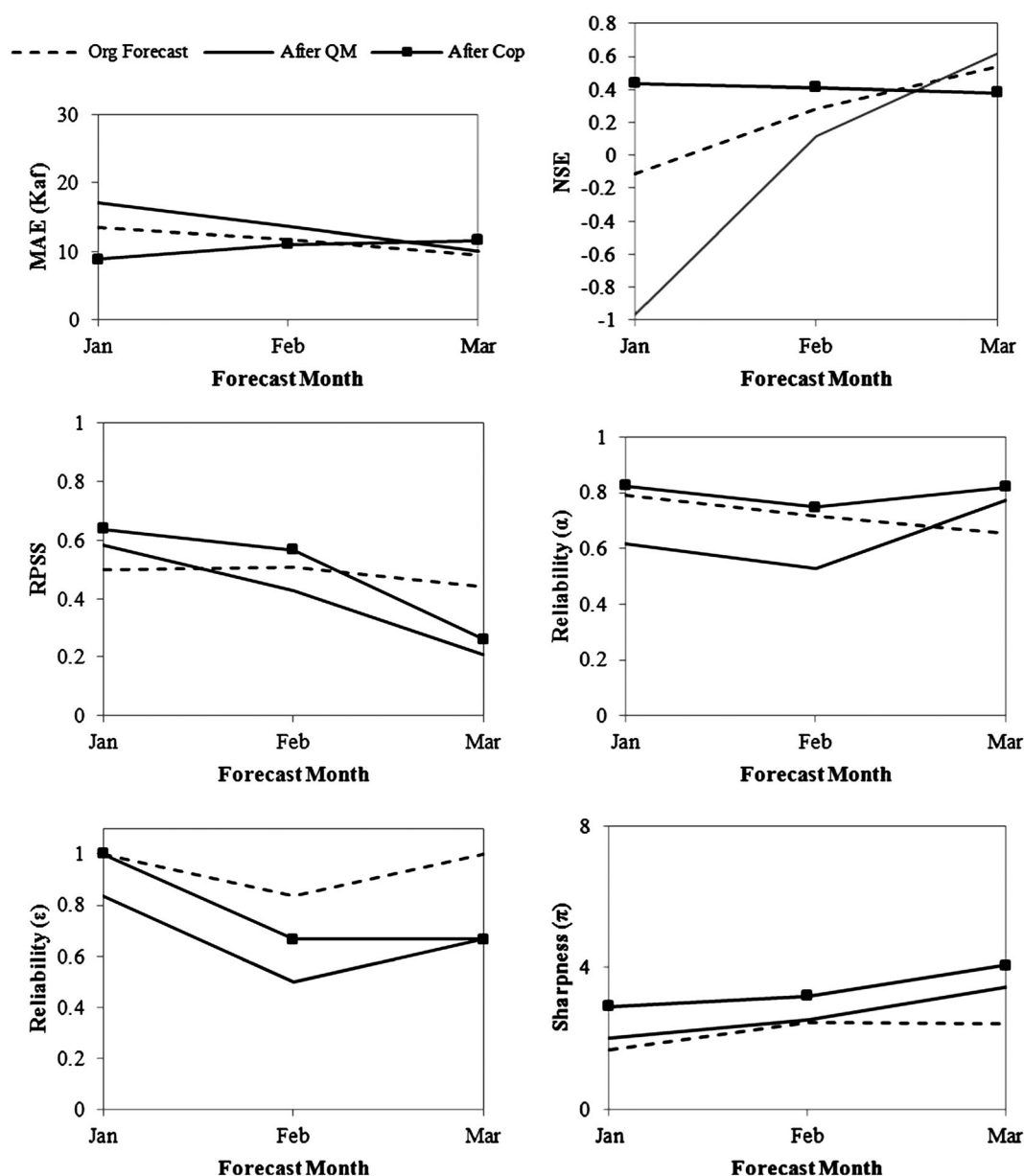


Figure 10. Comparing the performance of post-processing methods in adjusting the monthly streamflow hind-cast starting from different forecast dates in 2003. The forecast period of each forecast ensemble is 6 months



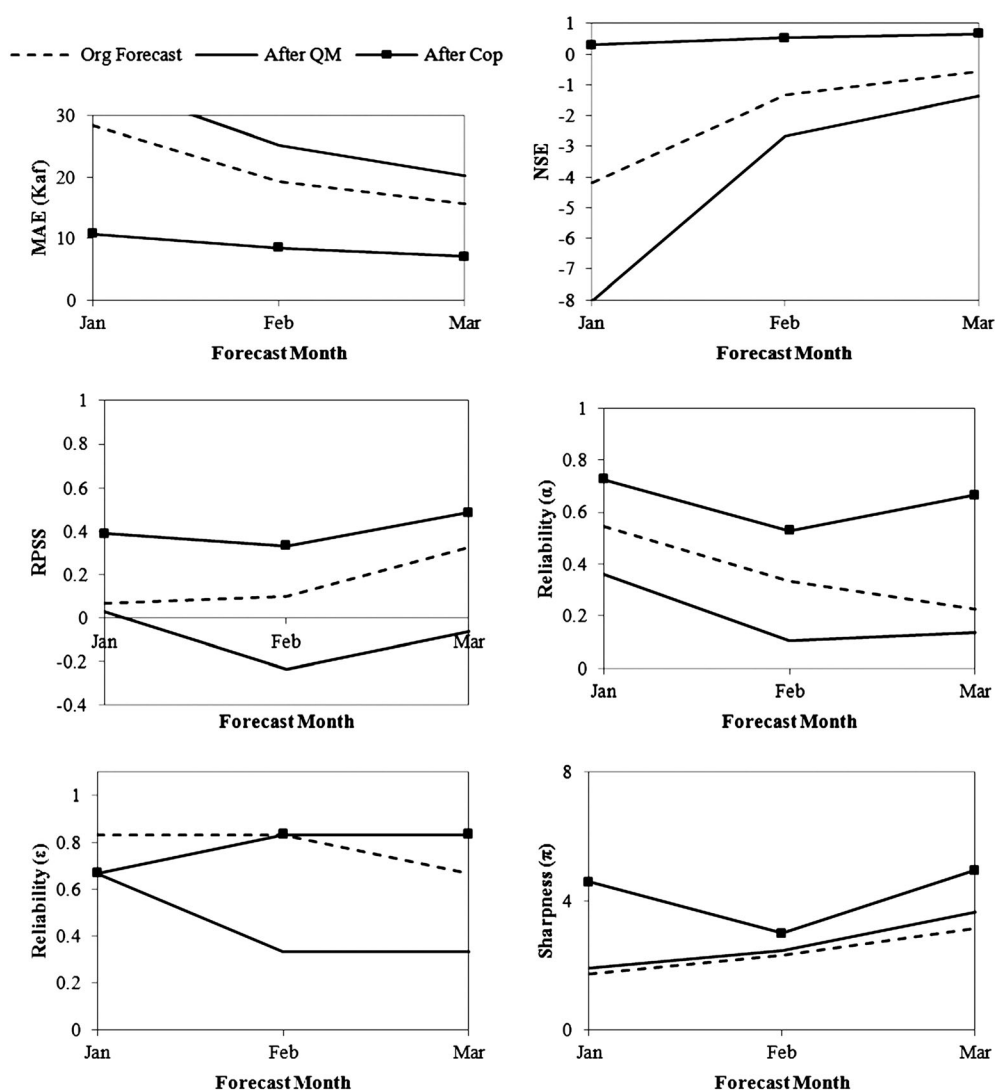


Figure 11. Comparing the performance of post-processing methods in adjusting the monthly streamflow hind-cast starting from different forecast dates in 2002. The forecast period of each forecast ensemble is 6 months

gets closer to that of the original forecast; however, it is still better than the QM results. The reliability metric derived from a QQ plot,  $\alpha$ , shows that QM adjustments are not reliable compared to original forecasts, while the proposed copula method performs better than the original forecasts. Regarding the reliability metric  $\epsilon$ , none of the employed methods is constantly effective in improving forecasts. The value of  $\epsilon$  reflects the adequacy of the ensemble spread to encompass all the observations during the forecast period. Generally, neither QM nor copula-based methods are able to adjust the original forecasts so as to embrace all the observations within the ensemble range. The next metric ( $\pi$ ) measures the precision (sharpness) of the ensemble. The sharpness of the adjusted ensemble after QM application is higher than the original forecast ensemble; however, the reliability of QM corrections is less than the others. The subplots of  $\epsilon$  and  $\pi$  indicate that a large sharpness of the forecast ensemble after copula application is at the expense of missing some observations to be inside the ensemble spread, implying overconfidence of the ensemble prediction. The QM method, on the other hand, results in better

ensemble sharpness (precision) than that of the original forecast; however, as long as a specific method is steadily proved to be unreliable, comparing its precision with other methods is rather trivial and misleading. In other words, if an 'inaccurate' forecast ensemble has high 'precision', it cannot be accredited as a preferred forecast. Therefore, the evaluation of methods with respect to the sharpness metric should be done by first ensuring a satisfactory reliability of the methods.

For better understanding of the performance of the post-processing methods proposed in this study, the ensemble range and mean forecasts of monthly flow volumes for the forecast periods in 2002 are shown in Figure 13. As can be seen, the mean forecast after copula post-processor is close to the observation for all three observation periods while after QM application, the mean forecasts go even further away from the observed volumes. The MAE and NSE results shown in Figure 11 verify the close distance between the observations and the mean forecast after copula-based post-processing. Moreover, reliable and precise forecast after copula post-processor as expected from the probabilistic measures in

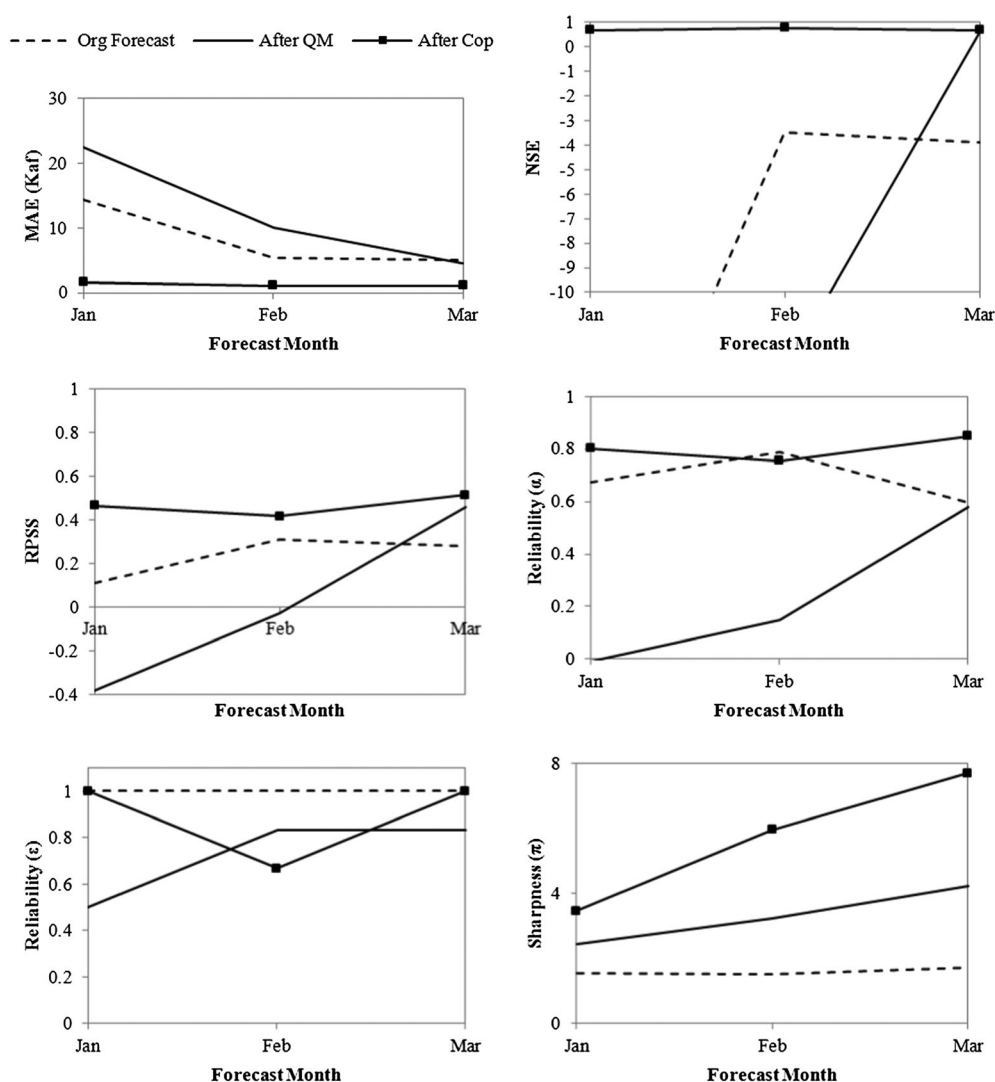


Figure 12. Comparing the performance of post-processing methods in adjusting the monthly streamflow hind-cast starting from different forecast dates in 2001. The forecast period of each forecast ensemble is 6 months

Figure 11 are reflected in the ensemble ranges. Also, from Figure 13, it can be seen that the error spread reduces significantly by the application of copula post-processor with the exception of few occasions where the observed volume falls outside the ensemble range after copula post-processing. The overall conclusion from Figure 13 is that the QM method is not an effective method to adjust the original forecasts while the multivariate copula-based post-processor is a more effective method that can be used operationally.

In general, a well-fitted copula function to the marginal distribution is a better choice than the QM method (especially in cases with large  $\gamma$ ). The results shown in Figures 10–13 also illustrate that the evaluation of different methods should not be merely based on the probabilistic metrics; they may be misleading if not being compared along with the point-wise measures.

## CONCLUSION

An auxiliary index, the so called failure index ( $\gamma$ ), was introduced to predict the overall performance of the post-processing methods before stepping into the forecast

mode. The failure index applied to the QM technique reflects the consistency of QM adjustments and corresponding observations; it varies between 0 and 1, with  $\gamma=0$  being the perfect-adjustment case. The forecast skill of QM shows that this statistical bias correction technique is not always successful in improving initial forecast trajectories. Testing 2500 hypothetical case studies indicates that the performance of the QM technique constantly degrades as  $\gamma$  increases. An alternative method based on the multivariate joint distribution was introduced for post-processing of forecasts. In the hypothetical case study, the proposed multivariate copula-based post-processor generally outperformed the QM technique with regard to forecast verification metrics; and unlike QM, it did not show specific sensitivity to  $\gamma$ . Post-processing of a real case study was also tested. Using a distributed parameter hydrologic model, PRMS, several ensembles of monthly streamflow forecasts of the Sprague River basin in southern Oregon were generated with a forecast horizon of 6 months. Generally, the forecast skill of the post-processed ensembles was effectively improved when the multivariate post-processor was applied, but it became

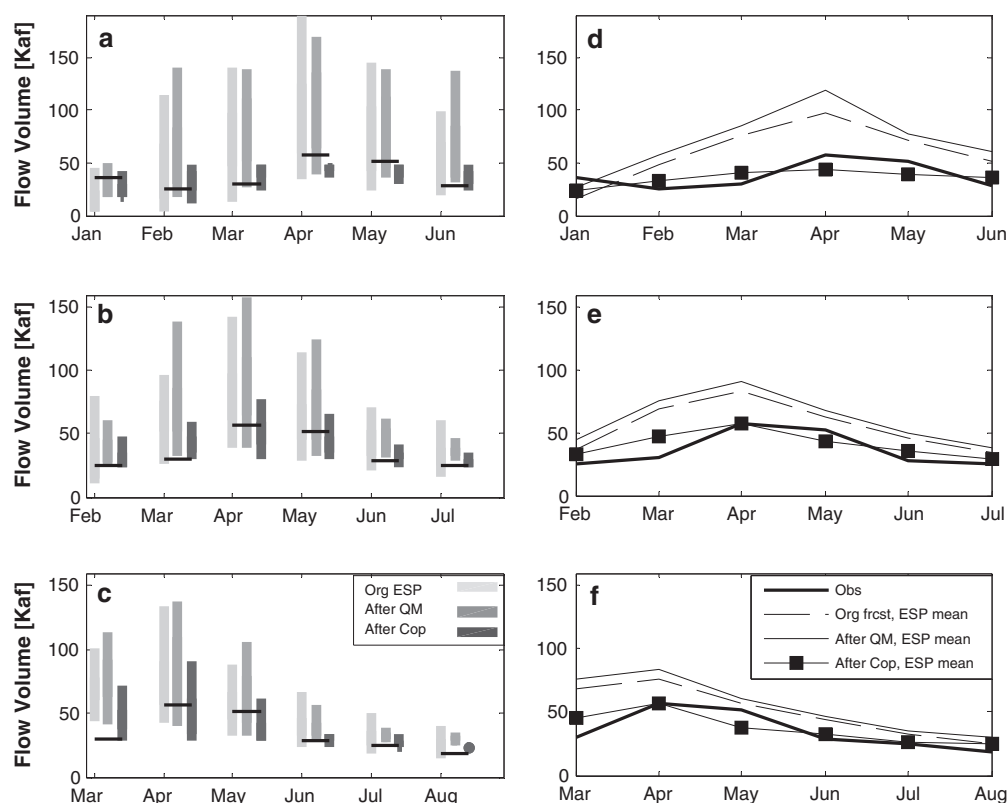


Figure 13. Comparison of the ensemble range before and after post-processing for three forecast periods in 2002 starting from a) Jan, b) Feb, and c) Mar, with the solid lines representing the monthly observations. Corresponding ESP mean are shown in subplots d–f

even worse when QM technique was used. The performance metrics indicated that QM was the dominated technique; however, weak performance of the QM technique was predictable from the failure ratio of the analysis period ( $\gamma=0.3$ ). Superiority of a multivariate copula-based method accounting for the joint behaviour of forecasts and observations as dependent variables was evidently demonstrated in the post-processing of hydrological forecasts. Complicated hydrological processes make it difficult to establish a multivariate function that directly models the dependent variables; therefore, a copula approach that combines the CDFs of variables instead of joining the raw variables showed a successful contribution in adjusting hydrological forecasts.

#### ACKNOWLEDGEMENTS

The first two authors would like to acknowledge the partial financial support for this research from USDA, NRCS Cooperative Agreement Number 68-7482-11-522, NOAA-MAPP, grant NA11OAR4310140, and NOAA-CSTAR, grant NA11NWS4680002.

#### REFERENCES

- Berg D. 2009. Copula goodness-of-fit testing: An overview and power comparison. *European Journal of Finance* **15**: 675–701.
- Biagorria GA, Jones JW, Shin DW, Mishra A, O'Brien JJ. 2007. Assessing uncertainties in crop model simulations using daily bias-corrected Regional Circulation Model outputs. *Climate Research* **34**: 211–222.
- Brookner E. 1998. *Tracking and Kalman Filtering Made Easy*. John Wiley & Sons: New York.
- Brown JD, Seo DJ. 2010. A Nonparametric Postprocessor for Bias Correction of Hydrometeorological and Hydrologic Ensemble Forecasts. *Journal of Hydrometeorology* **11**: 642–665. DOI: <http://dx.doi.org/10.1175/2009JHM1188.1>
- Brown JD, Seo DJ. 2012. Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions. *Hydrological Processes accepted article*. DOI: 10.1002/hyp.9263.
- Candille G, Beauregard S, Gagnon N. 2010. Bias Correction and Multiensemble in the NAEFS Context or How to Get a “Free Calibration” through a Multiensemble Approach. *Monthly Weather Review* **138**: 4268–4281.
- Cherubini U, Luciano E, Vecchiato W. 2004. *Copula methods in finance*. Wiley: New York.
- Clark MP, Hay LE. 2004. Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *Journal of Hydrometeorology* **5**(1): 15–32.
- Cui B, Toth Z, Zhu Y, Hou D. 2008. Statistical downscaling approach and its application, Preprints, 19th Conf. on Probability and Statistics, New Orleans, LA. *American Meteorological Society* **11**: 2.
- Day GN. 1985. Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management* **111**(2): 157–170.
- De Michele C, Salvadori G, Canossi M, Petaccia A, Rosso R. 2005. Bivariate statistical approach to check adequacy of dam spillway. *Journal of Hydrologic Engineering* **10**(1): 50–57.
- DeChant C, Moradkhani H. 2011. Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. *Hydrology and Earth System Sciences* **15**: 3399–3410. DOI: 10.5194/hess-15-3399.
- Diks C, Panchenko V, van Dijk D. 2010. Out-of-sample comparison of copula specifications in multivariate density forecasts. *Journal of Economic Dynamics and Control* **34**: 1596–1609.
- Djalalova I, Wilczak J, McKeen S, Grell G, Peckham S, Pagowski M, DelleMonache L, McQueen J, Tang Y, Lee P, McHenry J, Gong W, Bouchet V, Mathur R. 2010. Ensemble and bias-correction techniques for air quality model forecasts of surface O<sub>3</sub> and PM<sub>2.5</sub> during the

- TEXAQS-II experiment of 2006. *Atmospheric Environment* **44**(4): 445–467.
- Duan Q, Sorooshian S, Gupta VK. 1994. Optimal use of the SCE-UA global optimization method for calibrating watershed models. *Journal of Hydrology* **158**: 265–284.
- Dupuis DJ. 2007. Using Copulas in Hydrology: Benefits, Cautions, and Issues. *Journal of Hydrologic Engineering* **12**(4): 381–393.
- Embrechts P, Lindskog F, McNeil AJ. 2003. *Modelling dependence with copulas and applications to risk management*, Handbook of heavy tailed distributions in finance, Rachev ST (ed). Elsevier Science: Amsterdam, The Netherlands.
- Favre AC, Adlouni SE, Perreault L, Thiéonge N, Bobée B. 2004. Multivariate hydrological frequency analysis using copulas. *Water Resources Research* **40**: W01101. DOI: 10.1029/2003WR002456
- Genest C, Rémillard B. 2008. Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri-Poincaré* **44**(6): 1096–1127.
- Genest C, Quessy JF, Rémillard B. 2006. Goodness-of-fit procedures for copula models based on the probability integral transform. *Scandinavian Journal of Statistics* **33**: 337–366.
- Genest C, Rémillard B, Beaudoin D. 2009. Goodness-of-fit tests for copulas: A review and a power study, Insurance: *Mathematics and Economics* **44**: 199–213.
- Hashino T, Bradley AA, Schwartz SS. 2006. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrology and Earth System Sciences Discussions* **3**: 561–594.
- Hay LE, Umemoto M. 2006. Multiple-objective stepwise calibration using Luca, U.S. Geological Survey Open-File Report, 2006–1323, 25.
- Ines AVM, Hansen JW. 2006. Bias correction of daily GCM rainfall for crop simulation studies. *Agricultural and Forest Meteorology* **138**: 44–53.
- Isaaks EH, Srivastava RM. 1989. *An Introduction to Applied Geostatistics*. Oxford University Press: New York; 561.
- Joe H. 1997. *Multivariate Models and Dependence Concepts*. Chapman & Hall: London.
- Kao S, Govindaraju RS. 2008. Trivariate statistical analysis of extreme rainfall events via Plackett family of copulas. *Water Resources Research* **44**: W02415. DOI: 10.1029/2007WR006261
- Kao S, Govindaraju RS. 2010. A copula-based joint deficit index for droughts. *Journal of Hydrology* **380**: 121–134.
- Kelly KS, Krzysztofowicz R. 1997. A bivariate meta-Gaussian density for use in hydrology. *Stochastic Hydrology and Hydraulics* **11**: 17–31.
- Klamath Basin Ecosystem Foundation (KBEF). 2007. *Upper Sprague Watershed Assessment*. Klamath Falls: OR.
- Krzysztofowicz R. 1997. Transformation and normalization of variates with specified distributions. *Journal of Hydrology* **197**: 286–292.
- Laio F, Tamea S. 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences* **11**: 1267–1277. DOI: 10.5194/hess-11-1267-2007
- Leavesley GH, Lichty RW, Troutman BM, Saindon LG. 1983. *Precipitation-Runoff Modeling System: User's Manual*. U.S. Geological Survey Water-Resources Investigations 83–4238, 207.
- Li H, Luo L, Wood EF, Schaake J. 2009. The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *Journal of Geophysical Research* **114**(D4): D04114.
- Madadgar S, Moradkhani H. 2011. Improving the Ensemble Streamflow Prediction by Adjusting Hydrologic Ensemble Traces. In *the proceedings of ASCE World Environmental and Water Resources Congress 2011*. Palm Spring: CA.
- Madadgar S, Moradkhani H. 2012. Drought Analysis under Climate Change Using Copula. *Journal of Hydrologic Engineering*. DOI: 10.1061/(ASCE)HE.1943-5584.0000532
- McEnery J, Ingram J, Duan QY, Adams T, Anderson L. 2005. NOAA's advanced hydrologic prediction service, Building pathways for better science in water forecasting. *Bulletin of the American Meteorological Society* **86**: 375–385.
- Moradkhani H, Meier M. 2010. Long-Lead Water Supply Forecast using Large-scale Climate Predictors and Independent Component Analysis. *Journal of Hydrologic Engineering* **15**(10): 744–762. DOI.org/10.1061/(ASCE)HE.1943-5584.0000246
- Moradkhani H, Sorooshian S. 2008. General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis, in *Hydrological Modeling and Water Cycle, Coupling of the Atmospheric and Hydrological Models*. Springer, *Water Science and Technology Library* **63**: 1–23.
- Moradkhani H, DeChant CM, Sorooshian S. in review. Evolution of Ensemble Data Assimilation for Uncertainty Quantification using the Particle Filter-Markov Chain Monte Carlo Method. *Water Resources Research*.
- Najafi MR, Moradkhani H, Piechota T. 2012. Ensemble Streamflow Prediction: Climate Signal Weighting vs. Climate Forecast System Reanalysis. *Journal of Hydrology* **442–443**: 105–116. <http://dx.doi.org/10.1016/j.jhydrol.2012.04.003>.
- Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models part I- a discussion of principles. *Journal of Hydrology* **10**(3): 282–290.
- Nelsen RB. 1999. *An Introduction to Copulas*. Springer: New York.
- Olsson J, Lindström G. 2008. Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. *Journal of Hydrology* **350**: 14–24.
- Parrish M, Moradkhani H, DeChant CM. 2012. Towards reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation. *Water Resources Research* **48**: W03519, DOI: 10.1029/2011WR011116.
- Piani C, Haerter JO, Coppola E. 2010. Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology* **99**: 187–192.
- Renard B, Kavetski D, Kuczera G, Thyer M, Franks SW. 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research* **46**: W05521. DOI: 10.1029/2009WR008328
- Rosenblatt M. 1952. Remarks on a multivariate transformation. *Annals of Mathematical Statistics* **23**: 470–472.
- Roulin E, Vannitsem S. 2005. Skill of medium-range hydrological ensemble predictions. *Journal of Hydrometeorology* **6**(5): 729–744.
- Salvadori G, De Michele C. 2006. Statistical characterization of temporal structure of storms. *Advances in Water Resources* **29**(6): 827–842.
- Salvadori G, De Michele C. 2010. Multivariate multiparameter extreme value models and return periods: A copula approach. *Water Resources Research* **46**: W10501. DOI: 10.1029/2009WR009040
- Savu C, Trede M. 2008. Goodness-of-fit tests for parametric families of Archimedean copulas. *Quantitative Finance* **8**: 109–116.
- Schaake JC, Demargne J, Hartman R, Mullusky M, Welles E, Wu L, Herr H, Fan X, Seo D-J. 2007. Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrology and Earth System Sciences Discussions* **4**: 655–717.
- Scheweppe FC. 1973. *Uncertain Dynamic Systems*. Englewood Cliffs, Prentice-Hall: New Jersey, USA; 576.
- Shiau JT. 2006. Fitting drought duration and severity with two-dimensional copulas. *Water Resources Management* **20**(5): 795–815.
- Sklar A. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de L'Institut de Statistique de L'Université de Paris* **8**: 229–231.
- Smith JA, Day GN, Kane MD. 1992. Nonparametric framework for long-range streamflow forecasting. *Journal of Water Resources Planning and Management* **118**: 82–91.
- Thirel G, Rousset-Regimbeau F, Martin E, Habets F. 2008. On the impact of short-range meteorological forecasts for ensemble streamflow predictions. *Journal of Hydrometeorology* **9**(6): 1301–1317.
- Thyer M, Renard B, Kavetski D, Kuczera G, Franks SW, Srikanthan S. 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research* **45**: W00B14. DOI: 10.1029/2008WR006825
- Todini E. 2008. A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management* **6**(2): 123–137.
- Twedt TM, Schaake JC, Peck EL. 1977. National Weather Service extended streamflow prediction, In: Proc. 45th Western Snow Conference, Albuquerque 52–57.
- Weiß GNF. 2011. Are Copula-GoF-tests of any practical use? Empirical evidence for stocks, commodities and FX futures. *The Quarterly Review of Economics and Finance* **51**: 173–188.
- Wilks DS. 2006. *Statistical Methods in the Atmospheric Sciences*, 2nd edn. Elsevier Academic Press: Amsterdam.
- Wong G, Lambert MF, Leonard M, Metcalfe AV. 2010. Drought Analysis Using Trivariate Copulas Conditional on Climatic States. *Journal of Hydrologic Engineering* **15**(2): 129–141.
- Wood AW, Lettenmaier DP. 2006. A test bed for new seasonal hydrologic forecasting approaches in the western United States. *Bulletin of the American Meteorological Society* **87**(12): 1699–1712.

- Wood AW, Lettenmaier DP. 2008. An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophysical Research Letters* **35**: L14401. DOI: 10.1029/2008GL034648
- Zhang L, Singh VP. 2007a. Trivariate Flood Frequency Analysis Using the Gumbel–Hougaard Copula. *Journal of Hydrologic Engineering* **12**(4): 431–439.
- Zhang L, Singh VP. 2007b. Gumbel–Hougaard Copula for Trivariate Rainfall Frequency Analysis. *Journal of Hydrologic Engineering* **12**(4): 409–419.
- Zhao L, Duan Q, Schaake J, Ye A, Xia J. 2011. A hydrologic post-processor for ensemble streamflow predictions. *Advances in Geosciences* **29**: 51–59.