Probabilistic Assessment of Monthly River Flow Discharge Using Copula And OSVR Approaches

Mohammad Nazeri Tahroudi (m_nazeri2007@yahoo.com)
University of Birjand https://orcid.org/0000-0002-6871-2771
Rasoul Mirabbasi
Shahrekord University
Yousef Ramezani
University of Birjand
Farshad Ahmadi
Shahid Chamran University of Ahvaz

Research Article

Keywords: ARCH Models, ARMA, Conditional heteroscedasticity, Copula-GARCH, Rainfall- Runoff Modeling

Posted Date: June 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-440240/v1

Probabilistic assessment of monthly river flow discharge using copula and OSVR approaches

Mohammad Nazeri Tahroudi[1*], Rasoul Mirabbasi [2], Yousef Ramezani[3], Farshad Ahmadi[4]

1. Ph.D Candidate, Department of Water Engineering, University of Birjand, Birjand, Iran. Email: m_nazeri2007@yahoo.com, m_nazeri2007@birjand.ac.ir

2. Associate Professor, Department of Water Engineering, Shahrekord University, Shahrekord, Iran, Email: mirabbasi@agr.sku.ac.ir

3. Associate Professor, Department of Water Engineering, University of Birjand, Birjand, Iran. Email: y.ramezani@birjand.ac.ir

4-Assistant Professor, Department of Hydrology and Water Management, Shahid Chamran University of Ahvaz, Ahvaz, Iran. Email: f.ahmadi@scu.ac.ir

* Corresponding Author

Probabilistic assessment of monthly river flow discharge using copula and OSVR approaches

Abstract

Simulation of flow discharge based on monthly precipitation values as inputs is one of the important issues in hydrology and water resources studies, especially in areas where data with the shorter time scales are not available. In this study, the applicability of support vector regression (SVR) model optimized by Ant colony and Copula-GARCH algorithms was investigated and compared to simulate the flow discharge based on total monthly rainfall in Talezang Basin, Iran. Entropy theory was used to select a suitable meteorological station corresponding to a hydrometric station. The vector autoregressive model was also used as the base model in Copula-GARCH simulations. The correlation results of the studied paired variable confirmed the possibility of using copula-based models. The simulation results were evaluated using R^2, Nash-Sutcliffe Efficiency (NSE) and root mean square error (RMSE) statistics. According to the 99% confidence intervals of the simulations, the accuracy of both models was confirmed. The simulation results showed that the Copula-GARCH model was more accurate than the optimized SVR (OSVR) model. Considering the 90% efficiency (NSE = 0.90) of Copula-GARCH approach, the results show a 36% improvement of RMSE statistics by Copula-GARCH model compared to OSVR model in simulating the flow discharge on a monthly scale. The results also showed that by combining nonlinear ARCH models with the copula-based simulations, the reliability of the simulation results increases, which was also confirmed using the violin plot. The results also showed an increase in the accuracy of the Copula-GARCH model at the minimum and maximum values of the data.

Keywords: ARCH Models, ARMA, Conditional heteroscedasticity, Copula-GARCH, Rainfall-Runoff Modeling.

1. Introduction

River flow discharge as an input to water resources systems is one of the important parameters in managing the operation of dam reservoirs. Simulation and forecasting the annual discharge can be done using stochastic methods. On a monthly scale, which is more important in dam management, due to the seasonal and periodic approach of stochastic methods, the number of parameters increases and as a result, the estimation error will be increased. Therefore, by including other effective parameters in discharge values, the accuracy of the simulations can be increased. Accurate calculation of flow discharge and flood is a special priority for many river engineering and flood control projects. On the other hand, various studies in different parts of Iran indicate the occurrence of climate change and changes in meteorological parameters in recent decades (Ramezani and Tahroudi, 2020). This will also have a more impact on the flow discharge. Meteorological studies also show that changes in precipitation pattern in different parts of Iran are irregular. These changes have increased the frequency of extreme events, especially precipitation and it led to increase the maximum runoff occurrences. Since the design of most dams in the country does not take into account the conditions of climate change and the occurrence of such rainfalls, usually the runoff from these rainfalls is not stored and becomes unavailable as floods (Ramezani et al. 2020; Tahroudi et al. 2019 a & b; Khozeymehnezhad and Tahroudi, 2019). River flow discharge is also affected by base flow and rainfall in each region. Simulation of flow discharge given by rainfall will greatly assist in the water resources management and flood control in catchments. In addition to the importance of modeling and simulations in estimation of flow discharge, sometimes different simulation models are used to overcome the problem of data shortage or lack of data on a short time scale. One of the most widely approaches in this field is using the linear time series models. One of the most important models of this family is vector

autoregressive (VAR) models that are able to model several time series with different delays. Using these models, the effect of different variables on each other or the simultaneous effect of a particular variable in several stations can be examined.

The VAR model is one of the linear time series models that can be combined with nonlinear models and the residuals of these models can be considered and modeled. The most important nonlinear models are autoregressive conditional heteroscedasticity (ARCH) family models such as generalized autoregressive conditional heteroskedasticity (GARCH) model. These models have a high ability to connect with the residuals of the linear models. In most cases, while combining with linear models, they have increased the accuracy of calculations. Recent developments in the use of copula functions and their integration with non-linear ARCH models have led to new research that the effect of various parameters can be considered in terms of multivariate analysis based on the copula function. Copulas are functions that create a multivariate distribution function by connecting univariate margin functions (Nelsen, 2006; Sklar, 1959). Separate analysis of marginal distributions and their dependence structure is one of the most important advantages of copula functions (Serinaldi et al. 2009; Salvadori et al. 2007; Ramezani et al. 2019). In recent years, the use of hybrid models and copula based models for simulations has attracted the attention of researchers. Using the Copula-GARCH approach, Yoo et al. (2016) analyzed and simulated precipitation values at 12 rain gauge stations in South Korea. The results showed that the proposed method will be very effective in quantifying the uncertainty of bivariate drought frequency curves. Various studies can also be mentioned in this field (Li and Zheng, 2016; Guo and Wang, 2017; Abdi et al. 2017; Ayantobo et al. 2018; Kim et al. 2019). Yuan et al (2020) used the Copula-GARCH approach to study price fluctuations in agricultural commodities. Their results showed that the proposed approach has a high ability to describe the interaction of parameters with each

other. Tahroudi et al. (2020a) investigated the changes and joint frequency analysis of meteorological and hydrological droughts in the Zarrineh-Rud catchment in the south of Lake Urmia in Iran using copula functions. The results showed that by using copula functions, for a specific meteorological drought duration in a station, the duration of hydrological drought in an existing hydrometric station can be determined based on the probability of conditional occurrence as well as certain return periods. Tahroudi et al. (2020b) analyzed the conditional and joint behavior of groundwater level deficiency and rainfall deficiency signature in the Naqadeh sub-basin located in the Lake Urmia Basin in Iran using copula functions. Considering the maximum groundwater level deficiency produced, the relationship between rainfall deficiency and groundwater level deficiency was obtained to estimate the groundwater level deficiency signature values. In addition to the proposed approaches that have been used recently, there are various other models in this field, which can be referred to as support vector machine models, regression-based models, genetic programing, etc. The optimized support vector regression (OSVR) method has high efficiency in simulating meteorological and hydrological parameters, which has been mentioned in the studies of Tahroudi and Ramezani (2020). Optimizing the parameters of the support vector regression method using various algorithms such as the ant colony improves the performance and accuracy of the model in the simulations. Nazeri Tahroudi and Ramezani (2020) simulate the dew point temperature values in different climates of Iran by optimizing the parameters of the support vector regression model. Copula-based simulation and simulation based on OSVR are both optimized and enhanced models, that the first model can be upgraded using the GARCH nonlinear model and the second one can be upgraded using optimization algorithms such as the ant colony optimization (ACO) algorithm. With the simulation approach based on the Copula-GARCH hybrid model, the values obtained from the bivariate frequency curve will be

more reliable. The Copula-GARCH model integrates conditional variability dependencies between variables in the simulation. Also, due to the existence of heteroscedasticity in hydrological time series, including rainfall (Wang et al. 2005; Modarres and Ouarda, 2012; Yusof and Kane, 2012), the Copula-GARCH hybrid model will be suitable for simulating the rainfall time series. Accordingly, in this study, while examining the performance of Copula-GARCH and OSVR models, a simulation of mean monthly discharge (MMD) given by monthly rainfall (MR) is performed. In addition, an attempt based on the optimized entropy theory was used to select the rain gauge station corresponding to the hydrometric station.

2. Material and Methods

2.1 Case Study

Dez basin is located between two meridians of eastern 48°,10' and 50°, 21' and two orbits of northern 31°, 34' and 34°, 7' in the western regions of Iran. The area of this basin is equal to 21720 square kilometers and have an average elevation 1600 meters above mean sea level. In this study, Copula-GARCH and OSVR models were used for modeling the monthly rainfall-runoff in the mentioned basin. The data used belong to the Dez basin at the Talezang hydrometric station in the period 1988- 2018. Fig 1 shows the location of the studied basin and its hydrometric and rain gauge stations. The summery of statistical characteristics of the studied data are also presented in Table 1.

Fig. 1.

Table 1.

2.2 Kendall's tau correlation

The first step in copula-based analysis is to examine the dependence structure between the studied variables. The usual method for examining correlation is the Kendall's tau correlation coefficient. The Kendall's tau, denoted by τ, is defined as the probability of concordance minus the probability of discordance between the two paired random variables X_1 and X_2. The Kendall's tau value between the continuous random variables X_1 and X_2 is defined as Equation (1):

$$\tau(X_1, X_2) = P((X_{11} - X_{21})(X_{12} - X_{22}) > 0) - P((X_{11} - X_{21})(X_{12} - X_{22}) < 0) \quad (1)$$

where, $(X_{11} - X_{12})$ and $(X_{21} - X_{22})$ are independent and regular distributed differences of (X_1, X_2) (Hollander et al. 2014).

2.3 Vector Autoregressive (VAR) Models

VAR is one of the flexible models for analyzing the multivariate time series. The VAR model was introduced to describe the dynamic behavior of the economic and financial series and their prediction. VAR model predictions are quite flexible, since they can bet on the future path of potential variables. In addition to describing and forecasting data, the VAR model is also used for structural inferences and analyzing the policies. In structural analysis, specific hypotheses are imposed on the structure of the data under investigation, and the effects of unexpected shocks or innovations are summed up with the model variables. These effects are usually summarized with impact reaction and predicted error variance analysis functions. This model focuses on the analysis of constant multivariate covariance. VAR models in economic analysis have been introduced by Sims (1980). In this study, the base model was selected to create the Copula-GARCH model, is a VAR model. This model has been selected according to the number of studied parameters and their dependencies. For more information, see Shahidi et al. (2020).

153

154 **2.4 Autoregressive model with conditional heteroskedasticity**

155 The GARCH model was developed as an alternative to time series models that, based on the
156 assumption of linearity between variables at different time stages, could not take into account the
157 conditional dependence of the variance or heteroskedasticity. Because of its power in modeling
158 variables in which changes are significant, it has been widely used, especially in the financial and
159 economic fields (Duan, 1996; Tse and Tsui, 2002; Floros et al. 2007; Watanabe, 2012). The
160 GARCH model has many applications in hydrological time series simulation (Wang et al. 2005;
161 Modarres and Ouarda, 2012; Yusef and Kane, 2012; Ramezani et al. 2019). In this paper, the
162 GARCH model is developed to detect heteroskedasticity in the flow discharge time series. The
163 parameters of the GARCH model were estimated using the maximum likelihood method.

164 The GARCH model is actually a generalized model of the ARCH family. This model was first
165 introduced and developed in economic studies by Engle (1982). All different modes of this model
166 are based on the ARCH model. In general, the ARCH model is as follows:

$$\varepsilon_t = \sigma_t z_t \quad and \quad \sigma_t^2 = a_0 + \sum_{i=1}^{m} b_i \varepsilon_{t-i}^2 \tag{2}$$

167 where, σ_t^2 is the conditional variance, ε_t is the error term or the remainder of the model with
168 zero mean and variance of 1, $a_0 \geq 0, b_i \geq 0$ are the model parameters, *m* is equal to the order of
169 model, and Z_t is also the time series of the desired variable (Engle, 1982).

170 To better understand the model, the structure of the ARCH model was considered here.

$$a_t = \sigma_t \varepsilon_t, \ \sigma_t^2 = a_0 + a_1 a_{t-1}^2 \tag{3}$$

171 where, $a_1 \geq 0, a_0 \geq 0$. First of all, the conditional mean a_t must be zero. Because:

$$E(a_t) = E[E(a_t \mid F_{t-1})] = E[\sigma_t E(\varepsilon_t)] \tag{4}$$

Then the conditional variance is obtained from the following equation:

$$Var(a_t) = E(a_t^2) = E[E(a_t^2 \mid F_{t-1})] = E[a_0 + a_1 a_{t-1}^2] = a_0 + a_1 E(a_{t-1}^2) \quad (5)$$

Since, according to $E(a_t = 0)$ and $Var(a_t) = E(a_{t-1}) = E(a_{t-1}^2)$, a_t is a static and trend free, we will have:

$$Var(a_t) = a_0 + a_1 Var(a_t) \quad (6)$$

$$Var(a_t) = \frac{a_0}{(1-(a_0))} \quad (7)$$

Since the variance of a_t should be positive, thus the range of a_1 should be between 0 and 1. In some applications, above values (a_t) should also exist and so, α_1 should provide some extra moments. For example, in studying the behavior of sequences, it is necessary to limit the fourth moment (a_t). Assuming that ε_t is normal, we will have the following equation (Shahidi et al. 2020):

$$E[E(a_t^4 \mid F_{t-1})] = 3[E(a_t^2 \mid F_{t-1})]^2 = 3E(a_0 + a_1 a_{t-1}^2)^2 \quad (8)$$

Therefore:

$$E(a_t^4) = E[E(a_t^4 \mid F_{t-1})] = 3E(a_0 + a_1 a_{t-1}^2)^2 = 3E(a_0^2 + 2a_0 a_1^2 a_{t-1}^2 + a_1^2 a_{t-1}^4) \quad (9)$$

If a_t is considered as the fourth moment and $m_4 = E(a_t^4)$, then:

$$m_4 = 3E(a_0^2 + 2a_0 a_1 Var(a_t) + a_1^2 m_4) = 3a_0^2(1 + 2\frac{a_1}{1-a_1}) + 3a_1^2 m_4 \quad (10)$$

Eventually:

$$m_4 = \frac{3a_0^2(1+a_1)}{(1-a_1)(1-3a_1^2)} \quad (11)$$

Although the ARCH model is simple, it often requires a lot of parameters to obtain the proper modeling process. For this reason, we have to look for alternative models (Moffat et al. 2017). Bollerslev (1992) proposed the developed ARCH model as follows:

$$a_t = \sigma_t e_t$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \alpha_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \quad (12)$$

where, e_t is a random series with zero mean and variance of one. Also, the EGARCH model is a natural logarithmic model of the GARCH model, which was presented by Nelson (1991).

2.5 Copula functions and Sklar theorem

Copulas are a flexible way to create joint distributions with different type of margins. Copulas are multivariate distribution functions whose one-dimensional margins are uniform over the range (0,1). A Copula definition is attributed to Sklar (1959) who describes in a theory how univariate distribution functions can be combined in the form of multivariate distributions. Sklar showed that for continuous d dimension random variables $\{X_1,....,X_d\}$ with $u_j=F_{X_j}(x_j)$ marginal CDFs that $j=1,...,d$, there is a d-dimensional copula of $C_{U_1,...,U_d}$ such that:

$$C_{U_1,\dots,U_D}(U_1,\dots,U_D) = H_{X_1,\dots,X_D}(X_1,\dots,X_D) \quad (13)$$

where, u_j is j^{th} margin and $H_{X_1,...,X_d}$ is the joint CDF of $H_{X_1,...,X_d}$. Because for continuous random variables, the CDF function of the margins are non-decreasing from 0 to 1, the $C_{U_1,...,U_d}$ copula can be considered as a $H_{X_1,...,X_d}$ transformation from $[-\infty,\infty]^d$ to $[0,1]^d$. The result of this transformation is that marginal distributions are separated from $H_{X_1,...,X_d}$ and, therefore, $C_{U_1,...,U_d}$ reflects only the relationship between the variables and provide a complete description of the overall dependence

structure (Nelsen, 2006). The selection of the copula function first depends on determining the degree of dependence of the studied paired variables. In the present study, for estimating the copula dependence parameter, in this study, the method of inference functions for margins (IFM), which is the most common method for estimating the copula parameters, was used (Joe, 1997).

In the case of two variables, it was assumed that the two correlated random variables X and Y were distributed as functions $f_X(x;\alpha_1,\alpha_2,\mathrm{K},\alpha_p)$ and $f_Y(y;\lambda_1,\lambda_2,\mathrm{K},\lambda_r)$, respectively, that $\alpha_1,\alpha_2,\mathrm{K},\alpha_p$ are the parameters of $f_X(x)$ and $\lambda_1,\lambda_2,\mathrm{K},\lambda_r$ are the parameters of $f_Y(y)$. The actual number of parameters depends on the type of univariate margin distributions. For n pairs of independent observations, the likelihood logarithmic functions for X and Y, i.e. $\ln L_X(x;\alpha_1,\alpha_2,\mathrm{K},\alpha_p)$ and $\ln L_Y(y;\lambda_1,\lambda_2,\mathrm{K},\lambda_r)$, were maximized separately to estimate the parameters. $\hat{\alpha}_1,\hat{\alpha}_2,\mathrm{K},\hat{\alpha}_p$ and $\hat{\lambda}_1,\hat{\lambda}_2,\mathrm{K},\hat{\lambda}_r$ are the estimated parameters. The log-likelihood function of the joint probability density function of $f_{X,Y}(x,y)$ was considered as follows:

$$Ln\,L(x,y;\hat{\alpha}_1,\hat{\alpha}_2,\ldots,\hat{\alpha}_p,\hat{\gamma}_1,\hat{\gamma}_2,\ldots,\hat{\gamma}_r,\theta) = \quad (14)$$

$$Ln\,L_C(x,y;F_x(x),F_y(y),\theta) + Ln\,L_X(x;\hat{\alpha}_1,\hat{\alpha}_2,\ldots,\hat{\alpha}_p) + Ln\,L_Y(y;\hat{\gamma}_1,\hat{\gamma}_2,\ldots,\hat{\gamma}_r)$$

where, $\ln L_C$ is the log-likelihood function of the copula density. By replacing the estimated values for $\hat{\gamma}_1,\hat{\gamma}_2,...,\hat{\gamma}_r$ and $\hat{\alpha}_1,\hat{\alpha}_2,\mathrm{K},\hat{\alpha}_p$ in Equation (14), the log-likelihood function is maximized to obtain the estimated dependence parameter $\hat{\theta}$. Then, by comparing the results of each copula with the results of the empirical probability, the appropriate copula is selected for the paired variables. Copula-based simulations were also first discussed in the research of Bedford and Cook (2001 and 2002). To obtain the sample $u_1,\ldots,u_d$ from a d variable copula, the following method is performed:

$$w_j \overset{i,i,d}{:} U[0;1],\ j=1,...,d \tag{15}$$

Then;

$$\begin{aligned} u_1 &= w_1 \\ u_2 &:= C_{2|1}^{-1}(w_2 \mid u_1) \\ &M \\ u_d &:= C_{d|d-1,...,1}^{-1}(w_d \mid u_{d-1},...,u_1) \end{aligned} \tag{16}$$

To determine the conditional distribution functions $C_{j|j-1,...,1},\ j=1,...,d$ required for the pair copulas structure, the recursive relationship for the conditional distribution function with the h function is used. For a two-variable copula $C_{ij}(u_i, u_i; \theta_{ij})$ with parameter θ_{ij}, the h functions are defined as follows:

$$h_{i|j}(u_i \mid u_j; \theta_{ij}) := \frac{\partial}{\partial u_j} C_{ij}(u_i, u_j; \theta_{ij}) \tag{17}$$

$$h_{j|i}(u_j \mid u_i; \theta_{ij}) := \frac{\partial}{\partial u_i} C_{ij}(u_i, u_j; \theta_{ij}) \tag{18}$$

2.6 Entropy Theory

The studied basin in this research is the Dez basin, which has 11 rain gauge stations with a sufficient data length. In the present study, the entropy theory was used to select the most suitable rain gauge station that reflects the general characteristics of the studied basin and its information can be generalized to the whole basin. See Tahroudi et al (2019c) for more information on entropy theory and how stations are selected and ranked.

2.7 Optimized Support Vector Regression (OSVR)

In this study, the SVR optimized by the ant colony optimization algorithm was used in two cases. In the first case, it used to investigate the interaction effect of the studied stations in entropy theory and the second case it used to simulate mean monthly discharge (MMD) values affected by monthly rainfall (MR). The first application of this method in water problems was presented by Dibike et al. (2001) for rainfall-runoff modeling (Nazeri-Tahroudi and Ramezani, 2020). SVR is an efficient learning system based on a constrained optimization theory which is used from the principle of induction the minimization of structural error, and leads to a general optimal answer. In the SVM regression model, a function related to the dependent variable y, which is a function of several independent variables x, is estimated. Similar to other regression models, it assumes that the relation between independent and dependent variables was determined by an algebraic function such as f (x) plus some amount of confusion (allowed error ε):

$$f\left(x\right)=W^{T}.\phi\left(x\right)+b \tag{19}$$

$$y=f\left(x\right)+noise \tag{20}$$

If W (vector of coefficients) and b (constant) are the properties of the regression function and ϕ is also the kernel function, then the object is finding the functional form for f(x). This topic is achieved by training the SVM model through a set of samples (training set). Therefore, to calculate w and b, it is necessary to optimize the error function in the ε *-SVM* model by considering the conditions listed in Equation 22.

$$\frac{1}{2}W^{T}W+C\sum_{i=1}^{N}\xi_{i}+C\sum_{i=1}^{N}\xi_{i}^{*} \tag{21}$$

$$
\begin{aligned}
& W^T.\phi(x_i)+b-y_i \le \varepsilon+\xi_i^* \\
& y_i-W^T.\phi(x_i)-b \le \varepsilon+\xi_i \\
& \xi_i,\xi_i^* \ge 0 \quad , \quad i=1,...,N
\end{aligned} \tag{22}
$$

where, C is a positive integer that determines the penalty when the model training error occurs. ϕ is the kernel function, N is the number of samples, and the two indices of ξ_i and ξ_i^* are the Slack variables, which determine the upper and lower limit of the training error associated with the allowed error value ε. In problems, it is predicted that the data is within the boundary range ε. Now, if the data is out of range ε, then there will be an error equal to ξ_i and ξ_i^*. It is worth mentioning that the SVM model solves the problems caused by the under fitting and over fitting by simultaneously minimizing two terms $W^T W/2$ and training errors, namely $C\sum_{i=1}^{N}(\xi_i+\xi_i^*)$, in Equation 23. Therefore, by introducing 2 Lagrange coefficients a_i and a_i^*, the optimization problem will be solved with the numerical maximization of the following quadratic function:

$$
\begin{aligned}
& \sum_{i=1}^{N} y_i(a_i+a_i^*)-\varepsilon\sum_{i=1}^{N}(a_i+a_i^*)- \\
& 0.5\sum_{i,j=1}^{N}(a_i+a_i^*)(a_j+a_j^*)\phi(x_i)^T\phi(x_j)
\end{aligned} \tag{23}
$$

$$
\begin{aligned}
& \sum_{i=1}^{N}(a_i+a_i^*)=0 \\
& 0\le a_i \le C,\ 0\le a_i^* \le C,\ i=1,2,...,N
\end{aligned} \tag{24}
$$

The above objective function in Equation 25 is a convex function and therefore, the solution of Equation 25 would be unique and optimal. After defining the Lagrange coefficients in Equation

25, the characteristics w and b in the SVM regression model is calculated using the Karush-Kan-Tucker theory, where $W = \sum_{j=1}^{N} (a_i + a_i^*)\phi(x_i)$. As a result, for the SVM regression model there is:

$$W = \sum_{i=1}^{N} (a_i + a_i^*)\phi(x_i)^T \phi(\mathrm{x}) + b \tag{25}$$

It should be noted that the Lagrange terms ($(a_i + a_i^*)$) can be zero or non-zero. Therefore, only data sets whose coefficients $\overline{a_i}$ are non-zero are entered in the final regression equation and this data set is known as the support vectors. In simple terms, support vectors are data that help to create a regression function. Among the vectors mentioned, those whose $\left|\overline{a_i}\right|$ values are less than C are called margin support vectors. When the value $\left|\overline{a_i}\right|$ of the support vectors is equal to C, it is known as an error support vector or a bounded support vector. Margin support vectors are found on the margin of the insensitive boundary while error support vectors are out of range. Finally, the regression SVM function can be rewritten in the following form:

$$f(x) = \sum_{i=1}^{N} a_i \phi(x_i)^T \phi(x_j) + b \tag{26}$$

In Equation 26, the calculation of $\phi(x)$ in its characteristic space may be very complicated. To solve this problem, the regular approach in the SVM regression model is the selection of a kernel function as $K(x_i, x) = \phi(x_i)^T \phi\sqrt{b^2 - 4ac}$. Various kernel functions can be used to construct different types of ε-SVM models. The most commonly used kernel functions available in the vector regression model are: A polynomial kernel with 3 target characteristics, a sigmoid kernel containing 2 target characteristics and kernel of radial base functions (RBF) with a target characteristic.

2.8 Ant colony optimization algorithm (ACO)

The ant colony optimization (ACO) algorithm was first proposed by *Colorni et al. (1991)*. One of the first applications of the ACO algorithm has been to solve the traveling salesman problem. Since the ACO algorithms depend on the type of use and similarity of the ants moving on the graph, the use of the traveling salesman problem to explain the basic principles of ant algorithms was highly logical, and it was originally a typical example for introducing this algorithm. For more information, see (Colorni et al. 1991, Dorigo, 1992; Nazeri-Tahroudi and Ramezani, 2020).

3. Results and Discussion

There are 11 rain gauge stations in the Dez basin with recorded data in the period of 1988-2018. In this study, entropy theory was used to select the best and most valuable rain gauge station that clearly shows the characteristics of the studied basin. By applying the entropy theory and calculating its indicators, the studied stations were ranked. The results of ranking and information transfer index (ITI) of the studied stations are presented in Table 2. The results of the study and ranking of the studied stations based on the N(i) index of entropy theory showed that Vanaei station is the best and most valuable rain gauge station in the region that the rainfall values of this station can be generalized to the whole basin area. It should be noted that the support vector regression method optimized with the ant colony optimization (ACO) algorithm was used to evaluate the interaction of rain gauge stations (Tahroudi et al. 2019c). The results showed that the information transfer of Vanaei station is established with other stations and is related to all existing stations. By selecting Vanaei station as the representative station, the correlation between the MMD values of the Talezang hydrometric station and the MR values of Vanaei station was investigated.

Table 2.

The results of evaluating the Information Transfer Index (ITI) also showed that, except of the two stations Kamandan and Sepidasht Sezar, other studied rain gauge stations are located in the surplus area in the basin in terms of concentration of stations. In this study, in order to simulate the MMD values given by the occurrence of MR values, the performance of ARCH-based models (Copula-GARCH models) and the OSVR-based model was investigated and compared. First, the correlation between the mentioned paired variables was investigated. The scatter diagram of the studied data at the real scale is presented in Figure 2. In this figure, the empirical contour lines, histograms of the studied data and Kendall's tau coefficient are presented. The value of the Kendall's tau statistic that reflect the correlation between the studied variables is 0.34 which is acceptable for further analysis.

Fig. 2.

3.1 Simulation of MMD values of Talezang hydrometric station given monthly rainfall values of Vanaei rain gauge station using Copula-GARCH model

Next step, in order to present the Copula-GARCH model, it is first necessary to fit time series models on the studied data. In this regard, VAR model was used (Shahidi et al. 2020). The data were analyzed using the VAR model with different delays and finally the residual series of the VAR model was extracted for the paired variables. The basis of modeling based on the Copula-GARCH model is the VAR model. This model actually models the average part of the time series and due to the existence of heteroskedasticity in the time series of rainfall and flow discharge, modeling the variance of data was also investigated. In this regard, the residual series was extracted

335 for the paired variables and fitted by the GARCH model. The residual series were produced from
336 the GARCH model. Prior to fitting the GARCH model, the structural stability of the residual series
337 was investigated using the Ordinary least squares (OLS) residuals and Cusum tests and presented
338 in Figure 3 (Ploberger and Krämer, 1992).

339 **Fig. 3.**

340 The Cusum test calculates the process of empirical fluctuations using a specific method from
341 a generalized statistical framework. According to Figure 3 and the confidence intervals, the results
342 of this test were also confirmed (Kramer et al. 1991). As can be seen, the OLS-based CUSUM
343 process has not exceeded the confidence intervals. Hence, there is no evidence of structural change.
344 In the next step, for joint frequency analysis of the paired residual series of the VAR model, the
345 fitness of different copula functions and their rotational form were examined. Akaike information
346 criterion (AIC), Bayesian information criterion (BIC) and Log-Likelihood (Log-Like) were used
347 to select the superior copula function. The results of goodness of fit tests were performed for the
348 paired residual series confirmed the superiority of Gaussian copula to other copulas. Accordingly,
349 the Gaussian copula function was selected with AIC, BIC and Log-Like values of -90.9, -87 and
350 46.4, respectively, with a copula parameter of 0.51. After selecting the Gaussian copula as the
351 superior copula, the simulation based on the bivariate copula of the paired residual series was
352 investigated. The results of simulation of the mean monthly discharge (MMD) of Talezang station
353 given recorded monthly rainfall (MR) at the Vanaei rain gauge station using Copula-GARCH
354 model in the period of 1988-2018 were presented as Figures 4 and 5.

355

356

357 **Fig 4.**

According to Figure 4, it can be seen that the simulated MMD (m^3/s) data has a good agreement with the observed data in both training and test stages. The maximum and minimum points are well simulated in most cases. In some cases, overestimation is seen, but on average, there is a good correlation between the observed and simulated values based on the Copula-GARCH model. The calibration results of the Copula-GARCH model are also shown in Figure 5, except for two overestimated cases, all the simulated data are in the 99% confidence intervals and the accuracy of the model is also confirmed. The simulation based on Copula-GARCH has the RMSE of 79.64 m^3/s on a monthly scale. According to the correlation values and Nash-Sutcliffe statistics, this RMSE is acceptable and the performance of the model in the simulation of MMD values (m^3/s) given the occurrence of MR values is confirmed. The results presented in Figures 4 and 5 are based on modeling the mean and variance of the MMD series and also include considering its heteroskedasticity, which is also modeled by the copula-based model. According to the statistics presented in Figure 5, it can be concluded that the proposed hybrid model has a high ability to simulate MMD values given by MR values.

Fig 5.

3.3 Simulation of MMD values of Talezang hydrometric station given monthly rainfall values of Vanaei rain gauge station using OSVR model

Simulation of the MMD of Talezang hydrometric station based on the OSVR model approach was performed by considering the MR values at the Vanaei rainfall station as input. The parameters of the OSVR model (Epsilon, c and Sigma) were optimized using the ant colony optimization (ACO) algorithm to achieve higher accuracy, and with 100 iterations, the values of Epsilon, c and Sigma were estimated 21.473, 1000 and 12.91, respectively. The results showed that in iteration 13, the OSVR model reached the optimal state and the cost function obtained the lowest value. The results of the objective function (RMSE) are shown in Figure 6. As can be seen from Figure

6, the OSVR model was able to reduce the simulation RMSE of MMD values from about 200 m^3/s to 105.8 m^3/s during iterations. There was no improvement in the results from 13 to 100 repetitions. Finally, using the optimized coefficients of the OSVR model, the simulation of the MMD values was performed given MR values and the simulation results are presented in Figures 7 and 8.

Fig 6.

Fig 7.

As it can be seen from Figure 7, there is good agreement between observed MMD values and corresponding values simulated by the OSVR model. The minimum points in these simulations are not well estimated due to the presence of zero MR data. Also, in some months, the maximum values are not well simulated. According to Figure 8, the results of the study and the RMSE of the OSVR model in simulating MMD values showed that the RMSE of the model is about 106 (m^3/s) and the NSE of the model is 78%. According to the 99% confidence intervals, the results showed that about 15 points of the simulated MMD values (m^3/s) are outside the confidence range, which indicates overestimation or underestimation of the target points. However, the performance of the model is acceptable according to the presented results.

Fig 8.

By using and implementing the Copula-GARCH model, which actually consists of the VAR, VAR-GARCH models and the joint residual series of the VAR-GARCH model, the NSE of this model compared to the OSVR model was improved by about 15%. On the other hand, compared to the OSVR model, the R^2 and RMSE statistics of the simulations by the Copula-GARCH model have been improved by 11% and 36%, respectively. The accuracy of simulation results of MMD values (m^3/s) improved about 36%, which indicates a decrease of RMSE by 26.16 (m^3/s). Finally, Violin Plot and Taylor Diagram were used to compare the performance of the used models and also to evaluate their certainty in simulating the MMD values of the Talezang hydrometric station.

406 The Violin plot of observed and simulated data is presented in Figure 10 and Taylor diagram of
407 Copula-GARCH and OSVR models also presented in Figure 11.

408 **Fig. 9.**

409 **Fig. 10.**

410 As can be seen from Figures 9 and 10, the results of the studied models in the MMD simulation
411 of the Talezang hydrometric station are very close to each other based on the evaluation criteria.
412 In other words, evaluation criteria cannot provide any information on how data are distributed.
413 Therefore, using plots such as Violin and Taylor diagrams can be very useful. The violin plot is
414 another form of box plot. Box plots show only the minimum, maximum, mean, and quarters of the
415 data, but the violin plot is used to visualize the data distribution and its probability density.
416 According to the presented violin plot, it can be seen that the simulated time series with the OSVR
417 model has slightly overestimating and the maximum values are simulated more than the
418 corresponding observed values. But in general, the Copula-GARCH model simulates MMD values
419 better than the OSVR model. According to Figure 9, the range of variation of MMD values, as
420 well as the range of 5 and 95% of the values simulated by the Copula-GARCH model, is closer to
421 the observed values, which increases the reliability of the model. As it can be seen from Figure 9,
422 the Copula-GARCH model better covers the range of data changes. Appropriate certainty in
423 copula-based simulation of the different meteorological and hydrological values has been
424 investigated and confirmed in various studies. Recent studies in simulation of different hydro-
425 meteorological variables using two and multidimensional copula functions show that these
426 functions have high ability and accuracy in bivariate simulations of meteorological and
427 hydrological phenomena (Bezak et al. 2017; Ahmadi et al. 2018; Nazeri Tahroudi at al. 2020 a &
428 b; Kim et al. 2019). Regarding the simulation of other variables using copula functions, Bazak et

al. (2017) also stated that the copula functions have acceptable accuracy in the simulation of sediment values and have a higher accuracy compared to the regression method. In addition, the results of this study showed that by using the copulas, the accuracy and efficiency of the hybrid VAR-GARCH model can be improved, that are consistent with the results of Yoo et al. (2016). Yuan et al. (2020) also confirmed the acceptable accuracy of the Copula-GARCH model in agricultural commodity price modeling by combining copulas with GARCH models. The Taylor diagram (Figure 10) is based on the geometric relationship among the correlation coefficient (R), the time series standard deviation, and the root mean square difference. This diagram is a good tool for evaluating various methods. The reference point (hollow circle between standard deviation 200 and 300) indicates the position of the observed values based on the standard deviation of the time series. The position of the simulated data with OSVR and Copula-GARCH models is plotted based on the root mean square difference, its correlation coefficient with the observed values and the time series standard deviation in Figure 10. Due to the fact that the Copula-GARCH location is closer to the reference point, it can be concluded that the Copula-GARCH model have a higher accuracy than the OSVR model in simulating the MMD values.

4. Conclusion

In this study, a Copula-GARCH model and a VAR model were used to simulate MMD values of the Talezang hydrometric station given MR values of Vanaei rain gauge in the Dez basin in southwestern Iran. Due to the influence of MR time series on MMD values on a monthly scale, vector autoregressive model (VAR) was used as the base model. The VAR model provides relatively good results due to the involvement of an effective parameter in the modeling. The correlation of the studied variables using Kendall's tau test was 0.34, which satisfies the initial

condition of using copula-based models. The results of the joint frequency analysis of the paired variables showed that the Gaussian copula has the best fitness on the residual series of VAR model. The results of simulation of MMD values based on Copula-GARCH model showed that this model has a high ability to simulate the studied variables and has high accuracy. Appropriate estimation of MMD values is observed in all months and in the whole range of changes in MMD values. The maximum and minimum points of the MMD values are also well estimated throughout the studied period. In this study, in addition to the copula model, the OSVR model was applied to simulate monthly runoff. After estimating the coefficients of the SVR model using the ant colony optimization algorithm, the MMD values were simulated according to the MR values. The results showed that the OSVR model did not simulate the minimum points or base flow well due to the presence of zero data in the rainfall time series. The values of R^2, RMSE and NSE statistics also confirmed the superiority of the Copula-GARCH model over the OSVR model. In addition to the mentioned statistics, the results of Violin plot and Taylor diagram also showed higher certainty of Copula-GARCH model than OSVR model. The proposed approach, due to the use of the copula function and the appropriate marginal distribution functions for each series, brought the simulation results closer to the observed values, which increased the model accuracy. Using this approach will improve performance in simulation of important meteorological and hydrological variables, which will be very useful in water resources planning and management.

5. Acknowledgements

The authors would like to thank Politecnico di Milano for providing the facilities to the first author as a visiting researcher. Also, the authors would like to thank Iran Water Resources Management Company for providing the data.

6. Authors' Contributions Conceptualization:

Methodology: M. Nazeri Tahroudi and F. Ahmadi; **Formal analysis and investigation:** M. Nazeri Tahroudi, R. Mirabbasi, Y. Ramezani and F. Ahmadi; **Writing—original draft preparation**: M. Nazeri Tahroudi; **Writing—review and editing:** M. Nazeri Tahroudi, R. Mirabbasi, Y. Ramezani and F. Ahmadi; **Supervision:** R. Mirabbasi, Y. Ramezani.

7. Declarations

Funding (Not applicable)

Conflicts of interest/Competing interests (The authors declare no conflicts of interest)

Availability of data and material (The data and material used in this work are available from the corresponding author by request)

Ethical Approval (Not applicable)

Consent to participate (Not applicable)

Consent to Publish (Not applicable)

Consent for Publication (Not applicable)

8. References

Abdi A, Hassanzadeh Y, Talatahari S, Fakheri-Fard A, Mirabbasi R (2017a) Regional bivariate modeling of droughts using L-comoments and copulas. Stochastic Environmental Research and Risk Assessment 31(5):1199-1210.

Ahmadi F, Radmanesh F, Sharifi M R, Mirabbasi R (2018) Bivariate frequency analysis of low flow using Ccopula functions (Case study: Dez River Basin, Iran). Environmental Earth Sciences 77: 643 DOI: 10.1007/s12665-018-7819-2.

497 Ayantobo O O, Li Y, Song S, Javed T, Yao N (2018) Probabilistic modelling of drought events in
498 China via 2-dimensional joint copula. Journal of Hydrology 559: 373-391.
499 Bedford T, Cooke R M (2001) Probability density decomposition for conditionally dependent
500 random variables modeled by vines. Annals of Mathematics and Artificial intelligence 32(4):
501 245-268.
502 Bedford T, Cooke R M (2002) Vines: A new graphical model for dependent random variables.
503 Annals of Statistics 30(4): 1031–1068.
504 Bollerslev T, Chou R Y, Kroner K F (1992) ARCH modeling in finance. A selective review of
505 the theory and empirical evidence. Journal of Econometrics 52: 5-59.
506 Colorni A, Dorigo M, Maniezzo V (1991) Ant system: An autocatalytic optimizing process.
507 Dipartimento Di Elet-tronica, Politecnico Di Milano, Milan, Italy.
508 Dibike Y B, Velickov S, Solomatine D, Abbott M B (2001) Model induction with support vector
509 machines: introduction and applications. Journal of Computing in Civil Engineering 15(3):
510 208-216.
511 Dorigo M (1992) Optimization, learning and natural algorithms. PhD Thesis, Politecnico di Milano
512 Duan J C (1996) A unified theory of option pricing under stochastic volatility-from GARCH to
513 diffusion. Hong Kong University of Science and Technology, Research project
514 Engle R F (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of
515 United Kingdom inflation. Econometrica 50(4): 987-1007.
516 Floros C, Jaffry S, Lima G V (2007) Long memory in the Portuguese stock market. Studies in
517 Economics and Finance 24(3): 220-232.

518 Guo, A, Wang, Y (2017). Assessment of variability in the hydrological cycle of the Loess Plateau,
519 China: examining dependence structures of hydrological processes. AGUFM, 2017,
520 GC41C-1028.
521 Hollander M, Wolfe D A, Chicken E (2014) Nonparametric statistical methods. New York: Wiley.
522 Khozeymehnezhad H, Nazeri-Tahroudi M (2020) Analyzing the frequency of non-stationary
523 hydrological series based on a modified reservoir index. Arabian Journal of Geosciences
524 13(5): 1-13.
525 Khozeymehnezhad H, Tahroudi M N (2019) Annual and seasonal distribution pattern of rainfall
526 in Iran and neighboring regions. Arabian Journal of Geosciences 12(8): 1-11.
527 Kim J E, Yoo J, Chung G H, Kim T W (2019) Hydrologic risk assessment of future extreme
528 drought in South Korea using bivariate frequency analysis, Water 11(10): 2052.
529 Krämer W, Ploberger W, Schlüter I (1991) Recursive vs. OLS residuals in the CUSUM test. In
530 Economic Structural Change (pp. 35-47). Springer, Berlin, Heidelberg.
531 Li F, Zheng Q (2016) Probabilistic modelling of flood events using the entropy copula. Advances
532 in Water Resources, 97: 233-240.
533 Modarres R, Ouarda T B (2013) Generalized autoregressive conditional heteroscedasticity
534 modelling of hydrologic time series. Hydrological Processes 27(22): 3174-3191.
535 Moffat I U, Akpan E A, Abasiekwere U A (2017) A time series evaluation of the asymmetric
536 nature of heteroscedasticity: An EGARCH approach. International Journal of Statistics and
537 Applied Mathematics 2(6): 111-117.
538 Nazeri-Tahroudi M, Ramezani, Y (2020) Estimation of Dew Point Temperature in Different
539 Climates of Iran Using Support Vector Regression. Idojaras 124(4): 521-539.
540 Nelsen R B (2006) An introduction to copulas. Springer, New York, 269p.

Nelson D B (1991) Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59(2): 347-370.

Ploberger W, Krämer W (1992) The CUSUM test with OLS residuals. Econometrica 60(2): 271-285.

Ramezani Y, Tahroudi M N (2020) Improving the performance of the SPEI using four-parameter distribution function. Theoretical and Applied Climatology 139(3): 1151-1162.

Ramezani Y, Tahroudi M N, Ahmadi F (2019) Analyzing the droughts in Iran and its eastern neighboring countries using copula functions. IDOJARAS 123(4): 435-453.

Salvadori G, De Michele C, Kottegoda N T, Rosso R (2007) Extremes in nature: an approach using copulas (Vol. 56). Springer Science & Business Media.

Serinaldi F, Bonaccorso B, Cancelliere A, Grimaldi S (2009) Probabilistic characterization of drought properties through copulas. Physics and Chemistry of the Earth, Parts A/B/C 34(10-12): 596-605.

Shahihi A, Ramezani Y, Nazeri-Tahroudi N, Mohammadi S (2020) Application of vector autoregressive models to estimate pan evaporation values at the Salt Lake Basin, Iran. Idojaras 124(4): 463-482, DOI:10.28974/idojaras.2020.4.3.

Sims C A (1980) Macroeconomics and reality. Econometrica 48(1): 1-48.

Sklar A (1959) Fonctions de Repartition and Dimensions et LeursMarges. Publications de L'Institute de Statistique, Universite' de Paris, Paris 8: 229–231.

Tahroudi M N, Khalili K, Ahmadi F, Mirabbasi R, Jhajharia D (2019a) Development and application of a new index for analyzing temperature concentration for Iran's climate. International Journal of Environmental Science and Technology 16(6): 2693-2706.

563 Tahroudi M N, Pourreza-Bilondi M, Ramezani Y (2019b) Toward coupling hydrological and
564 meteorological drought characteristics in Lake Urmia Basin, Iran. Theoretical and Applied
565 Climatology 138(3-4): 1511-1523.
566 Tahroudi M N, Ramezani Y De Michele C, Mirabbasi R (2020a) A new method for joint frequency
567 analysis of modified precipitation anomaly percentage and streamflow drought index based
568 on the conditional density of copula functions. Water Resources Management 34(13): 4217-
569 4231.
570 Tahroudi M N, Ramezani Y, De Michele C, Mirabbasi R (2020b) Analyzing the conditional
571 behavior of rainfall deficiency and groundwater level deficiency signatures by using copula
572 functions. Hydrology Research 51(6): 1332-1348.
573 Tahroudi M N, Siuki A K, Ramezani Y (2019c) Redesigning and monitoring groundwater
574 quality and quantity networks by using the entropy theory. Environmental Monitoring and
575 Assessment 191(4): 250.
576 Tse Y K, Tsui A K. C (2002) A multivariate generalized autoregressive conditional
577 heteroscedasticity model with time-varying correlations. Journal of Business Economic
578 Statistics 20(3): 351-362.
579 Wang W, Van Gelder P H A J M, Vrijling J K, Ma J (2005) Testing and modelling autoregressive
580 conditional heteroskedasticity of streamflow processes. Nonlinear Processes in Geophysics,
581 European Geosciences Union (EGU) 12(1): 55-66. hal-00302471
582 Watanabe T (2012) Quantile forecasts of financial returns using realized GARCH models. The
583 Japanese Economic Review 63(1): 68-80.

584 Yoo J, Kim D, Kim H, Kim, T W (2016) Application of copula functions to construct confidence
585 intervals of bivariate drought frequency curve. Journal of Hydro-environment Research 11:
586 113-122.
587 Yuan X, Tang J, Wong W K, Sriboonchitta S (2020) Modeling co-movement among different
588 agricultural commodity Markets: a Copula-GARCH Approach. Sustainability 12(1): 393.
589 Yusof F, Kane I L (2013) Volatility modeling of rainfall time series. Theoretical and Applied
590 Climatology 113(1-2): 247-258.

Figures

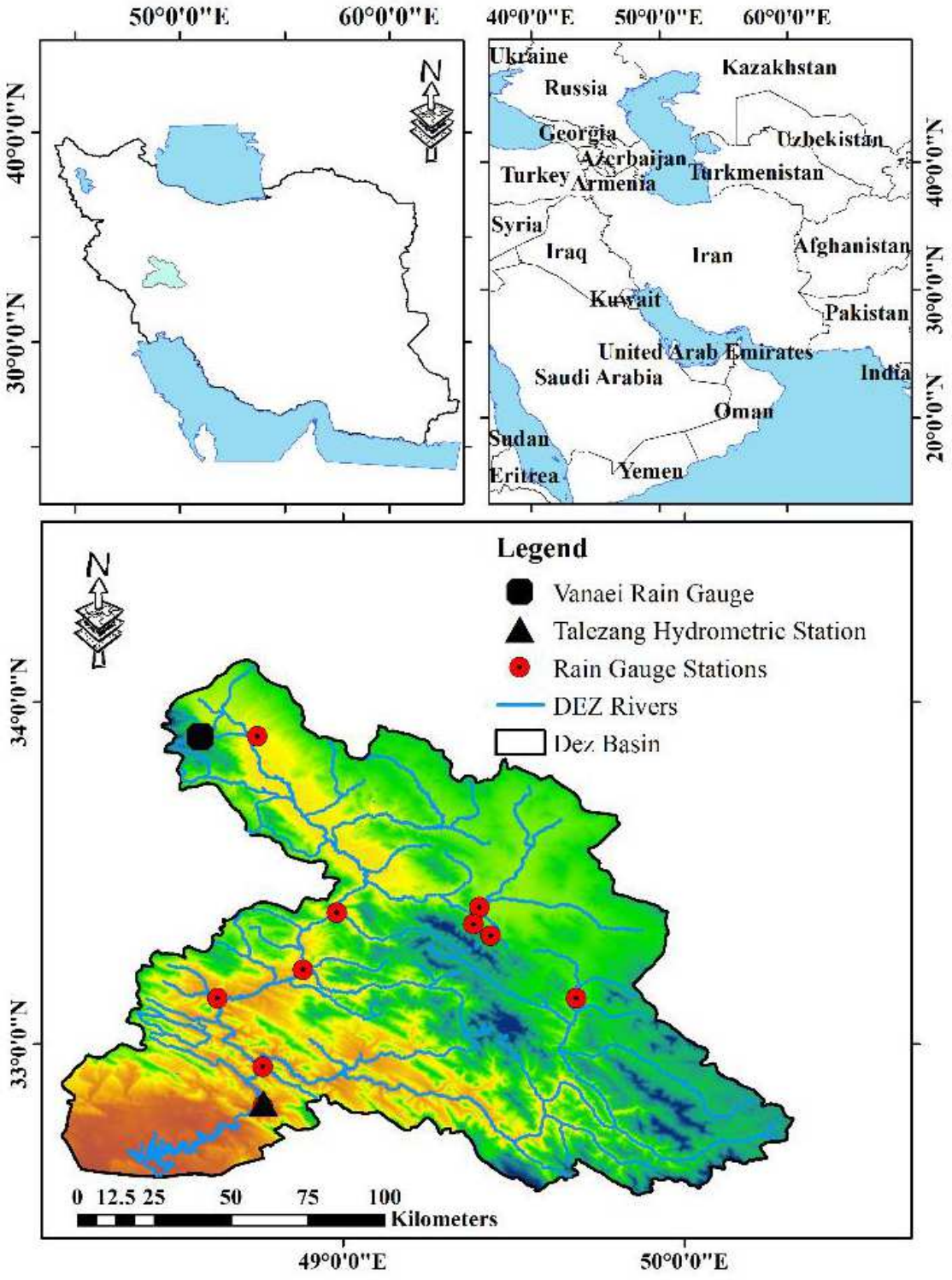

Figure 1

Location map of the studied rain gauge and hydrometric stations in Dez basin Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or

area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

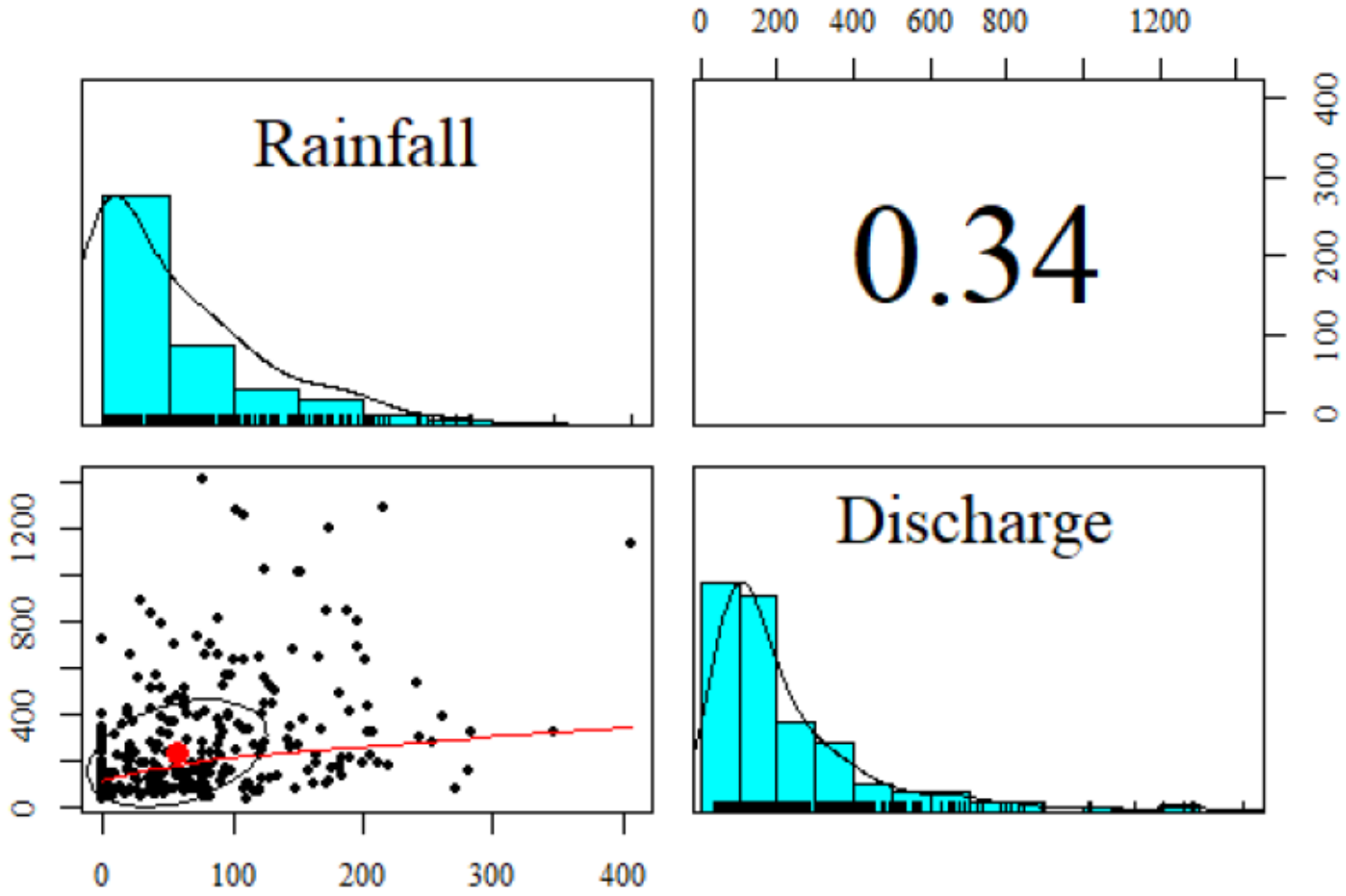

Figure 2

Kendall's tau, histogram and empirical contour lines of rainfall (mm) and flow discharge (m3/s) series

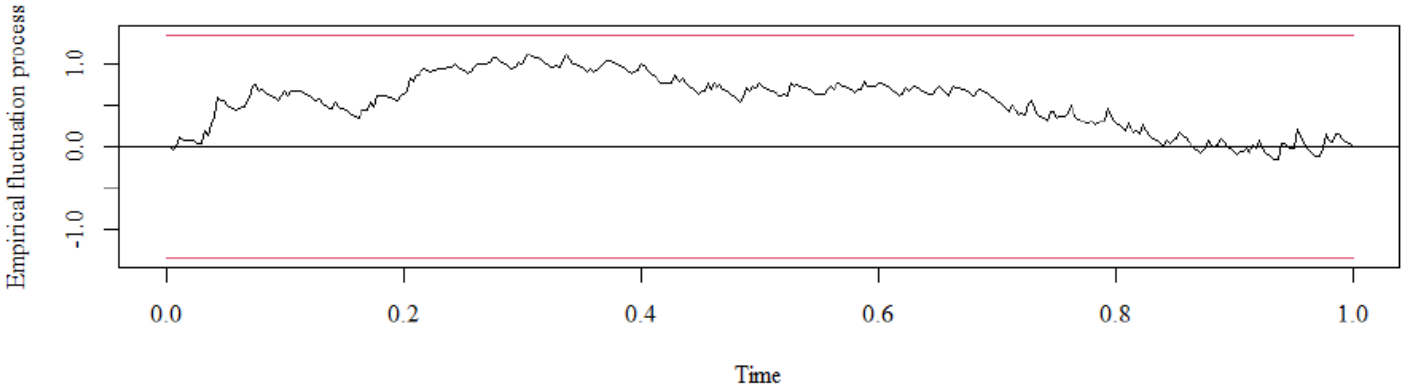

Figure 3

The OLS-based Cusum process of VAR(p) model (the red lines show the confidence intervals)

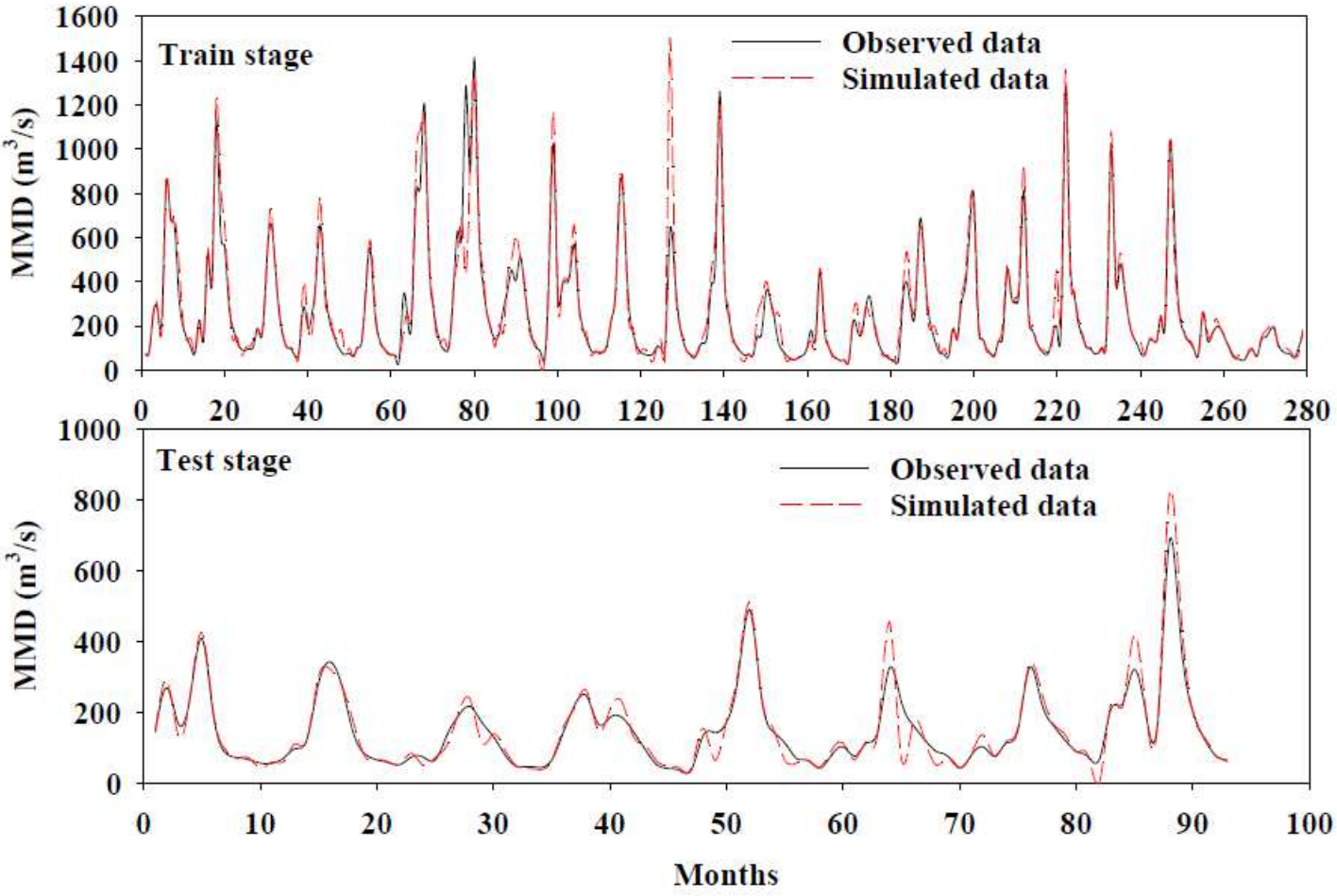

Figure 4

Results of simulation the MMD (m3 /s) of Talezang hydrometric station given MR values (mm) of Vanaei rain gauge station using Copula-GARCH model in two training and test stages (1988-2018)

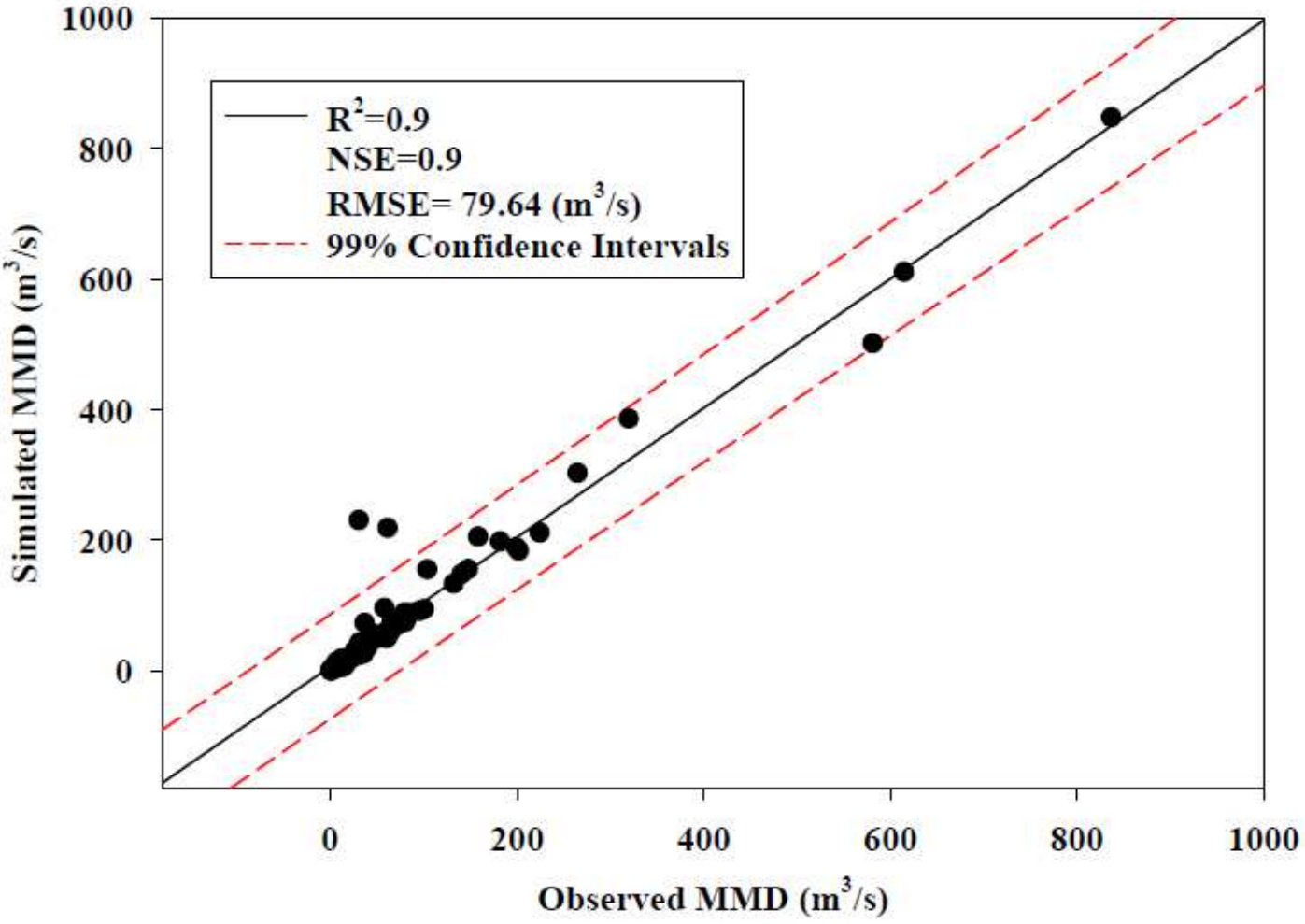

Figure 5

The simulated values of the MMD (m3/s) using Copula-GARCH model versus observed values at the Talezang hydrometric station

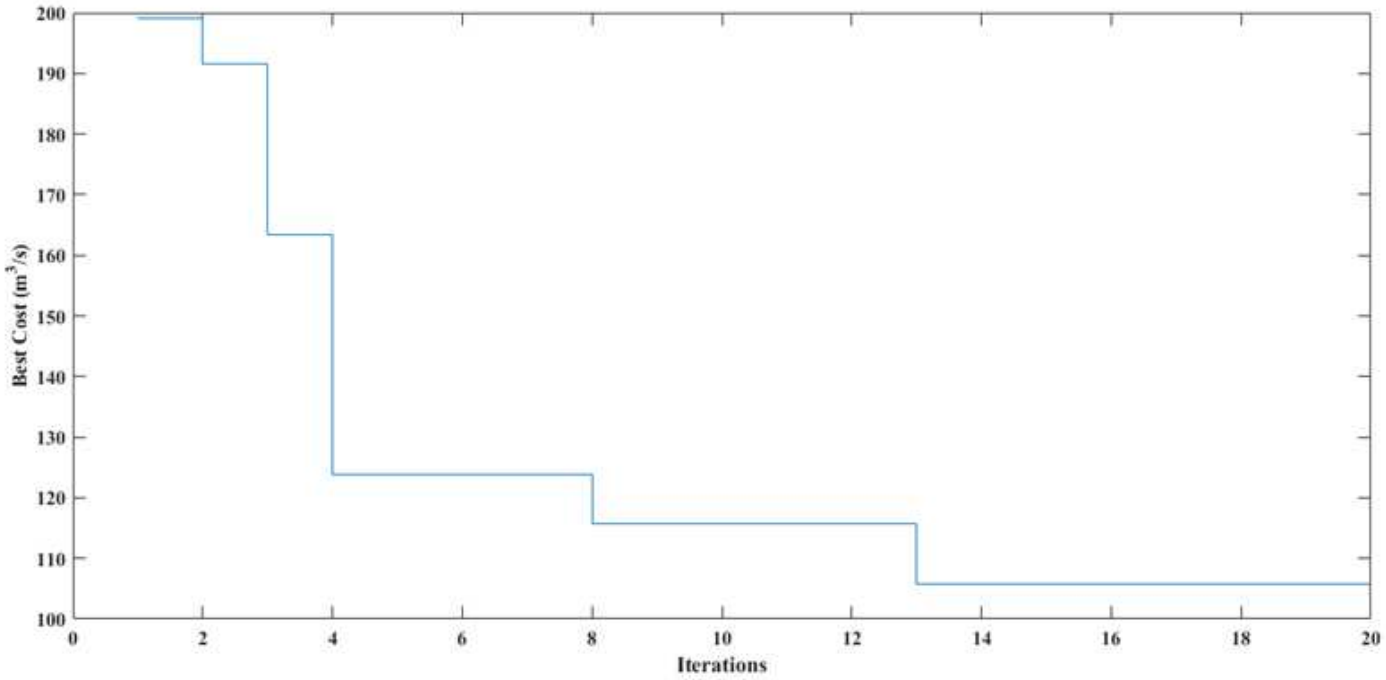

Figure 6

Results of objective function (RMSE) improvement in different iterations to optimize OSVR model coefficients

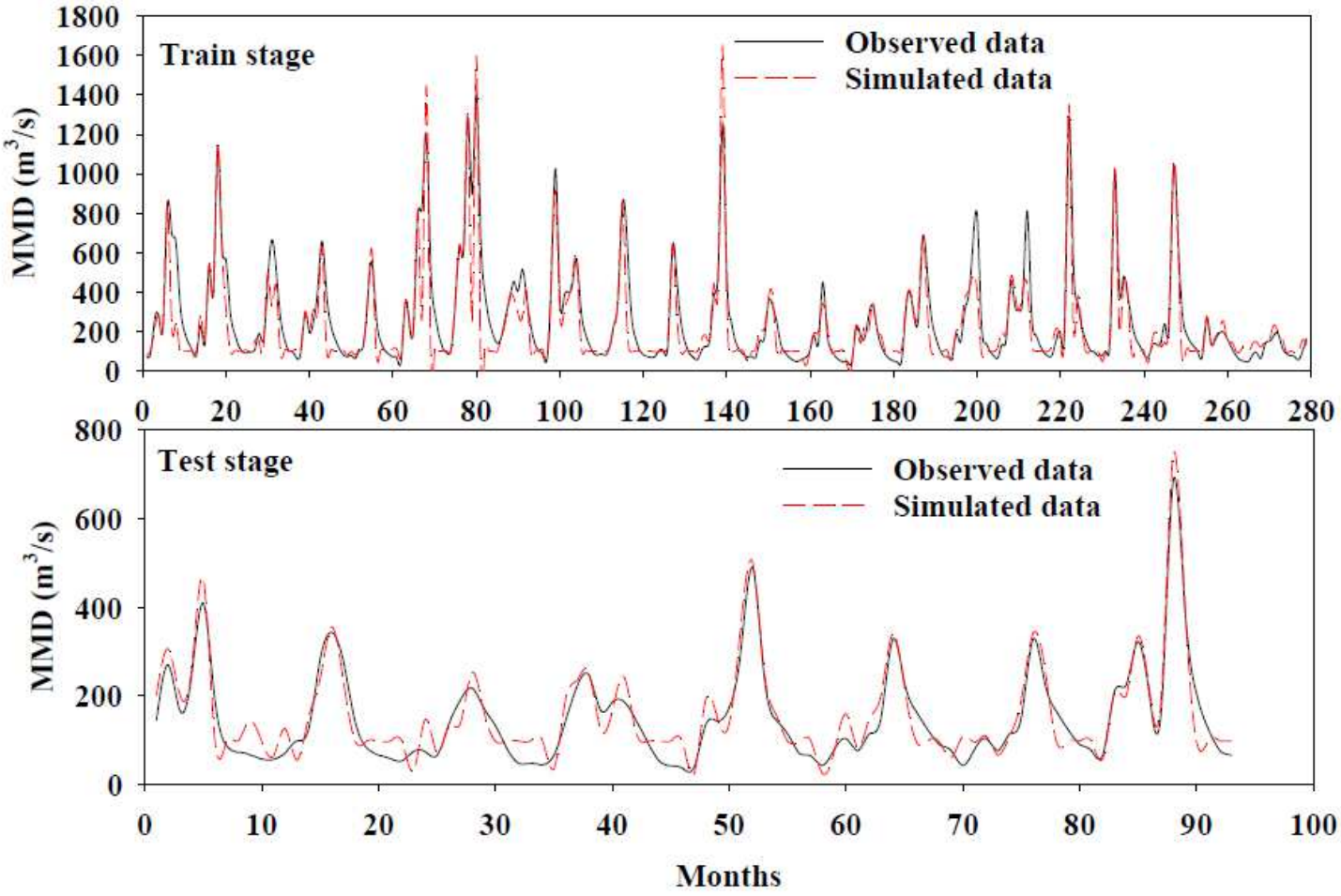

Figure 7

Results of simulation the MMD (m3 /s) of Talezang hydrometric station given MR values (mm) of Vanaei rain gauge station using OSVR model in two training and test stages (1988-2018)

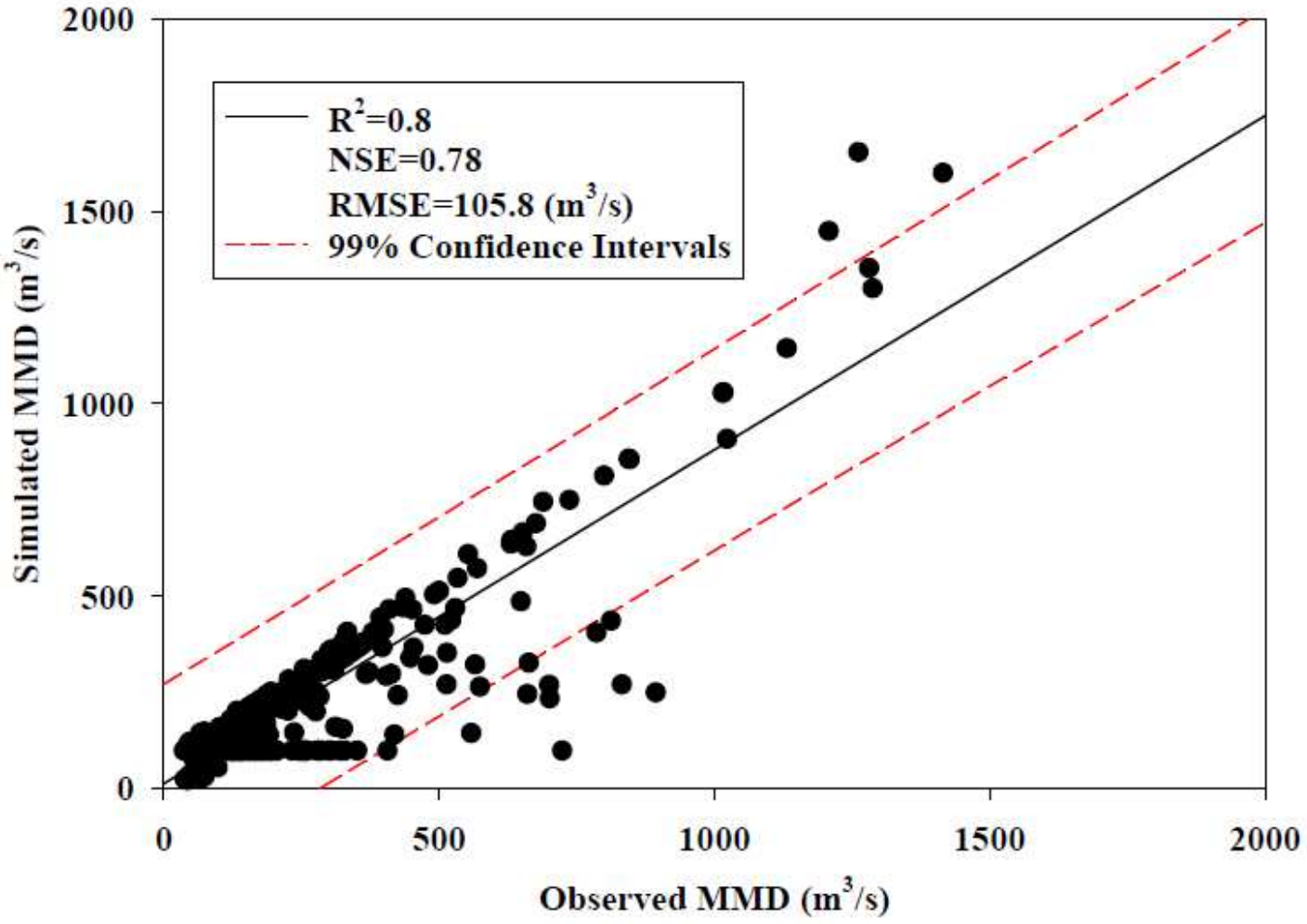

Figure 8

The simulated values of the MMD (m3/s) using OSVR model versus observed values at the Talezang hydrometric station

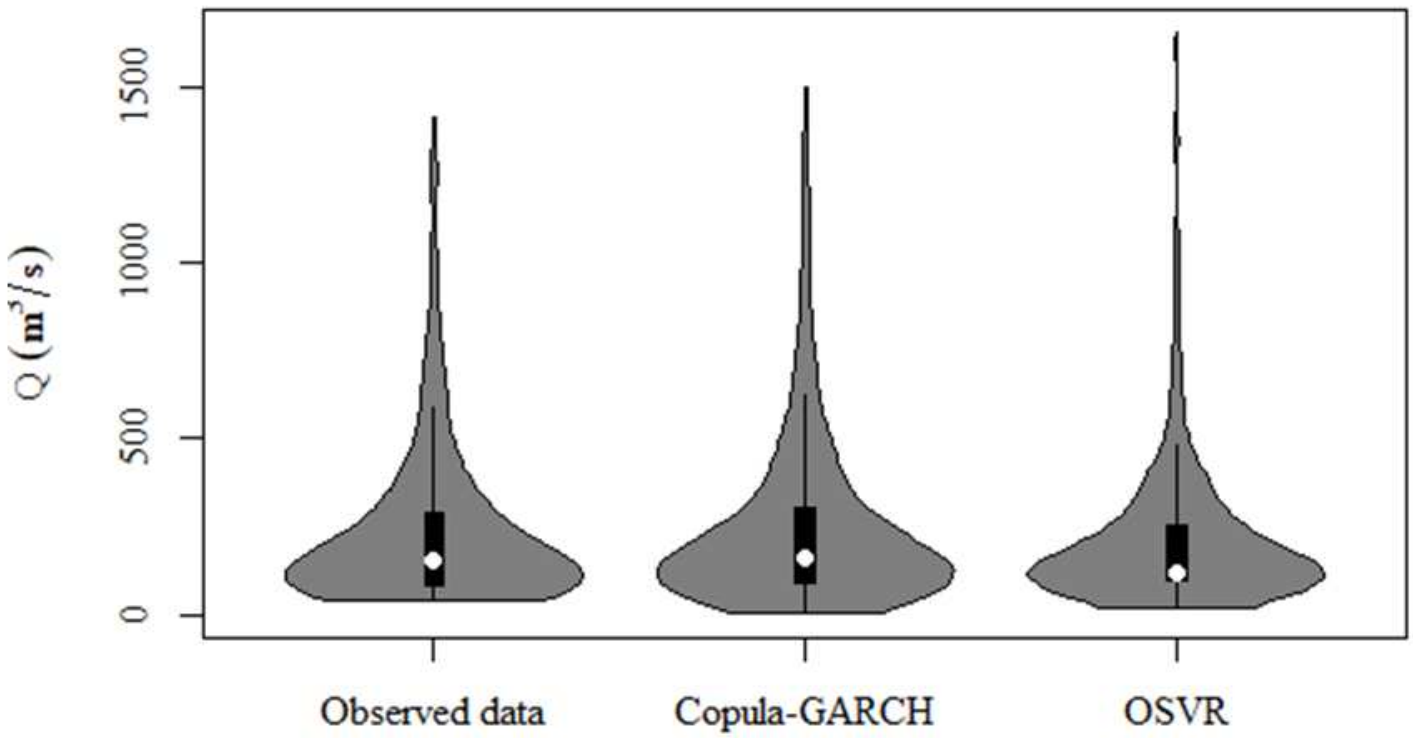

Figure 9

Violin plot of MMD (m3/s) series in the Talezang hydrometric station and corresponding simulated series by Copula-GARCH and OSVR models.

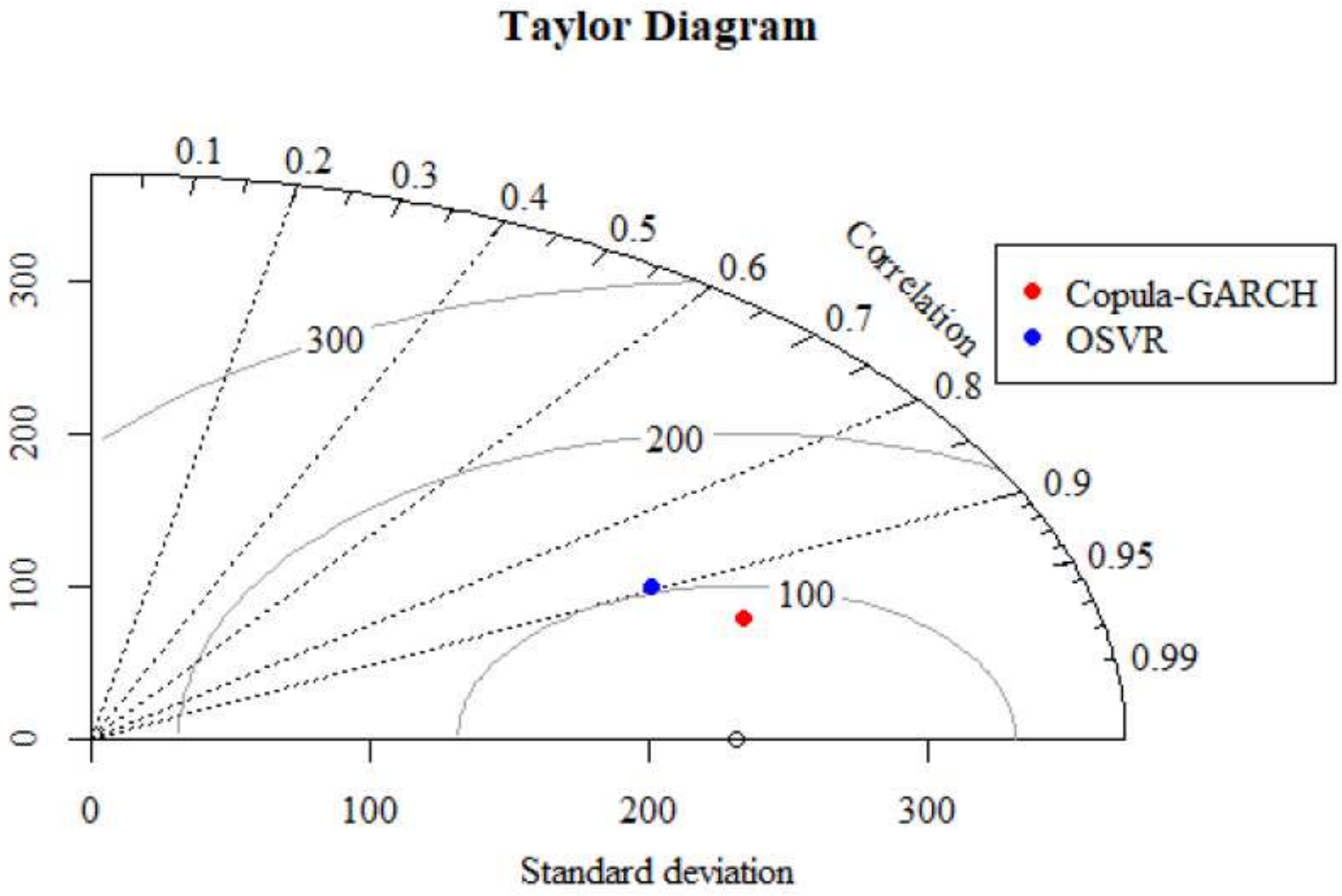

Figure 10

Taylor Diagram of MMD (m3/s) and corresponding simulated series using OSVR and Copula-GARCH models