Heliyon 7 (2021) e07877

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Copula based post-processing for improving the NMME precipitation forecasts

Farhad Yazdandoost [a,*], Mina Zakipour [a], Ardalan Izadi [b]

[a] Department of Civil Engineering, K. N. Toosi University of Technology, Tehran, Iran
[b] Multidisciplinary International Complex (MIC), K. N. Toosi University of Technology, Tehran, Iran

ARTICLE INFO

Keywords:
Copula
NMME
Post-processing
Precipitation forecast
Semi-parametric distribution

ABSTRACT

Using reliable and timely precipitation forecasts on a monthly or seasonal scale could be useful in many water resources management planning, especially in countries facing drought challenges. Amongst many, the North American Multi-Model Ensemble (NMME) is one of the most well-known models. In this study, a Bayesian method based on Copula functions has been applied to improve NMME precipitation forecasts. This method is based on the existence of a correlation between the raw forecast and observational data. Two main factors affect the results of rainfall improvement based on the selected method. This research has presented innovative methods in these regards namely; 1) the approach of selecting the appropriate statistical distribution for variables and 2) the selection method of improved data according to the conditional probability distribution functions (CPDF). To evaluate the effectiveness of the statistical distribution, firstly the precipitation forecast improvement model has been developed based on the application of parametric (Exponential, Normal, Gamma, LogNormal and General Exreteme Value (GEV)) and non-parametric distributions (Standard Normal Kernel). Then the novel mixed distribution function based on GEV parametric distribution and Standard Normal Kernel (non-parametric distribution) has been suggested. As the second aim, a new method for selecting improved data based on the center of mass of estimated CPDF is presented. The evaluation of the proposed method for estimating the statistical distribution of data and improving the forecast precipitation by the NMME model has been performed in Sistan and Baluchestan province in Iran. In this regard, the data of 1982–2010 for the calibration period and the data of 2012–2016 for the validation of the results have been used. According to the results, the non-parametric distribution best fitted with the data in the time series and selecting the appropriate bandwidth increased the efficiency of this distribution. Besides, due to the weakness of non-parametric distributions in the boundaries, the use of GEV distribution with a high ability to estimate boundary conditions as semi-parametric distribution, led to improved performance of the proposed distribution. Finally, the selection of the improved data based on the center of the mass method has efficiently provided much improvement compared to the maximum likelihood method commonly used.

1. Introduction

From the past to the present and towards the foreseeable future, droughts and floods, as the high frequency and destructive natural disaster events may be regarded as the two sides of the same coin (Roser and Ortiz-Ospina, 2016). Prediction of precipitation, as the major component of the climate system (water cycle), can be very effective in reducing the potential damages due to these natural disasters, while it can be essential in building resilience to climate (extreme water resources planning and management). During the past decade, responding to this necessity, numerous precipitation prediction dynamic models (Al Zawad and Aksakal, 2010; Fenta Mekonnen and Disse, 2018; Hattermann et al., 2017; Peng et al., 2018) have been developed and explored. These models are usually classified in the category of Global Circulation Models (GCMs).

Among different existing GCMs, the North American Multi-Model Ensemble (NMME) model is an effective seasonal precipitation model for precipitation prediction coupling models from the US and Canadian climate modeling centers. It is capable of providing timely and reliable seasonal precipitation prediction since 2007 with ease of access to data (Becker et al., 2020; Roy et al., 2020; Slater et al., 2019). Since then, finding the relation between the observed data and predicted data has

* Corresponding author.
E-mail address: yazdandoost@kntu.ac.ir (F. Yazdandoost).

https://doi.org/10.1016/j.heliyon.2021.e07877
Received 21 November 2020; Received in revised form 8 February 2021; Accepted 23 August 2021
2405-8440/© 2021 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

inspired many researchers around the world to evaluate the forecast skills of the NMME models.

Despite NMME's accuracy, climate variables that are strongly affected by changes in spatial scale (such as precipitation) still meet significant errors. Thus, it is propounded to not utilize the climate model's forecast products directly, due to three main reasons: 1) ensemble climate forecasts are not so accurate, 2) various assumptions are taken for the sake of overcoming the shortage of initial and boundary condition for every region, 3) due to some limitations the models usually forecast variables for large scale applications, hence at least they need spatial bias correction (as the most convenient approach of post-processing) even if the variables can be accurate enough (Rayner et al., 2005; Tao et al., 2014; Wu et al., 2011). Therefore, many works have tried to improve the accuracy of predictions by statistical methods such as bivariate joint distributions, Meta-Gaussian distribution function (Kelly and Krzysztofowicz, 1997) and Bayesian Joint Probability (BJP) (Robertson et al., 2013). These methods require to transform both observation and forecast variables into normal distribution. This forced process will reduce the accuracy of the estimated distribution between the variables (Brown and Seo, 2013; Madadgar and Moradkhani, 2014; Madadgar et al., 2014). Hence there is a demand for bivariate (or multivariate) distribution functions that can address variables according to the suitable estimated distributions. The construction of multivariate distributions in the notion of Copulas, which has been developed by Sklar, are applicable for this purpose (Sklar, 1959).

The copula functions are unit cube functions that relate the multi-dimensional distributions to their one-dimensional marginal (Sklar, 1959). At first, it was commonly used in financial approaches (Cherubini, 2004; de Melo Mendes and de Souza, 2004; Frees et al., 1996; Frees and Valdez, 1998; Hürlimann, 2004). The application of Copula functions in the hydrological domain started in 2003 (De Michele and Salvadori, 2003; Favre et al., 2004; Salvadori & De Michele, 2004a, 2004b). They have been widely utilized in different aspects such as drought, groundwater monitoring, etc (Bárdossy, 2006; Madadgar & Moradkhani, 2013, 2014; Salvadori and De Michele, 2010). Copula functions, as the post-processing method, was also used in several studies to reduce involved uncertainties. For example, a) illustrating the relation between two uncorrelated variables such as observation and simulation data of streamflow (Madadgar et al., 2014), b) applying Copula functions in the Bayesian model (Madadgar and Moradkhani, 2014; Schefzik et al., 2013), c) applying Ensemble Copula Coupling (ECC) to predictions of temperature, pressure, precipitation and wind (Schefzik, 2013).

One of the questions addressed in this study is the investigation of the different marginal distributions for Copula functions. However, while the effectiveness of using Copula-based methods for improving the accuracy of NMME precipitation forecasts has been investigated by many researchers, the use of non-parametric marginal distribution has been rarely investigated. In contrast, previously obtained results showed that nonparametric marginal distributions are precise, uniform and satisfactory rather than the parametric marginal distribution especially for multi-modal data used in hydrological investigations (Adamowski, 1985; Kim et al., 2003; Kim et al., 2006a, b; Kocsis et al., 2017). According to the literature, the accurate estimation of nonparametric distributions is imperative for research (Efromovich, 1999; Han et al., 2018; Xin et al., 2020). Among several used nonparametric density estimation methods, the kernel density is a widely accepted distribution for different hydrological variables (Ghosh and Mujumdar, 2007; Kim et al., 2006a, b; Kim et al., 2003; Lall et al., 1996).

Despite the ease and accuracy of kernel density estimation (KDE) usage, it suffers from a main challenge/limitation. KDE couldn't estimate the density in boundaries as well. Hence the extreme value estimation would be prone to bias. In other words, precise estimation of extreme precipitation values for this method is rarely expected. A combination of parametric and non-parametric distributions, called semiparametric distribution, can be introduced as a practical solution for this shortcoming. Hjort and Glad (1995) show that the semiparametric kernel estimator, which starts with normal distribution, would render better estimates than the traditional kernel density estimator. Several studies show that General Extreme Value (GEV) distribution is a suitable estimator for extreme event behavior (Ben Alaya et al., 2020; Fowler and Kilsby, 2003; Gao et al., 2016; Gilleland and Katz, 2006; Nadarajah, 2005). This distribution was also applied to GCM summer monsoon precipitation in India (Shashikanth & Sukumar, 2017). As one of the objectives of the research the authors seek to investigate the new proposed semi-parametric method based on GEV and KDE. Applying Copula function in Bayesian model as the Copula-Bayesian approach, results in a Conditional Probability Density Function (CPDF). The CPDF yields the likelihood of the specific amount of observation data given the particular value of the estimated precipitation data, therefore it is often called likelihood function (Reich and Cotter, 2015). In this regard, the selection of the most suitable method of applying this function has always been under debate. Consequently, how the application of various methods would yield different results of improved data must be investigated. Therefore, addressing this necessity is selected as the other aim of this research and a new selection method is introducing as the main contribution of the study.

Based on the presented review of literature, data improvement has always been discussed in different aspects. The main objectives of this study are focused mainly on the following three questions 1) how the proper marginal distribution for Copula functions can be selected. 2) how the combination of parametric & nonparametric distribution functions (named here as semi-parametric distribution) can present better fitting performance of marginal distributions and 3) how the best alternative can be selected amongst the possible choices based on CPDF. To illustrate the high ability of the novel method, it is compared with the Maximum Likelihood (ML) method as the most prevalent one. In the present study, both selection methods by different marginal distributions (parametric, non-parametric and proposed semi-parametric distribution) are implemented in R (Team, 2013), which allows fast implementation with access to various packages.

The rest of the paper is organized as the following, first, the used data and the study area is presented in section 2. Then, in section 3, the proposed methodology of research is introduced through three subsections including; 1) time-series preparation and marginal distribution, 2) the theory of Copula Function and 3) the evaluation of forecast accuracy. Section 4 presents the obtained results and finally, the concluding remarks are condensed in section 5.

2. Data and case study

2.1. Study area

Sistan and Baluchestan Province in south-east Iran is one of the driest regions in the world and is facing prolonged droughts (Yazdandoost et al., 2020). The prevailing climate of the region is barren and arid, with an average annual rainfall of about 183 mm and an average annual temperature of about 22.4 °C. These climatology variables besides the limited existing water resources would tremendously increase the importance of forecasting precipitation as the main water resource in this region. As a result, precipitation forecasts can be used in mid-term planning to manage exigent conditions. The observed long-term average monthly precipitation for this region is shown in Table 1.

2.2. Data sources

2.2.1. Observation data

The rain-gauge records are the primary source of observation precipitation. However, in Sistan and Baluchestan province as seen in Figure 1 the 20 existing rain-gauge stations do not provide proper distribution and coverage. Table 2 presents the detailed characteristic of these stations with discontinuous and sparse recordings over sufficient periods. Therefore, generating a reference gridded precipitation

Table 1. Observed long-term average monthly precipitation.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Prec. (mm)	25.94	29.48	33.79	16.44	8.04	6.9	15.93	14.28	4.6	3.35	4.53	19.47

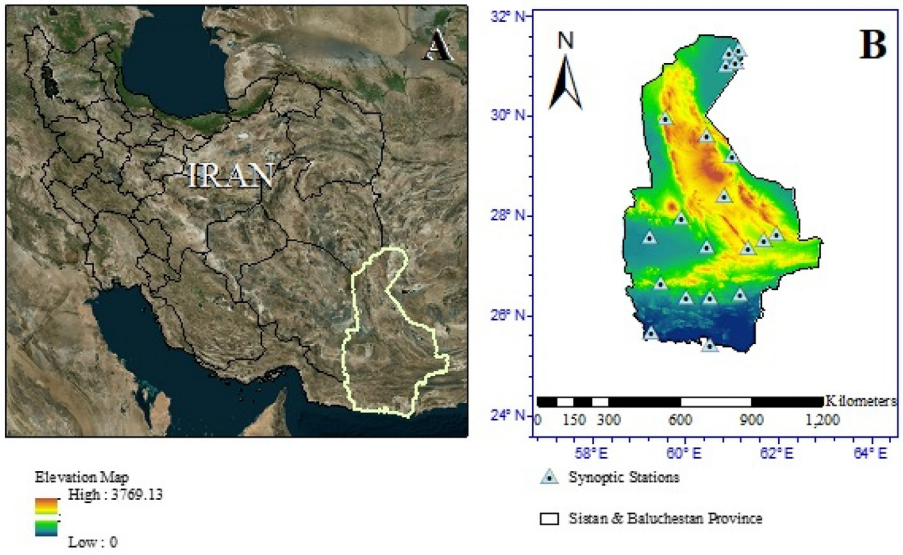

Figure 1. The location and topography of study area within A) Iran and B) The province of Sistan & Baluchestan.

Table 2. Coordinates of observational stations.

Station ID	Station Name	latitude	longitude	elevation (m)
99649	Bazman	27.85	60.18	957
40898	Chahbahar	25.28	60.65	8
40885	Delgan	27.48	59.45	391
19980	Founj	26.57	59.65	734
19992	Ghasreghand-looriani	26.22	60.73	500
19574	Hamon	30.85	61.45	473
19546	Hirmand	31.13	61.78	0
40879	Iranshahr	27.23	60.72	591.1
40870	Khash	28.23	61.19	1427
99608	Mirjaveh	29.02	61.43	836
40895	Nikshahr	26.23	60.20	510
99586	Nosratabad	29.85	59.98	1127
99650	Rask	26.23	61.40	406
40878	Saravan	27.39	62.32	1182
19942	Soran	27.28	62.00	0
40829	Zabol	31.09	61.54	489.2
40874	Zaboli	27.13	61.67	1271
99623	Zahak	30.90	61.68	495
40856	Zahedan	29.47	60.90	1370
19998	Zar Abad	25.58	59.40	35

dataset based on these non-evenly distributed stations across the entire area has hinted to consider the Global Precipitation Climatology Centre (GPCC) as the surrogate reliable reference gridded data for precipitation observations (Azizi et al., 2015; Darand and Zand, 2016; Rezayi et al., 2011).

The GPCC firstly has been established in 1989 by Germany's National Meteorological Service, the Deutscher Wetterdienst (DWD) on request of the World Meteorological Organization (WMO) (Yazdandoost et al., 2020). The GPCC provides different spatial resolution gridded precipitation data (2.5 ° × 2.5 °, 1.0 ° × 1.0 °, 0.5 ° × 0.5 °, and 0.25 ° × 0.25 ° resolution) based on around 80000 observational stations from several different sources. In this study, the monthly records of GPCC from 1982 to 2016 with the spatial resolution of 0.5 ° × 0.5° was used.

Table 3. Summary of the four NMME models and their characteristics, used in the study (modified from source: Yazdandoost et al., 2020).

Model	Hindcast period	Forecast period	Ensemble size	Reference
NCEP-CFSv2	1982–2010	2012-present	24 (28)*	(Saha et al., 2014)
CMC1-CanCM3	1981–2010	2011-present	10	(Merryfield et al., 2013)
CMC2-CanCM4	1981–2010	2011-present	10	(Merryfield et al., 2013)
NCAR-CCSM4	1982–2010	2011-present	10	(Gent et al., 2010; Kirtman and Min, 2009)

* Note: The value in the parenthesis presents the ensemble size for the forecast period.

2.2.2. Forecast data

Four NMME models are chosen as the forecast precipitation data. The NMME models have been presented at 1° spatial resolution, hence, they are gridded to 0.5° × 0.5°, using a bilinear method. These four NMME models included 54 ensemble members to calibrate the post-processing method in the hindcast period (1982–2010). For post-processing approach validation, these NMME models include 58 ensemble members for the forecast period (2012–2016). The more detailed information about the NMME models used in this research is condensed in Table 3. To improve the estimated data of the NMME models, the ensemble means of each model were calculated initially and then the Grand of the four mentioned models as the arithmetic mean of the predictions was used through the proposed post-processing process.

3. Methodology

To improve the NMME precipitation forecast data as the main objective of this research, a three-step post-processing method based on

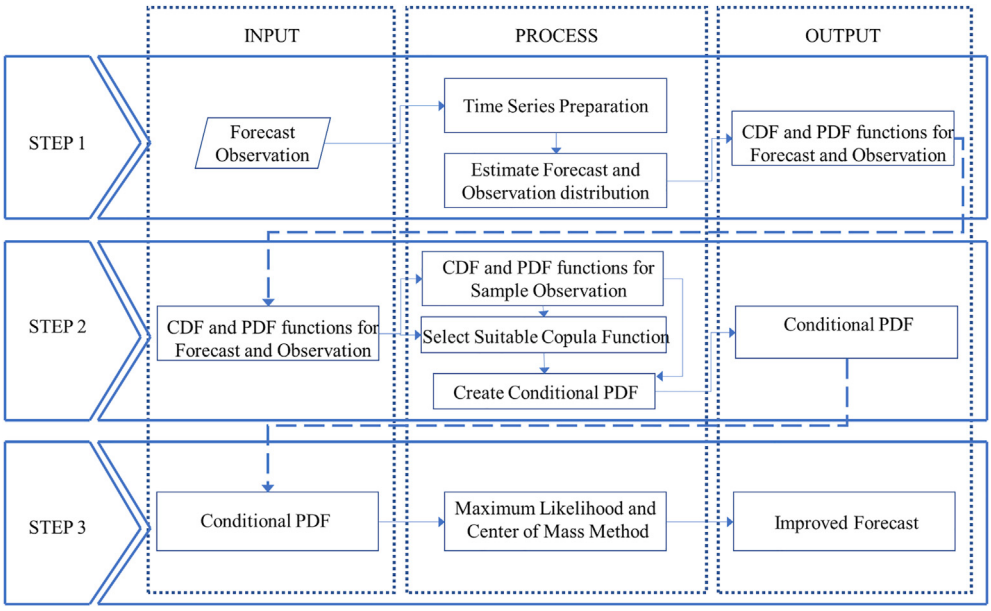

Figure 2. The dominant perspective of the Post-Processing.

Table 4. Summary of eight methods for bandwidth determination.

Bandwidth calculation method	Abbreviation	reference	R function
Mean Integrated Squared Error	MISE	(Nagler, 2016)	-
Asymptotic Mean Integrated Squared Error	AMISE	(Scott David, 1992)	h.amise
Maximum-Likelihood Cross-Validation	MLCV	(Habbema et al., 1974)	h.mlcv
Unbiased Cross-Validation	UCV	(Rudemo, 1982)	h.ucv
Biased Cross-Validation	BCV	(Scott and Terrell, 1987)	h.bcv
Complete Cross-Validation	CCV	(Jones and Kappenman, 1992)	h.ccv
Modified Cross-Validation	MCV	(Stute, 1992)	h.mcv
Trimmed Cross-Validation	TCV	(Feluch and Koronacki, 1992)	h.tcv

Table 5. Presented models.

Marginal distributions	Improved forecasts picked out method	Abbreviation name of model
Parametric	Maximum Likelihood	Par-ML
Parametric	Centre of Mass	Par-CM
Non-parametric	Maximum Likelihood	Ker-ML
Non-parametric	Center of Mass	Ker -CM
Semi-parametric	Maximum Likelihood	GEVKer-ML
Semi-parametric	Center of Mass	GEVKer- CM

the Copula-Bayesian approach is proposed (Figure 2). A detailed description of each step is described in the following. In this approach, the existence of the correlation between the historical observations and estimated forecasts (hindcast period) is supposed and it is expected that this assumed correlation would remain consistent in the future (forecast period) (Khajehei and Moradkhani, 2017). It is worth noting that, however there is no restriction for using the proposed post-processing framework for different lead times or period lengths, here, the improvement of monthly average forecast with 0-month lead time is considered as the main subject of research.

3.1. Time series preparation and marginal distribution

This step consists of two phases. In the first phase, at each 0.5-degree cell, the observed precipitation and forecast data during the study period were classified into 12 monthly time-series. In fact, for each month of the year, two monthly precipitation time series showing the 1) observed and 2) estimated precipitation data for that specific month during the past successive years (1982–2010) were generated.

In the second phase, the proper marginal distribution for each time-series is determined. In this regard, the Exponential, Normal, Gamma, LogNormal and GEV (as the parametric), Normal Kernel Density (as the non-parametric) and a newly introduced semi-parametric (combination of Kernel and GEV) distributions were investigated. For the last two types of mentioned distributions, finding the proper value of the Kernel distribution bandwidth is a prerequisite. To respond to this need, different choices are determined and evaluated for the bandwidth through assessment of eight methods (Guidoum, 2015). The detailed characteristics of the used methods are given in Table 4. Bandwidth calculation is done mostly by R Package for the Kernel Estimation of Bivariate Copula Densities (kdecopula). Afterward, the most suitable bandwidth is selected based on the best obtained result in the hindcast period.

After the determination of bandwidth for non-parametric and semi-parametric distributions, the best distribution is chosen by Kolmogorov Smirnov (K–S) test. The K–S test (Xiao, 2017; Zhao et al., 2017) is usually used to measure the difference among the CDF of forecast (or observations) data in the hindcast period and an empirical cumulative distribution (defined as a benchmark to investigate whether distribution looks exactly like the sample). If the maximum distance is less than a permitted tolerance, the referred type of distribution will be acceptable (Akramin et al., 2020). Tolerance is a function of the null hypothesis that is described as a table in different sources (Akramin et al., 2020). The Null hypothesis in this work was assumed to be 0.05 as previously suggested by Madadgar and Moradkhani (2012). Since the results of the K–S test can not necessarily introduce an option as the only final course of selection; the use of another criterion for the judging between the presented distributions by the K–S test seems necessary. In these circumstances the use of the Bayesian Information Criterion (BIC) test introduced by (Schwarz, 1978), as the secondary criterion is rational. This criterion is calculated as Eq. (1), where k is the

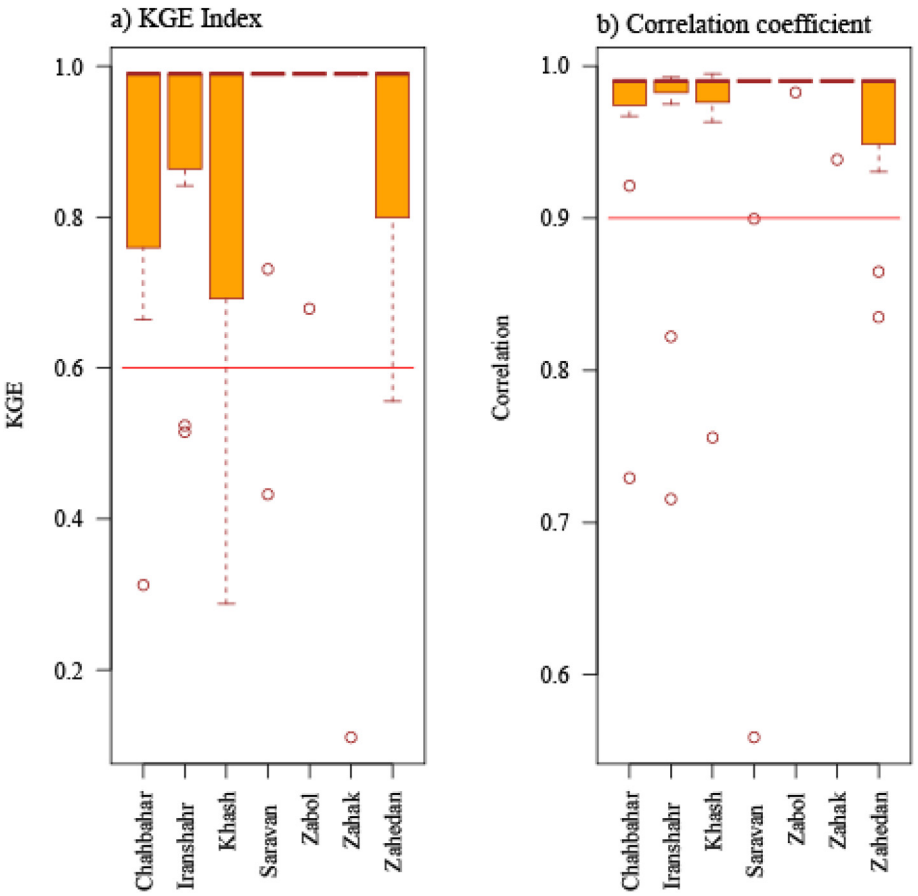

Figure 3. CC and KGE values for GPCC and rain-gauge data with a) KGE index and b) Correlation coefficient.

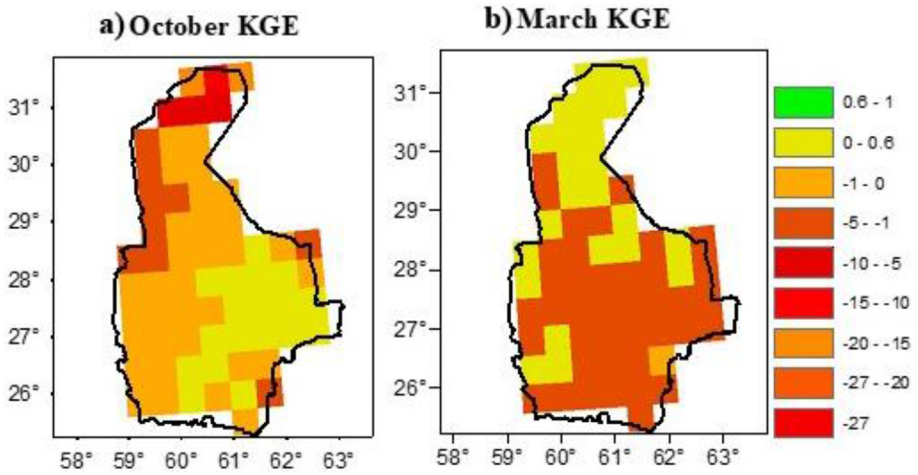

Figure 4. KGE values of raw forecast a) October b) March.

Table 6. K–S and BIC criterions value, parametric distributions.

Distribution	K–S			BIC		
	Observation	Hypothesis test	HINDCAST	Hypothesis test	Observation	HINDCAST
Exponential	0.35	Accept	0.065	Accept	-69.64	-77.89
Gaussian	0.93	Accept	0.68	Accept	-13.33	2.26
Gamma	1.55*10-5	Reject	1.11*10-16	Reject	-1.41	12.66
Lognormal	0.033	Reject	0.39	Accept	-2.32	14.23
GEV	0.99	Accept	0.9287	Accept	-127.61	-104.22

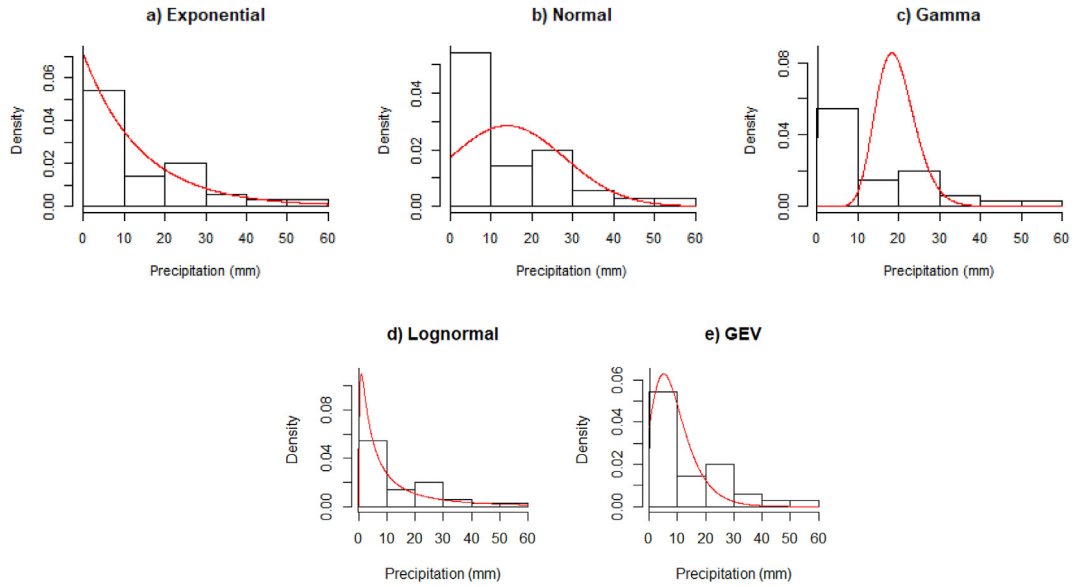

Figure 5. The example data histogram in front of parametric distribution a) Exponential, b) Normal, c) Gamma, d) Lognormal and e) GEV.

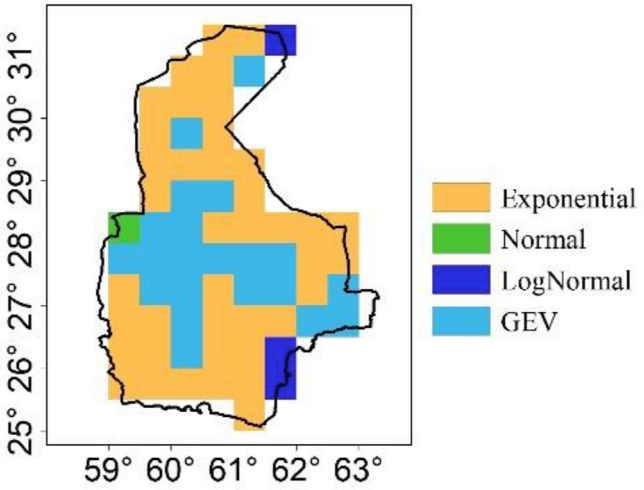

Figure 6. The dispersal of parametric distributions.

number of estimated parameters of investigated distribution, n is equivalent to the sample size, L is the maximized value of the likelihood function for estimated distribution.

$$\boldsymbol{BIC} = \boldsymbol{kln(n)} - 2\boldsymbol{ln(L)} \quad (1)$$

According to Rossi et al. (2020), BIC test can measure the efficiency of how any distribution function may fit empirical distribution function. Ideally, the lowest BIC would be the most desirable option.

For the sake of brevity, the current paper has only presented the proposed semi-parametric method and has avoided elaborating on the known parametric and non-parametric distributions. At the end of this section, the first aim of this research can be addressed completely.

3.1.1. Semi-parametric distribution

To overcome the deficiency of kernel distributions in the upper and lower boundaries, here a new applied semi-parametric distribution is proposed. This distribution benefits from the advantages of Kernel and GEV distributions simultaneously. In the proposed attitude, the kernel and GEV functions have been dedicated to the center and the (lower and upper) boundaries of any PDF respectively.

This distribution is built up in the following steps:

- Bandwidth determining: according to the description of the normal kernel distribution
- Substituting Kernel with GEV distribution in the bandwidth length of upper and lower boundaries
- Extending the GEV distribution to a point where it intersects with the kernel distribution. In this case, a continuous distribution can be provided.
- Finally, the best possible calculation for all the retrospective observed and forecast rainfall will be achieved from CDF and PDF

3.2. Post-processing based on copula functions

3.2.1. The theory of copula functions

The copulas (Sklar, 1959) are unit cube functions that relate the multi-dimensional distributions to their corresponding one-dimensional marginal distributions. Therefore, if $F_{x_i}(F_{x_i} = u_i)$ is the marginal distribution of each ith variable (x_i), the cumulative distribution function, F, expresses as Eq. (2):

$$\boldsymbol{F}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = \boldsymbol{C}[\boldsymbol{F}_{x_1}(\boldsymbol{x}_1), \boldsymbol{F}_{x_2}(\boldsymbol{x}_2), \ldots, \boldsymbol{F}_{x_n}(\boldsymbol{x}_n)] = \boldsymbol{C}(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n) \quad (2)$$

$$\boldsymbol{C}(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n) = \boldsymbol{Pr}\{\boldsymbol{U}_1 \leq \boldsymbol{u}_1, \boldsymbol{U}_2 \leq \boldsymbol{u}_2, \ldots, \boldsymbol{U}_n \leq \boldsymbol{u}_n \quad (3)$$

where C in Eq. (3) is the copula function of the joint distribution function F. Therefore, the probability distribution function $c(u_1, \ldots, u_n)$ can be calculated as Eq. (4).

Table 7. The calculated bandwidth of different methods in the selected cell.

Bandwidth selection method	MISE	AMISE	MLCV	UCV	BCV	CCV	MCV	TCV
Bandwidth value	17.61	17.77	37.93	24.41	23.44	17.74	23.67	24.42
Mean Improved data	61.093	61.093	61.32	61.19	61.18	61.09	61.18	61.19

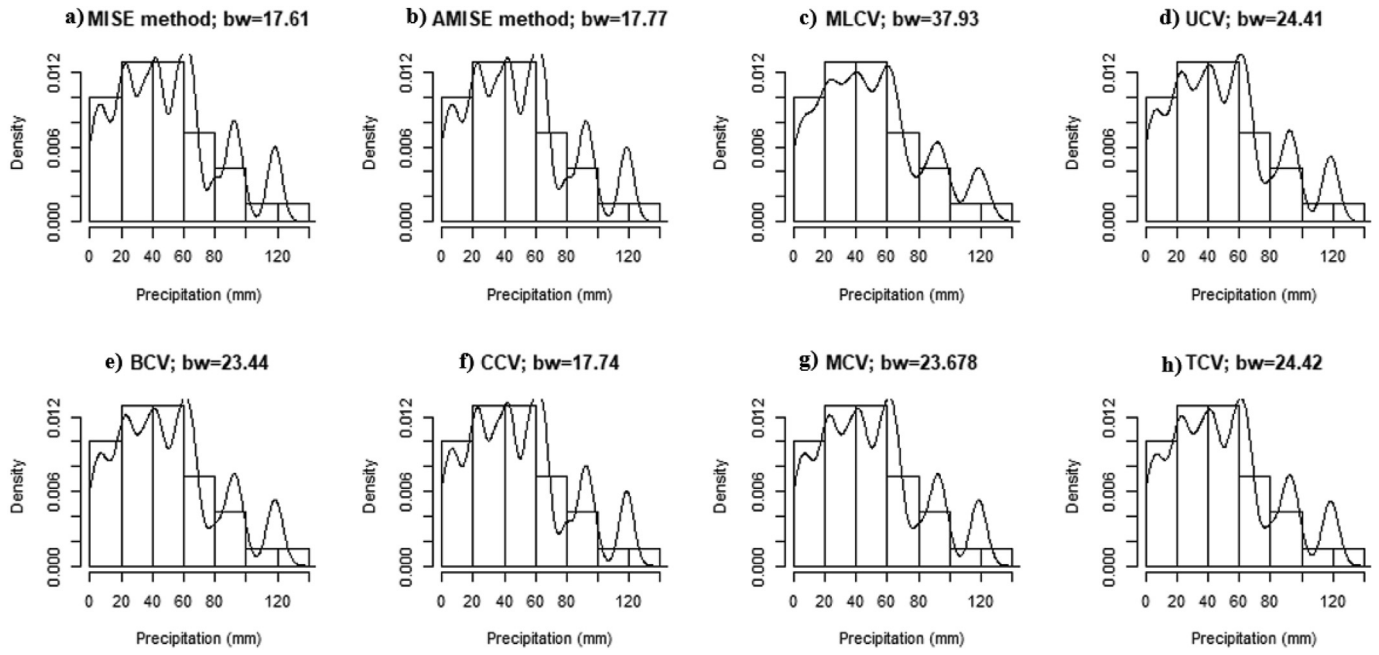

Figure 7. The example data histogram in front of Normal Kernel Density Distribution for different bandwidths calculated by a) MISE, b) AMISE, c) MLCV, d) UCV, e) BCV, f) CCV, g) MCV and, h) TCV methods.

$$\boldsymbol{c}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)=\frac{\partial^n \boldsymbol{C}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)}{\partial \boldsymbol{u}_1\ldots\partial \boldsymbol{u}_n} \tag{4}$$

Hence, the joint distribution function based on copula (f), which will be used as an input of the next session, can be obtained as Eq. (5):

$$\boldsymbol{f}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)=\boldsymbol{c}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)\prod_{i=1}^{n}\boldsymbol{f}_{\boldsymbol{x}_i}(\boldsymbol{x}_i) \tag{5}$$

Various Copula families have been introduced so far. Application of Archimedean (Tahroudi et al., 2020; Xu et al., 2019) and other Copula families are common in hydrological problems (Dehghani et al., 2019; Li et al., 2019; Zhang and Singh, 2007). Hence, this study used some of these families (i.e. Gaussian, student's T, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7, BB8, rotated Clayton, rotated Gumbel, rotated Joe, rotated BB1, rotated BB6, rotated BB7, rotated BB8 Copula functions). Among these functions, Clayton, Gumbel, Joe and BB suffer from this challenge that they couldn't consider the negative correlation. According to Eqs. (6), (7), and (8), using the 90,180 and 270-degrees rotation as an alternative solution may be considered to overcome this deficiency.

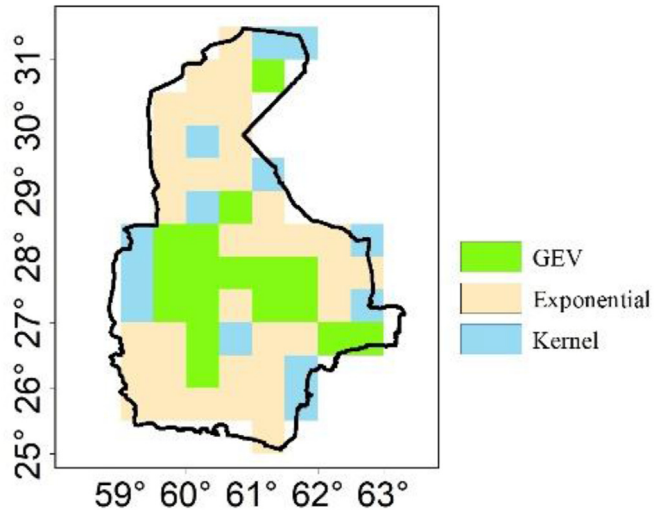

Figure 8. The dispersal of parametric and Normal Kernel Density Distributions.

$$\boldsymbol{C}_{90}(\boldsymbol{u}_1,\boldsymbol{u}_2)=\boldsymbol{u}_2-\boldsymbol{C}(1-\boldsymbol{u}_1,\boldsymbol{u}_2) \tag{6}$$

$$\boldsymbol{C}_{180}(\boldsymbol{u}_1,\boldsymbol{u}_2)=\boldsymbol{u}_1+\boldsymbol{u}_2-1+\boldsymbol{C}(1-\boldsymbol{u}_1,1-\boldsymbol{u}_2) \tag{7}$$

$$\boldsymbol{C}_{270}(\boldsymbol{u}_1,\boldsymbol{u}_2)=\boldsymbol{u}_1-\boldsymbol{C}(\boldsymbol{u}_1 1-,\boldsymbol{u}_2) \tag{8}$$

To identify each of the aforementioned functions' satisfaction rates in distribution over multi variables, the Akaike Information Criterion (AIC) method was implemented. AIC is a commonly used alternative criteria to discriminate among copula functions. After fitting available copulas using maximum likelihood, the criteria were computed for all copula families and the family with the minimum value was chosen. In bivariate copula families, the AIC is defined as Eq. (9). Where k is the number of parameters, θ is the parameter and other variables have been introduced before.

$$\boldsymbol{AIC}=-2\sum_{i=1}^{N}\ln[\boldsymbol{c}(\boldsymbol{u}_{i,1},\boldsymbol{u}_{i,2}|\boldsymbol{\theta})]+2\boldsymbol{k} \tag{9}$$

At the end of this stage, the selected copula function will be applied in the Bayesian equation to estimate CPDF.

3.2.2. Copula based Conditional Probability Density Function (CPDF)

At this phase, by using the best copula function identified in the previous stage and the Bayesian equation (Equation 10a), the creation of CPDF among the observational data and the raw forecast data will be attempted. In this regard, CPDF of observational data given each forecast data at its time step will be calculated. The joint distribution of forecast and observational data in Bayesian structure is expressed as follows:

$$\boldsymbol{f}(\boldsymbol{f},\boldsymbol{o})=\boldsymbol{f}(\boldsymbol{f})\boldsymbol{f}(\boldsymbol{o}|\boldsymbol{f}) \tag{10a}$$

$$\boldsymbol{f}(\boldsymbol{o}|\boldsymbol{f})=\frac{\boldsymbol{f}(\boldsymbol{f},\boldsymbol{o})}{\boldsymbol{f}(\boldsymbol{f})} \tag{10b}$$

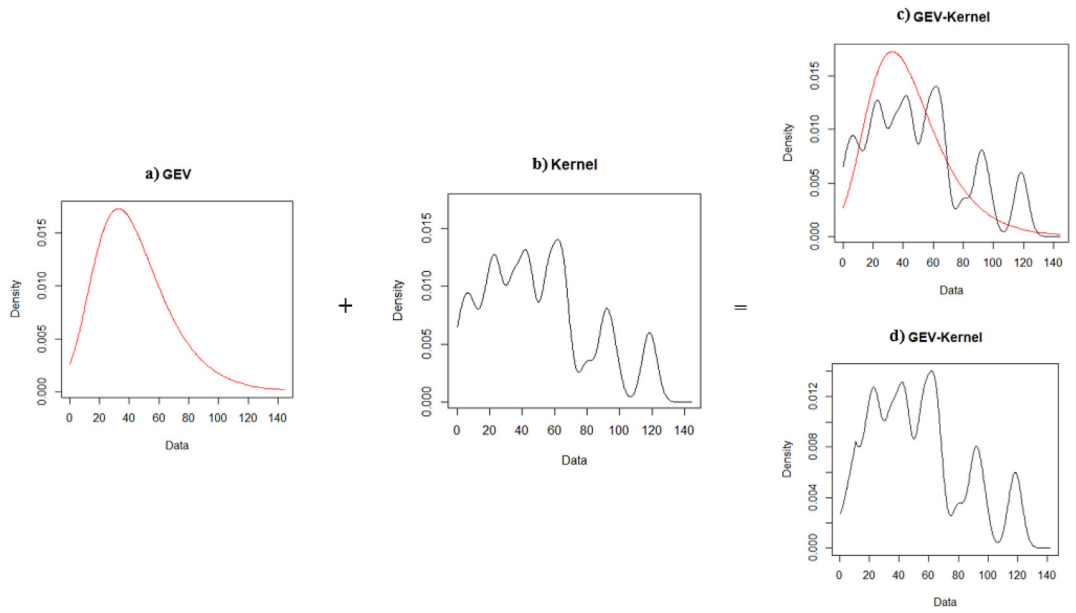

Figure 9. Steps of making GEVKer distribution a) GEV, b) Kernel, c) GEV (red) and Kernel (white) overlapping, and d) GEV-Kernel distributions.

Table 8. K–S and BIC criteria value, total distributions.

Distribution	K–S			BIC		
	Observation	Hypothesis test	Hindcast	Hypothesis test	Observation	Hindcast
GEV*	0.99	Accept	0.9287	Accept	-127.61	-104.22
Kernel	0.9984	Accept	0.9514	Accept	-125.73	-109.37
GEVKer	0.997	Accept	1	Accept	-147.2089	-139.95

* As seen in Table 6, the GEV distribution is the best parametric distribution in the selected cell.

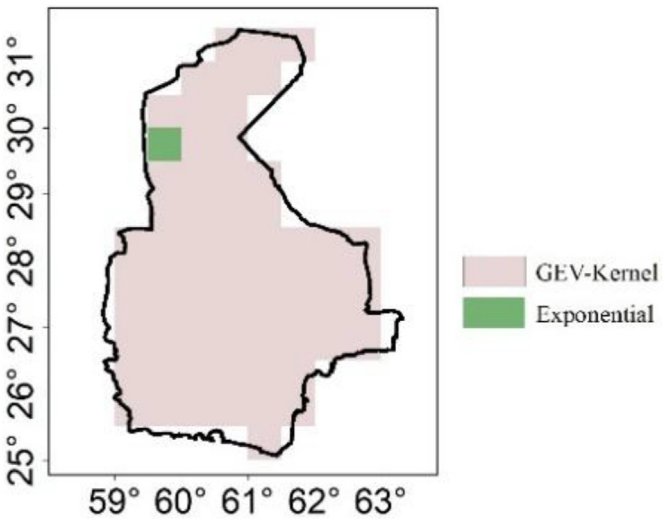

Figure 10. The best marginal distributions in the study area.

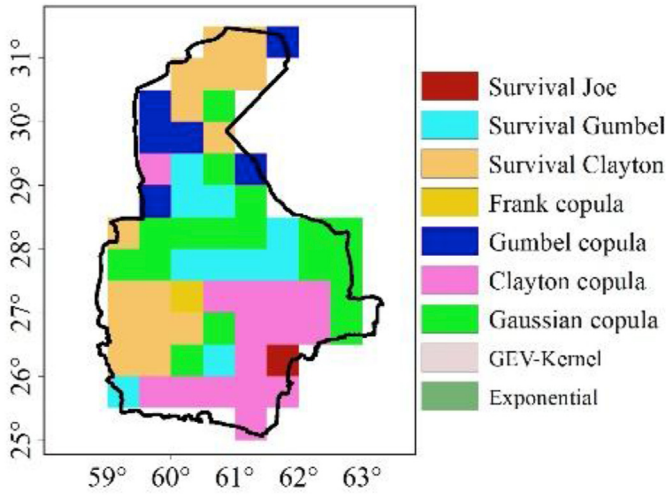

Figure 11. Appropriate copula functions for March observed data.

Therefore, with replacing/substituting Eq. (5) (modified for bivariate joint distribution function) in Eq. (10b) the CPDF can be calculated as Eq. (11):

$$\frac{f(s_t|f_t)}{} = \frac{c(U_s = u_s, U_f = u_f) f(f_t) f(s_t)}{f(f_t)} \quad = c(U_s = u_s, U_f = u_f) f(s_t) \tag{11}$$

In the last equation $f(s_t|f_t)$ is the CPDF in time t, $f(s_t)$, $f(f_t)$ is the marginal distributions of the sample from the observation and the forecast at time t. The sample data has 500 random data with the same distribution of observation data in the hindcast period (Khajehei and Moradkhani, 2017). For each specific raw forecast data, the above-mentioned process of creating CPDF will be carried out.

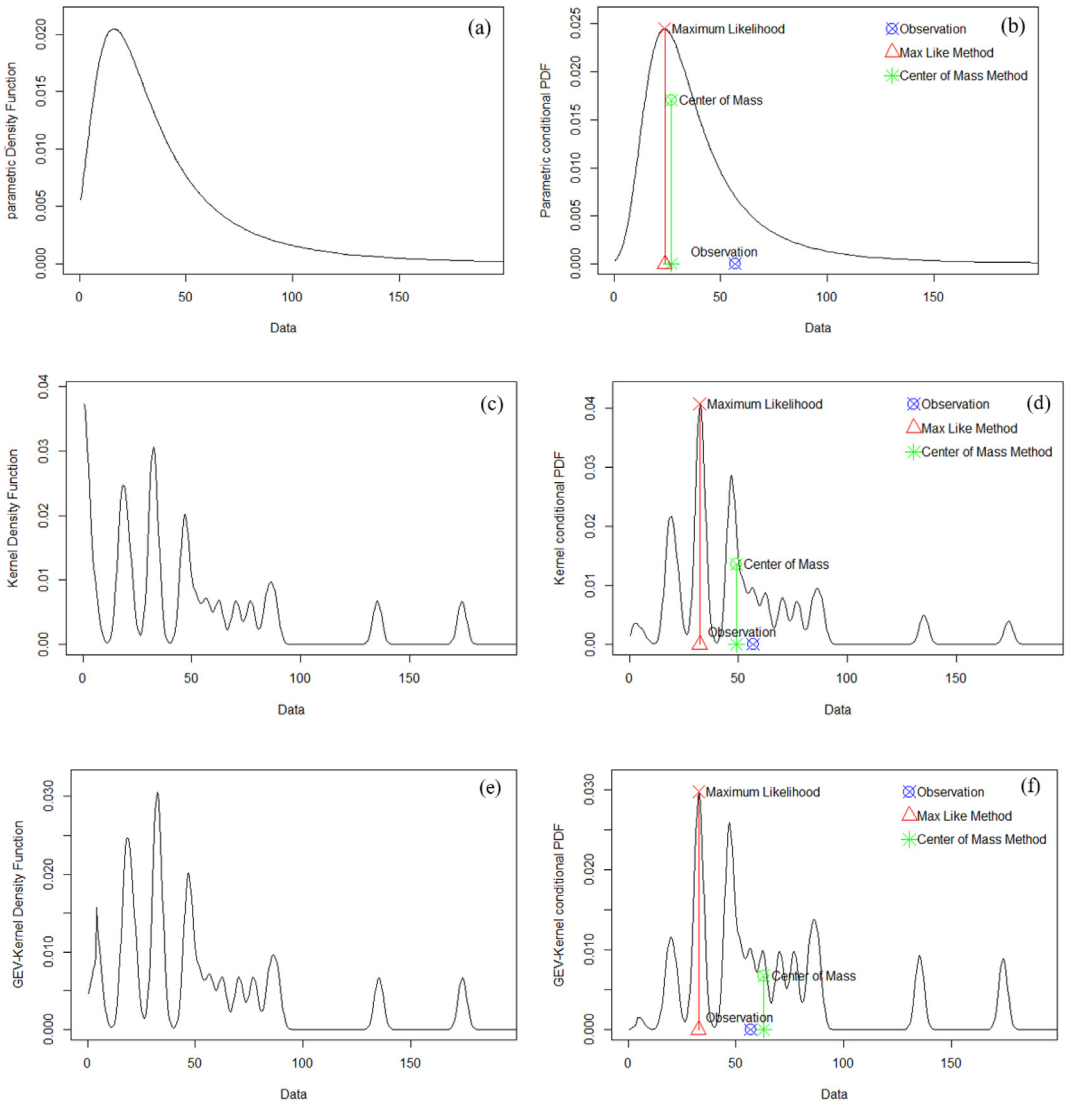

Figure 12. CPDF for different marginal distributions proposed by ML and CM methods, a) density function of parametric distribution, b) CPDF for parametric distribution, c) density function of Kernel distribution, d) CPDF for Kernel distribution, e) density function of gEV-Kernel distribution, f) CPDF for GEV-Kernel distribution.

3.2.3. Determination of improved forecast data

The CPDF shows the probability of the sample observational data after taking a single raw prediction data into account. There are diverse schemes for picking out the improved forecast found on CPDF (Khajehei and Moradkhani, 2017; Madadgar and Moradkhani, 2012). The CPDF based on parametric marginal distributions has a single maximum value. Previous researches have selected the maximum point of CPDF function (Maximum Likelihood) as the desired improved data. The non-parametric or semi-parametric marginal distributions cause the CPDF function to return more inadvertently than a single relative maximum point. To respond to this challenging issue, in this study a novel technique is introduced. Given the CDPF, each sample observation data has a probability of occurrence. Therefore, it can be inferred that each of the observational data has an effect on the selection of the improved data as a proportion of the probability of occurrence. Based on this logic, the Center of Mass (CM) of the CDPF curve can be introduced as the optimal value of improved data. Eq. (12) show how the Maximum Likelihood (ML) and the proposed method are quantified.

$$x_{im} = \begin{cases} x_{ML} & if\ f_c(x_{ML}) = \max(f_c) \\ \dfrac{\int_{x_{min}}^{x_{max}} x f_c(x)dx}{\int_{x_{min}}^{x_{max}} f_c(x)dx} \end{cases} \quad (12)$$

In the last equation f_c is CPDF, x is the sample observational data and x_{im} is the improved forecast data. In this research, the two previously mentioned improved forecasts picked out methods are used in different marginal distributions, listed in Table 5.

To evaluate/validate how much the NMME predictions (the raw ones and the improved ones) are consistent with the observation data, the Kling-Gupta Efficiency (KGE) test, is used here. The KGE (Equation 13) is

Table 9. The results of the ML and CM methods for each marginal distribution. (cell No:20,10. March).

Sample Cell	year	Observation	Hindcast	PAR (CM)	PAR (ML)	Kernel (CM)	Kernel (ML)	GEVKer (CM)	GEVKer (ML)
Prec. (mm)	1991	56.76	42.93	29.08	25.71	49.40	32.659	60.71	33.07

Table 10. The postprocessing results by using the ML and CM methods for each marginal distribution.

Month	Period	forecast	observation	Par-CM	Par-ML	Ker-CM	Ker-ML	GEVKer- CM	GEVKer-ML
Jan	Forecast	29.80	25.95	19.91	18.13	27.29	20.58	**27.25**	20.69
	Hindcast	18.26	17.79	16.09	12.06	19.67	12.64	**18.81**	**18.81**
Feb	Forecast	33.27	29.49	19.78	19.62	**31.23**	25.14	**31.28**	25.33
	Hindcast	33.77	44.38	20.81	20.23	31.73	23.79	**32.75**	**32.75**
Mar	Forecast	39.51	33.79	24.41	22.67	**35.60**	27.25	**35.77**	27.26
	Hindcast	29.64	45.11	21.70	18.00	**29.40**	20.43	**29.61**	**29.61**
Apr	Forecast	29.88	16.45	9.39	7.95	**18.62**	11.72	**18.48**	11.56
	Hindcast	28.33	33.19	9.50	7.89	**18.68**	11.30	18.62	18.62
May	Forecast	10.20	8.04	5.22	4.86	**9.48**	4.63	**9.41**	4.54
	Hindcast	10.98	13.76	5.54	5.33	**10.16**	4.97	**10.15**	**10.15**
Jun	Forecast	15.48	6.89	2.11	1.36	8.67	2.58	**8.51**	2.53
	hindcast*	-	-	-	-	-	-	-	-
Jul	Forecast	33.19	15.93	10.49	7.22	17.86	8.74	**16.87**	8.41
	Hindcast	28.83	12.35	**9.87**	6.23	15.98	6.57	15.03	15.03
Aug	Forecast	37.07	14.28	8.42	6.44	16.09	8.40	**15.55**	8.36
	hindcast*	-	-	-	-	-	-	-	-
Sep	Forecast	18.89	4.6	1.5	1.19	6.11	1.9	**6.04**	1.9
	Hindcast	21.19	15.89	1.64	1.66	**7.54**	6.44	7.35	7.35
Oct	Forecast	7.63	3.36	0.74	0.53	4.64	1.15	**4.61**	1.08
	Hindcast	8.75	12.67	0.83	0.70	**5.65**	2.44	5.58	5.58
Nov	Forecast	10.23	4.54	2.34	1.96	**5.41**	2.09	**5.42**	2.04
	Hindcast	13.53	11.98	2.79	2.71	**6.92**	3.40	6.91	6.91
Dec	Forecast	27.23	19.48	11.44	10.28	**22.31**	15.24	**22.33**	15.36
	Hindcast	21.03	11.80	10.58	8.27	19.07	**10.55**	19.37	19.37

* There are not enough data for the hindcast period.

the revised and enriched form of Nash-Sutcliffe Efficiency (NSE) (Gupta et al., 2009). It is a useful tool for the similarity assessment of the forecast and observational data.

$$\boldsymbol{KGE} = 1 - \boldsymbol{ED} \tag{13}$$

$$\boldsymbol{ED} = \sqrt{(\boldsymbol{r}-1)^2 - (\boldsymbol{\alpha}-1)^2 - (\boldsymbol{\beta}-1)^2} \tag{14}$$

Eq. (14), *ED*, is expressed as the Euclidean distance between estimated and observed data which is a function of correlation (r), the ratio of the variance of the forecast to the variance of observation (α) and the ratio bias (β) (Khajehei and Moradkhani, 2017).

4. Results

4.1. Validation of GPCC precipitation data

As mentioned in section 2.2.1 the GPCC precipitation data were used as the observation data during the study period. To evaluate the accuracy of using GPCC data, the KGE and CC values of rain-gauge stations with recorded precipitation values during 1981–2010 were compared with the corresponding GPCC precipitation values in the corresponding 0.5 by 0.5° cells.

Box plots of criteria values (KGE and CC) for each rain-gauge station over 12 months from 1981 to 2010 are shown in Figure 3. As seen in this figure, based on KGE values (all of them are above 0.6) there is a good agreement between the two datasets. Also, the CC with values greater than 0.95 for most time series at these stations, confirms that the GPCC data is consistent with the rain-gauge station data.

4.2. Raw forecast validation

To evaluate the accuracy of raw forecast data, first, the KGE criterion was used to find the reliability of raw NMME precipitation data due to the observation data (GPCC here). As suggested by (Khajehei and Moradkhani, 2017), the acceptable amount of this parameter was considered more than 0.6 here. Otherwise, the use of raw data without doing any post-processing is not logical. As a non-limiting example, the results of KGE for the rainiest (March) and driest (October) months of the study area are illustrated in Figure 4. As seen, in all extent of the study area the amount of KGE is lower than 0.6, hence, it is essential to do the post-processing cell by cell. In the following, to find out the best fitting distribution with the historical data, three types of distributions were examined.

4.3. Parametric distributions

The best (most suitable) parametric distribution in each cell is selected based on the following sequential stages. First, different parametric distributions are applied for each time-series. Then, the KS and BIC indexes are evaluated for each distribution. Finally, the best (most fitted) parametric distribution can be selected. As a non-limiting example, the observation data of a 0.5-degree cell in 62.25° of longitude and 33.25° of latitude coordinates are used here. Table 6 shows the results of fitting different parametric distributions to the specified time-series of the mentioned cell. According to this table, based on the K–S test, the Exponential, Normal, LogNormal, and GEV distributions are the acceptable ones (K–S>0.05) while the GEV distribution with the least BIC value can be introduced as the best among them.

Figure 5 shows the histogram of precipitation within the mentioned cell during the hindcast period (column bars) versus different fitted parametric distributions (red curves). This figure can present a better perception of judgment about the suitability of each distribution. In this way, the most fitted distribution for each cell can be selected as the most suitable distribution. For example, in the corresponding cell, the GEV distribution is the best one.

With repeating the same calculation in other cells, the EXP and GEV distributions were found dominant compared to the other parametric distributions across the study area (Figure 6).

4.4. Non-parametric distributions

As previously stated (section 3.1), determining the bandwidth is the first prerequisite to find the best fitting kernel distribution function. For the chosen cell, Table 7, presents the result of bandwidth values proposed by different methods. In this cell, the mean observation and forecast values are 49.27 and 70.25 mm respectively. As seen in Table 7, the suggested bandwidth values have led to almost similar results (around 61 mm). Figure 7 illustrates how the selection of different values of bandwidth affects the accuracy of the fitted kernel distribution. As a general rule, the best bandwidth is the one which is accompanied by the most improvement in the hindcast period. However, the obtained results show that the improved forecasts are not sensitive relative to the different bandwidth values, the bandwidth value of CCV method has the highest priority for the selection.

Figure 8 shows the dispersal of parametric distributions and Normal Kernel Density Distribution in the researched region. As seen in this figure, the kernel distribution function in only some limited cells could be more adapted to the hindcast data than the parametric distributions. Therefore, the use of a semi-parametric distribution function that can take advantage of parametric and non-parametric distributions simultaneously has been evaluated in the following.

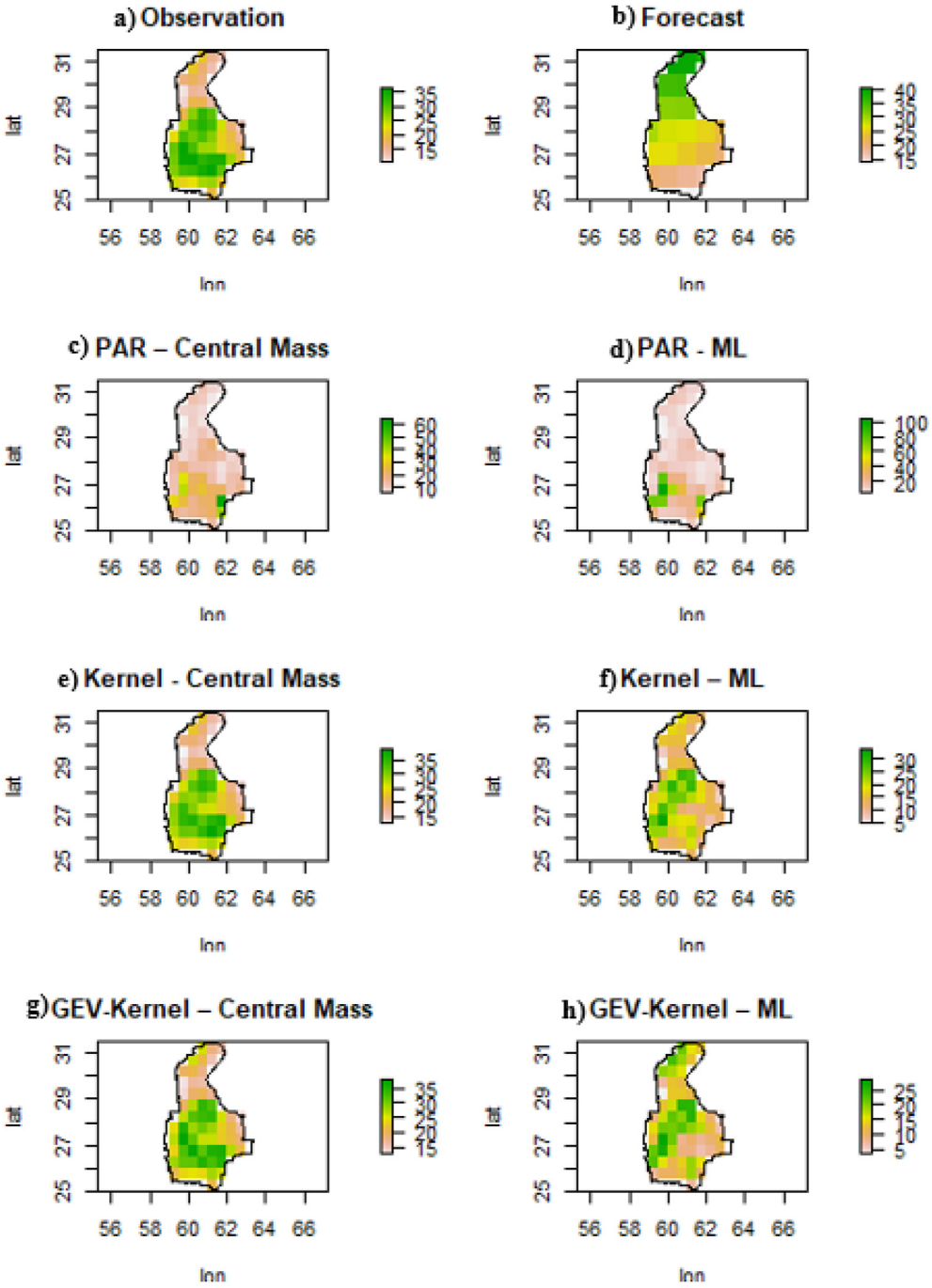

Figure 13. Spatial dispersion of the March precipitation (hindcast period) a) Observation, b) Forecast, c) PAR – Central Mass, d) PAR –ML, e) Kernel – Central Mass, f) Kernel – ML, g) GEV-Kernel – Central Mass, h) GEV-Kernel – ML.

4.5. Semi-parametric distribution

As seen in Figure 8, the GEV and EXP distributions have the most agreement with observation data in the region. As the GEV can efficiently cover the drawbacks of non-parametric in boundary limits (extreme values), it was used as a complementary function for integration with kernel distribution. In other words, at this stage, a combination of kernel and GEV functions was used to create a semi-parametric distribution function. The formation of this distribution is shown in Figure 9.

For the selected cell, Table 8 shows the K–S test and BIC criterion values for all the used parametric, non-parametric and semi-parametric distributions. As seen, based on the K–S test, all of the obtained results are acceptable but based on the BIC criterion, the GEVKer distribution is the best one. The same procedure was applied for all the cells. Across the study area, only one cell followed the parametric (exp) distribution and the others fitted with GEVKer distribution (Figure 10).

4.5.1. Determining the best copula functions and creating CPDF

As described in section 3.2.1, to find the best joint distribution function between the observation and raw forecast data in the hindcast period, the Archimedean and Elliptical Copula families were assessed. Then, the selected copula function for each cell in the hindcast period

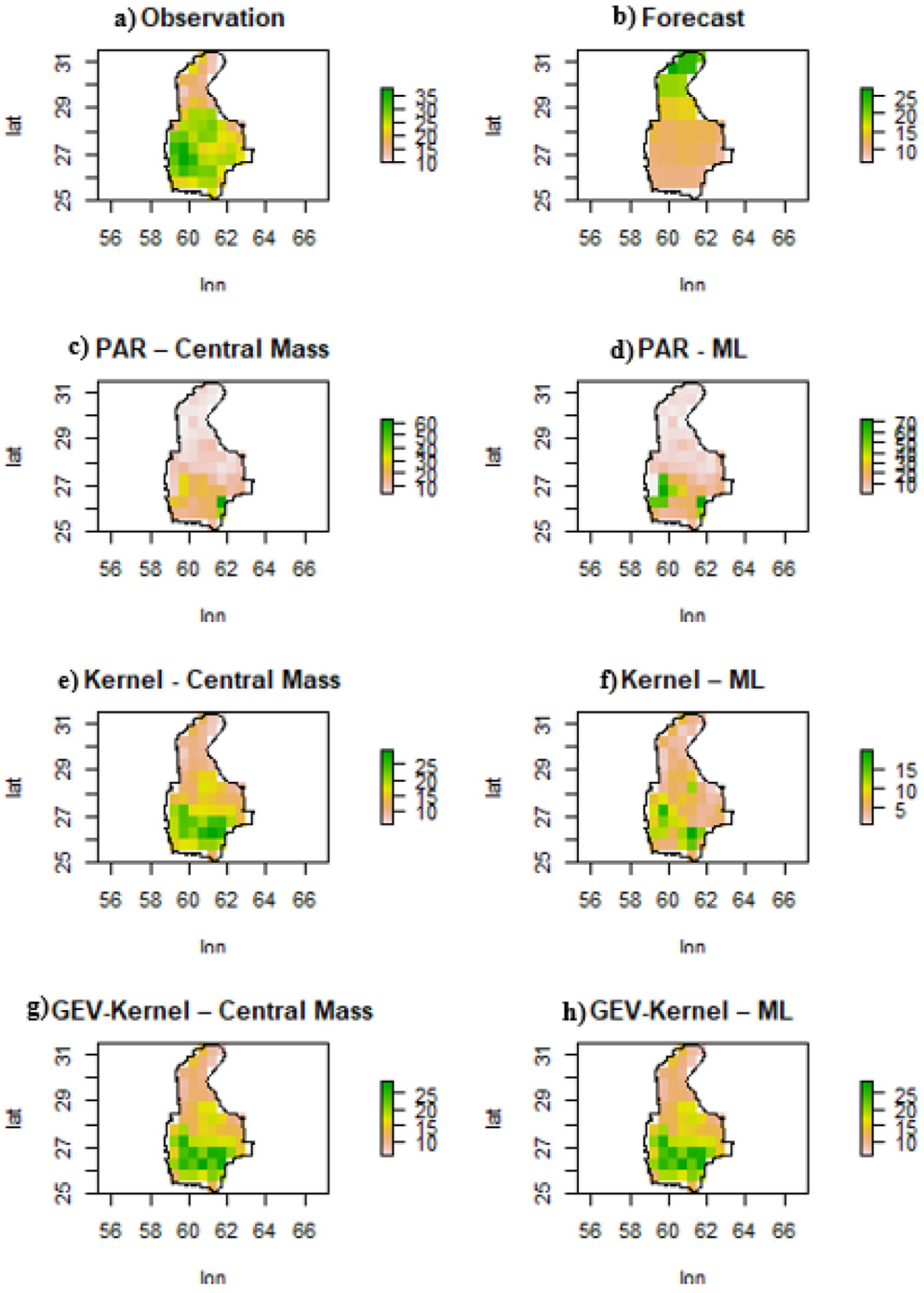

Figure 14. Spatial dispersion of the March precipitation (forecast period) a) Observation, b) Forecast, c) PAR – Central Mass, d) PAR –ML, e) Kernel – Central Mass, f) Kernel – ML, g) GEV-Kernel – Central Mass, h) GEV-Kernel – ML.

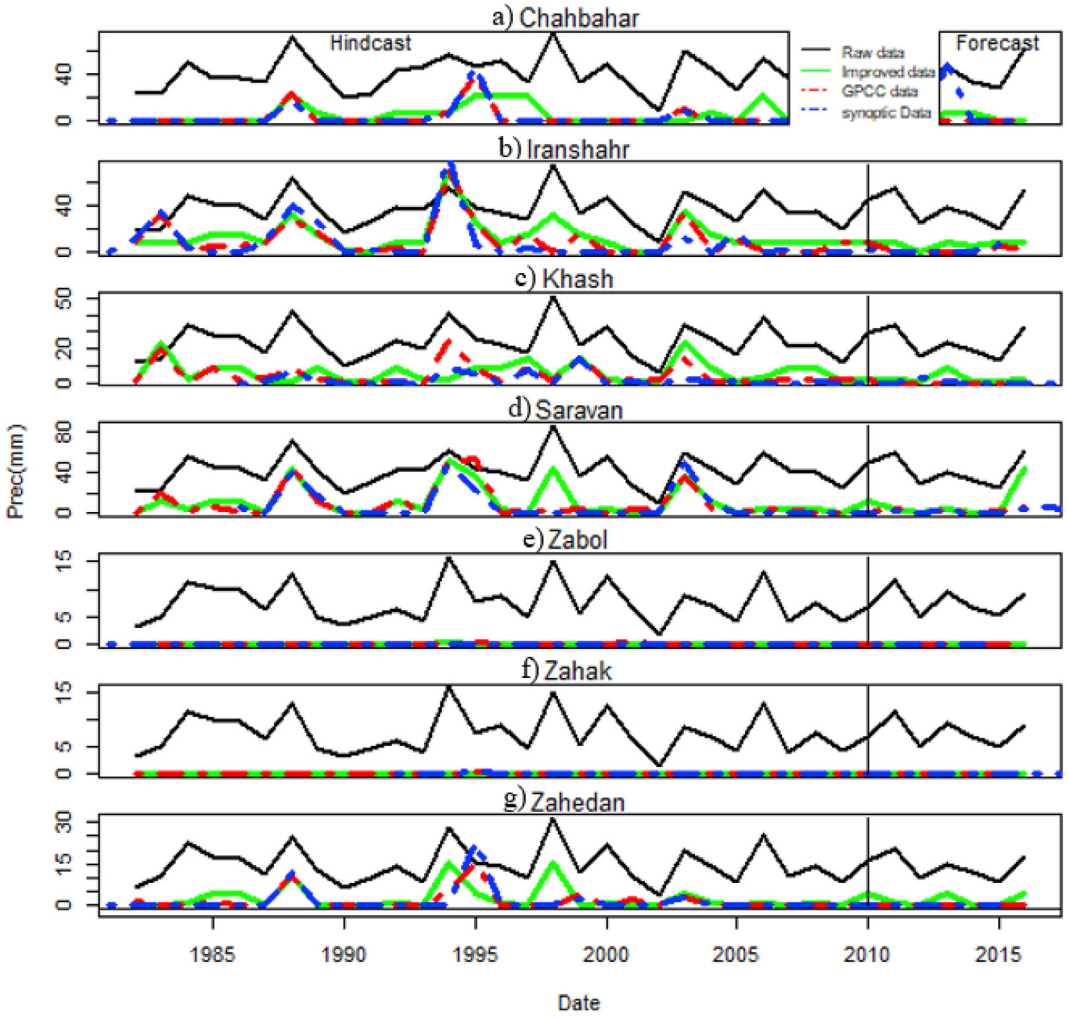

Figure 15. Comparison of estimated precipitation between real observation, GPCC, raw NMME models and improved NMME Models in observational stations: a) Chabahar, b)Iranshahr, c)Khash, d)Saravan, e)Zabol, f)Zahak, g)Zahedan.

Table 11. The correlation and mean values of time-dependent time series.

Data	Correlation	Mean Value
observation	-	10.0126
forecast	0.4312	16.8774
Par-ML	0.54462	3.27181
Par-CM	0.53696	3.51763
Ker-ML	0.55334	6.57218
Ker-CM	0.65302	11.6302
GEVKer-ML	0.56003	6.35002
GEVKer-CM	0.64978	11.6011

(based on the AIC test) is used to improve the estimated data in the hindcast and forecast periods. As a non-limiting example, for the rainiest month (March), the best copula functions introduced in each cell are illustrated in Figure 11. According to this figure, it can be seen that among different copula functions investigated in this study, only 7 were introduced as the best selected joint distribution across the area.

After the selection of the best joint distribution of the observed and forecast data, CPDF is created for each predicted precipitation. For this purpose, the selected copula function is applied in the Bayesian equation as discussed in section 3.2.2.

4.6. Improved forecast extraction

The essential part of post-processing is the selection of the best-improved precipitation among the sample observational data. To show the effect of using different marginal distributions and the method of selecting the best-improved forecast of precipitation, the sample cell has been examined. In Figure 12, the first column represents the sample observational data based on the parametric, non-parametric and semi-parametric marginal distributions during the rainy month (March). The second column of Figure 12 is equivalent to the CPDF of mentioned cell for predicted precipitationin 1991. Also, the results of the ML and CM methods are denoted here.

The results show that there is no significant difference in the results of the two mentioned methods for the parametric marginal distributions. On the other hand, these results are impressively different for non-parametric or semi-parametric marginal distributions because of their irregular CPDF shapes. Also, GEVKer marginal distribution may be able to overcome the over-estimating of kernel distribution deficiency in the lower boundary. The details of post-processing results for this cell is shown in Table 9.

The results of mean improved precipitation data calculated by the CM and ML methods in different months are presented in Table 10.

In Table 10, the closest values to the observed data are shown in bold. As seen in this table, the Kernel and GEVKer marginal distributions have

Figure 16. The result of forecasts for the spatial dependent time series a) the correlation coefficient b) KGE criterion.

Table 12. Anomaly table for the October.

	Forecast			Hindcast		
	Min negative anomaly	Mean anomaly	Max positive anomaly	Min negative anomaly	Mean anomaly	Max positive anomaly
forecast	-13.29	-3.07	6.26	-5.60	9.72	13.55
Par-ML	-29.72	9.74	3.06	-22.17	9.70	14.82
Par-CM	-13.59	9.48	7.71	-20.52	10.39	11.89
Ker-ML	-	7.00	7.00	-	16.89	16.89
Ker-CM	-1.26	0.67	-0.94	-1.95	8.14	8.456
GEVKer-ML	-1.67	10.20	9.11	-1.94	7.501	7.96
GEVKer-CM	-1.88	1.73	-1.10	-1.94	7.50	7.96

caused the best improvement. Meanwhile, the CM method has presented better values than the ML method. Figure 13 and Figure 14 display the improved predicted data of March for the total study area in the hindcast and forecast periods.

These results can confirm that the combined distribution has led to higher improvement. In summary, the introduced semi-parametric distribution is more flexible (like nonparametric distributions) and more accurate than the other ones especially in the boundaries (like parametric GEV distributions).

In terms of temporal variation of forecast accuracy, Figure 15 shows the comparison of precipitation time-series of the recorded observation, GPCC, raw NNME forecast models and the improved NMME data (GEV-Ker-CM) in location of 7 rain-gague station cells during the July (as a non-limiting example) in hindcast and forecast periods. As seen in this figure, the improved NMME model are much closer to the actual recorded values than the raw NMME data.

4.7. A deeper investigation of results

According to the methodology, the first step of post-processing was data classification into 12 monthly separated time series. Although most of the post-processing researches which are based on Copula functions use spatial and temporal dependent Copulas, however, operational comfort often leads researchers to classify database in spatial and temporal independent time series. These time series can provide the possibility of using spatial and temporal independent copula functions. Therefore, the necessity of examining the sensitivity of the results to the initial assumption has been studied. To achieve this target, new time series are built-in reliance on temporal dependency. These time series are consisting of whole data in 1981–2010 for each cell at the monthly timestep (384 members). Table 11 indicates the correlation coefficient and mean values of these time series.

As seen in Table 11 not only the time dependency of data is preserved but also it is improved after post-processing. The GEVKer-CM, Ker-CM, GEVKer-ML and Ker-ML methods have the maximum correlations respectively. Moreover, concerning the mean values, the GEVKer-CM and Ker-CM methods have the best performance in forecast rectification respectively.

For assessment of the spatial dependency, the raw and the improved data of each month in the entire study area are compared. The correlation coefficient and KGE criteria are demonstrated in Figure 16 for the months with the most (March) and the least (October) rainfall.

The results indicate that the correlation coefficient values of improved data set based on Ker-CM and GEVKer-CM methods in hindcast period are near 1. Also, in the forecast period these two methods were able to improve the correlation coefficient from negative values to above0.75 for rainy months and above 0.25 for the low rainfall months.

As mentioned earlier the KGE criterion is the best to scrutinize the post-processing influence and to measure its satisfaction. As seen in Figure 16b, the KGE values for the March in hindcast and forecast periods by two methods of Ker-CM and GEVKer-CM are satisfactory (KGE>0.6). The validity of this criterion is again confirmed for the October in hindcast period for the mentioned methods. However, the values of this criterion for October in the forecast period are negative. The reason for these unpleasant results is the low rainfall values in this month. Based on the formulation of KGE, the three terms of KGE equation are rational. Therefore, its values are highly sensitive to the denominator of those ratio terms. In this case, a small absolute variation in the precipitations' values can negatively impact the related ratio terms and thus produce negative KGE values (Santos et al., 2018). So, the anomaly and mean values of October precipitation are selected to evaluate the post-processing influence in low rainfall months (Madadgar et al., 2016). The monthly precipitation anomalies illustrate that the Ker-CM, GEVKer-ML and GEVKer-CM methods have smaller anomaly values (Table 12).

Despite initial assumption in time series classification, the results of monitoring the temporal and spatial dependency of improved data show the preservation and amelioration of these dependencies. Additionally, the Ker-CM and GEVKer-CM methods lead to the best results.

The Ker-ML and GEVKer-ML methods haven't led to acceptable mean values. However, according to the temporal and spatial dependent time series and relative CC and KGE amounts, they are more robust in an indication of the rate of the change.

5. Conclusion

This paper investigated the accuracy of the NMME forecast data and the usage of the statistical Copula method for improvement of the model output in Sistan and Baluchestan province, Iran. In this regard a three-step framework is proposed based on 1) Evaluating the conformity of GPCC data with rain-gauge recorded observational precipitation 2) Examining different statistical distributions and presenting a combined (semi-parametric) distribution 3) Examining the common method of determining the improved data based on the CPDF and proposing a new method. This research uses four NMME models containing 54 ensemble members for the hindcast period (1982–2010) and 58 ensemble members for the forecast period (2012–2016). To evaluate the accuracy of the ensemble's mean, the GPCC observational database is used as the observation data. The KGE values reveal the requirement of model output improvement. In this paper, the copula-based Bayesian approach is used as post-processing. This method results in a CPDF whose accuracy is a function of marginal distributions of Copula functions. Here, three kinds of marginal distributions as parametric (Exponential, Normal, Gamma, LogNormal and GEV), non-parametric distributions and a novel semi-parametric distribution are employed. After CPDF production, the choice of best-improved data is an essential step. This paper uses the common method based on Maximum Likelihood and suggests a novel method that chooses the center of mass of CPDF. Finally, the achievements of this paper may be summarized below.

1 The investigation of different marginal distributions shows that the GEVKer distribution has the most agreement with most of the time series. Furthermore, it can solve the kernel distributions' problem in boundaries.
2 Application of the Maximum Likelihood and Central Mass methods in parametric marginal distributions have nearly the same results. But they lead to significantly different results in the case of the kernel and GEVKer marginal distributions.
3 The non-parametric and semi-parametric distributions give better improvement.
4 The usage of non-parametric and semiparametric marginal distributions by the ML method, in comparison with parametric one, can present the proper rate of change (CC).
5 When the data values are low or the variation of data is small, the CC and KGE criteria wouldn't be the appropriate parameters to investigate the performance of post-processing. In these cases, the anomalies will be used.
6 Despite initial simplifying assumptions on time series classification based on temporal and spatial independence, the results represent amelioration of temporal and spatial independence.

Declarations

Author contribution statement

Farhad Yazdandoost: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Mina Zakipour: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Ardalan Izadi: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by Sistan and Baluchestan Regional Water Company.

Data availability statement

Data will be made available on request.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Adamowski, K., 1985. Nonparametric kernel estimation of flood frequencies. Water Resour. Res. 21 (11), 1585–1590.
Akramin, M., Takahashi, A., Husnain, M., Chuan, Z., 2020. KS test for crack increment in probabilistic fracture mechanics analysis. In: Paper Presented at the IOP Conference Series: Earth and Environmental Science.
Al Zawad, F.M., Aksakal, A., 2010. Impacts of climate change on water resources in Saudi Arabia. In: Global Warming. Springer, pp. 511–523.
Azizi, G., Miri, M., Mohamadi, H., Pourhashemi, M., 2015. Analysis of relationship between forest decline and precipitation changes in Ilam Province. Iran. J. For. Popl. Res. 23 (3).
Bárdossy, A., 2006. Copula-based geostatistical models for groundwater quality parameters. Water Resour. Res. 42 (11).
Becker, E., Kirtman, B.P., Pegion, K., 2020. Evolution of the North American multi-model ensemble. Geophys. Res. Lett. 47 (9), e2020GL087408.
Ben Alaya, M., Zwiers, F., Zhang, X., 2020. An evaluation of block-maximum-based estimation of very long return period precipitation extremes with a large ensemble climate simulation. J. Clim. 33 (16), 6957–6970.

Brown, J.D., Seo, D.J., 2013. Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions. Hydrol. Process. 27 (1), 83–105.
Cherubini, U., 2004. Pricing swap credit risk with copulas. In: Paper Presented at the EFMA 2004 Basel Meetings Paper.
Darand, M., Zand, K.S., 2016. Evaluation of the Accuracy of the Global Precipitation Climatology Center (GPCC) Data over Iran.
de Melo Mendes, B.V., de Souza, R.M., 2004. Measuring financial risks with copulas. Int. Rev. Financ. Anal. 13 (1), 27–45.
De Michele, C., Salvadori, G., 2003. A generalized Pareto intensity-duration model of storm rainfall exploiting 2-copulas. J. Geophys. Res.: Atmosphere 108 (D2).
Dehghani, M., Saghafian, B., Zargar, M., 2019. Probabilistic hydrological drought index forecasting based on meteorological drought index using Archimedean copulas. Nord. Hydrol 50 (5), 1230–1250.
Efromovich, S., 1999. Nonparametric regression for small samples. Nonparam. Curve Estim.: Methods Theory Appl. 118–180.
Favre, A.C., El Adlouni, S., Perreault, L., Thiémonge, N., Bobée, B., 2004. Multivariate hydrological frequency analysis using copulas. Water Resour. Res. 40 (1).
Feluch, W., Koronacki, J., 1992. A note on modified cross-validation in density estimation. Comput. Stat. Data Anal. 13 (2), 143–151.
Fenta Mekonnen, D., Disse, M., 2018. Analyzing the future climate change of Upper Blue Nile River basin using statistical downscaling techniques. Hydrol. Earth Syst. Sci. 22 (4), 2391–2408.
Fowler, H., Kilsby, C., 2003. A regional frequency analysis of United Kingdom extreme rainfall from 1961 to 2000. Int. J. Climatol.: J. Royal Meteorol. Soci. 23 (11), 1313–1334.
Frees, E.W., Carriere, J., Valdez, E., 1996. Annuity valuation with dependent mortality. J. Risk Insur. 229–261.
Frees, E.W., Valdez, E.A., 1998. Understanding relationships using copulas. North Am. Actuar. J. 2 (1), 1–25.
Gao, M., Mo, D., Wu, X., 2016. Nonstationary modeling of extreme precipitation in China. Atmos. Res. 182, 1–9.
Gent, P.R., Yeager, S.G., Neale, R.B., Levis, S., Bailey, D.A., 2010. Improvements in a half degree atmosphere/land version of the CCSM. Clim. Dynam. 34 (6), 819–833.
Ghosh, S., Mujumdar, P., 2007. Nonparametric methods for modeling GCM and scenario uncertainty in drought assessment. Water Resour. Res. 43 (7).
Gilleland, E., Katz, R.W., 2006. Analyzing seasonal to interannual extreme weather and climate variability with the extremes toolkit. In: Paper Presented at the 18th Conference on Climate Variability and Change, 86th American Meteorological Society (AMS) Annual Meeting.
Guidoum, A.C., 2015. Kernel Estimator and Bandwidth Selection for Density and its Derivatives. The Kedd Package, Version, 1.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377 (1-2), 80–91.
Habbema, J., Jdf, H., Van den Broek, K., 1974. A Stepwise Discriminant Analysis Program Using Density Estimation.
Han, Q., Hao, Z., Hu, T., Chu, F., 2018. Non-parametric models for joint probabilistic distributions of wind speed and direction data. Renew. Energy 126, 1032–1042.
Hattermann, F., Krysanova, V., Gosling, S.N., Dankers, R., Daggupati, P., Donnelly, C., Flörke, M., Huang, S., Motovilov, Y., Buda, S., 2017. Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins. Climatic Change 141 (3), 561–576.
Hjort, N.L., Glad, I.K., 1995. Nonparametric density estimation with a parametric start. Ann. Stat. 882–904.
Hürlimann, W., 2004. Multivariate Fréchet copulas and conditional value-at-risk. Int. J. Math. Math. Sci. 2004.
Jones, M., Kappenman, R., 1992. On a class of kernel density estimate bandwidth selectors. Scand. J. Stat. 337–349.
Kelly, K., Krzysztofowicz, R., 1997. A bivariate meta-Gaussian density for use in hydrology. Stoch. Hydrol. Hydraul. 11 (1), 17–31.
Khajehei, S., Moradkhani, H., 2017. Towards an improved ensemble precipitation forecast: a probabilistic post-processing approach. J. Hydrol. 546, 476–489.
Kim, S.-J., Magnani, A., Boyd, S., 2006a. Optimal kernel selection in kernel Fisher discriminant analysis. In: Paper Presented at the Proceedings of the 23rd International Conference on Machine Learning.
Kim, T.-W., Valdés, J.B., Yoo, C., 2003. Nonparametric approach for estimating return periods of droughts in arid regions. J. Hydrol. Eng. 8 (5), 237–246.
Kim, T.-W., Valdés, J.B., Yoo, C., 2006b. Nonparametric approach for bivariate drought characterization using Palmer drought index. J. Hydrol. Eng. 11 (2), 134–143.
Kirtman, B.P., Min, D., 2009. Multimodel ensemble ENSO prediction with CCSM and CFS. Mon. Weather Rev. 137 (9), 2908–2930.
Kocsis, T., Kovács-Székely, I., Anda, A., 2017. Comparison of parametric and non-parametric time-series analysis methods on a long-term meteorological data set. Central Europ. Geol. 60 (3), 316–332.
Lall, U., Rajagopalan, B., Tarboton, D.G., 1996. A nonparametric wet/dry spell model for resampling daily precipitation. Water Resour. Res. 32 (9), 2803–2823.
Li, H., Wang, D., Singh, V.P., Wang, Y., Wu, J., Wu, J., Liu, J., Zou, Y., He, R., Zhang, J., 2019. Non-stationary frequency analysis of annual extreme rainfall volume and intensity using Archimedean copulas: a case study in eastern China. J. Hydrol. 571, 114–131.
Madadgar, AghaKouchak, A., Shukla, S., Wood, A.W., Cheng, L., Hsu, K.L., Svoboda, M., 2016. A hybrid statistical-dynamical framework for meteorological drought prediction: application to the southwestern United States. Water Resour. Res. 52 (7), 5095–5110.
Madadgar, S., Moradkhani, H., 2012. Towards an Improved Postprocessing of Hydrological Forecast Ensembles Using Copula. EGUGA, p. 13757.
Madadgar, S., Moradkhani, H., 2013. Drought analysis under climate change using copula. J. Hydrol. Eng. 18 (7), 746–759.
Madadgar, S., Moradkhani, H., 2014. Improved B ayesian multimodeling: integration of copulas and B ayesian model averaging. Water Resour. Res. 50 (12), 9586–9603.
Madadgar, S., Moradkhani, H., Garen, D., 2014. Towards improved post-processing of hydrologic forecast ensembles. Hydrol. Process. 28 (1), 104–122.
Merryfield, W.J., Lee, W.-S., Boer, G.J., Kharin, V.V., Scinocca, J.F., Flato, G.M., Ajayamohan, R., Fyfe, J.C., Tang, Y., Polavarapu, S., 2013. The Canadian seasonal to interannual prediction system. Part I: models and initialization. Mon. Weather Rev. 141 (8), 2910–2945.
Nadarajah, S., 2005. Extremes of daily rainfall in west central Florida. Climatic Change 69 (2-3), 325–342.
Nagler, T., 2016. kdecopula: an R package for the kernel estimation of bivariate copula densities. arXiv preprint arXiv:1603.04229.
Peng, Y., Zhao, X., Wu, D., Tang, B., Xu, P., Du, X., Wang, H., 2018. Spatiotemporal variability in extreme precipitation in China from observations and projections. Water 10 (8), 1089.
Rayner, S., Lach, D., Ingram, H., 2005. Weather forecasts are for wimps: why water resource managers do not use climate forecasts. Climatic Change 69 (2-3), 197–227.
Reich, S., Cotter, C., 2015. Probabilistic Forecasting and Bayesian Data Assimilation. Cambridge University Press.
Rezayi, B.M., Jahanbakhsh, S., Bayati, K.M., Zeinali, B., 2011. Forecast of Autumn and winter Precipitation of West IRAN by Use from Summer and Autumn Mediterranean Sea Temperature.
Robertson, D., Shrestha, D., Wang, Q., 2013. Post processing rainfall forecasts from numerical weather prediction models for short term streamflow forecasting. Hydrol. Earth Syst. Sci. Discuss. 10 (5).
Roser, M., Ortiz-Ospina, E., 2016. Literacy. Our World in Data. Our World in Data.
Rossi, R., Murari, A., Gaudio, P., Gelfusa, M., 2020. Upgrading model selection criteria with goodness of fit tests for practical applications. Entropy 22 (4), 447.
Roy, T., He, X., Lin, P., Beck, H.E., Castro, C., Wood, E.F., 2020. Global evaluation of seasonal precipitation and temperature forecasts from NMME. J. Hydrometeorol. 1–41.
Rudemo, M., 1982. Empirical choice of histograms and kernel density estimators. Scand. J. Stat. 65–78.
Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., 2014. The NCEP climate forecast system version 2. J. Clim. 27 (6), 2185–2208.
Salvadori, G., De Michele, C., 2004a. Analytical calculation of storm volume statistics involving Pareto-like intensity-duration marginals. Geophys. Res. Lett. 31 (4).
Salvadori, G., De Michele, C., 2004b. Frequency analysis via copulas: theoretical aspects and applications to hydrological events. Water Resour. Res. 40 (12).
Salvadori, G., De Michele, C., 2010. Multivariate multiparameter extreme value models and return periods: a copula approach. Water Resour. Res. 46 (10).
Santos, L., Thirel, G., Perrin, C., 2018. Pitfalls in Using Log-Transformed Flows within the KGE Criterion.
Schefzik, R., 2013. Ensemble copula coupling as a multivariate discrete copula approach. arXiv preprint arXiv:1305.3445.
Schefzik, R., Thorarinsdottir, T.L., Gneiting, T., 2013. Uncertainty quantification in complex simulation models using ensemble copula coupling. Stat. Sci. 28 (4), 616–640.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Scott David, W., 1992. Multivariate density estimation: theory, practice, and visualization. In: Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.
Scott, D.W., Terrell, G.R., 1987. Biased and unbiased cross-validation in density estimation. J. Am. Stat. Assoc. 82 (400), 1131–1146.
Shashikanth, K., Sukumar, P., 2017. Indian monsoon rainfall projections for future using GCM model outputs under climate change. Adv. Comput. Sci. Technol. 10 (5), 1501–1516.
Sklar, M., 1959. Fonctions de repartition an dimensions et leurs marges. Publ. Inst. Statist. Univ., Paris, pp. 229–231, 8.
Slater, L.J., Villarini, G., Bradley, A.A., 2019. Evaluation of the skill of North-American multi-model ensemble (NMME) global climate models in predicting average and extreme precipitation and temperature over the continental USA. Clim. Dynam. 53 (12), 7381–7396.
Stute, W., 1992. Modified cross-validation in density estimation. J. Stat. Plann. Inference 30 (3), 293–305.
Tahroudi, M.N., Ramezani, Y., De Michele, C., Mirabbasi, R., 2020. Analyzing the conditional behavior of rainfall deficiency and groundwater level deficiency signatures by using copula functions. Nord. Hydrol.
Tao, Y., Duan, Q., Ye, A., Gong, W., Di, Z., Xiao, M., Hsu, K., 2014. An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai river basin. J. Hydrol. 519, 2890–2905.
Wu, L., Seo, D.-J., Demargne, J., Brown, J.D., Cong, S., Schaake, J., 2011. Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. J. Hydrol. 399 (3-4), 281–298.
Xiao, Y., 2017. A fast algorithm for two-dimensional Kolmogorov–Smirnov two sample tests. Comput. Stat. Data Anal. 105, 53–58.

Xin, P., Liu, Y., Yang, N., Song, X., Huang, Y., 2020. Probability distribution of wind power volatility based on the moving average method and improved nonparametric kernel density estimation. Global Energy Interconn. 3 (3), 247–258.

Xu, P., Wang, D., Singh, V.P., Wang, Y., Wu, J., Lu, H., Wang, L., Liu, J., Zhang, J., 2019. Time-varying copula and design life level-based nonstationary risk analysis of extreme rainfall events. Hydrol. Earth Syst. Sci. Discuss. 1–59.

Yazdandoost, F., Moradian, S., Zakipour, M., Izadi, A., Bavandpour, M., 2020. Improving the precipitation forecasts of the North-American multi model ensemble (NMME) over Sistan basin. J. Hydrol. 125263.

Zhang, L., Singh, V.P., 2007. Bivariate rainfall frequency distributions using Archimedean copulas. J. Hydrol. 332 (1-2), 93–109.

Zhao, D., Bu, L., Alippi, C., Wei, Q., 2017. A Kolmogorov-Smirnov test to detect changes in stationarity in big data. IFAC-PapersOnLine 50 (1), 14260–14265.