Quantifying the effect of climate variability on seasonal precipitation using Bayesian clustering approach in Kebir Rhumel Basin, Algeria

Lazhar Belkhiri[1] · Nir Krakauer[2]

Abstract

This paper presents a Bayesian clustering approach that allows quantification of the effect of climate variability on seasonal precipitation data in Kebir Rhumel Basin (KRB). We applied this approach to simultaneously identify clusters of stations with similar characteristics and the climate variability associated with each cluster and for the individual stations within each cluster. Both full pooling Bayesian clustering (FPBC) and partial pooling Bayesian clustering (PPBC) models with nonstationary generalized extreme value (GEV) distribution are applied to each season. In these models, a climate index variable, namely the El Niño Southern Oscillation (ENSO), is included as a time-varying covariate with an appropriate basis function to potentially explain the temporal variation of one or more of the parameters of the distribution. Results reveal that the partial pooling Bayesian clustering model provided the best fit for the seasonal precipitation data. The significant effect of ENSO differs from one season to another. During spring and autumn, ENSO significantly affects precipitation across large parts of KRB. Furthermore, the southern part and northern part of KRB are positively and negatively influenced by ENSO during winter and summer, respectively. Moreover, almost all stations during spring and autumn are negatively and positively influenced by ENSO, respectively. Finally, we demonstrated that the proposed model helps to reduce the uncertainty in the parameter estimation and provides more robust results.

Keywords Seasonal precipitation · Climate variability · Bayesian clustering approach · Kebir Rhumel Basin

1 Introduction

Climate change and variability is one of the most serious challenges facing the global ecosystem in the 21st Century. Global warming and climate fluctuation have comprehensively affected the water cycle components, especially precipitation (Berghuijs et al. 2017; Konapala et al. 2020; Belkhiri and Kim 2021; IPCC, 2021). The El Niño Southern Oscillation (ENSO) is the single most influential climate phenomenon affecting the variability of precipitation (Ropelewski and Halpert 1987; Hoerling et al. 1997;

Dai et al. 1997; Ward et al. 2014). Although ENSO is defined based on Pacific surface pressure or sea surface temperature, it affects climate in various ways globally, particularly across the tropics and subtropics. Many studies have shown the complex effects of ENSO on the precipitation in many countries throughout Africa (Ropelewski and Halpert 1987; Kiladis and Diaz 1989; Nicholson and Kim 1997; Mason and Goddard 2001; Lüdecke et al 2021). Fewer studies investigated the effect of ENSO on precipitation in Algeria. For example, Meddi et al. (2010) analyzed the temporal variability of annual precipitation in the Macta and Tafna catchments of northwestern Algeria and showed a negative ENSO correlation with precipitation. Turki et al. (2016) studied the long-term variability of rainfall in the Soummam watershed and noted significant effect of ENSO on rainfall variability in northeastern Algeria. Zeroual et al. (2016) analyzed the climate indices' influence on temperature and annual as well as seasonal rainfall in the coastal region of northern Algeria from 1972 to 2013, the researchers found a positive correlation

✉ Lazhar Belkhiri
belkhiri.la@gmail.com; belkhiri_laz@yahoo.fr

Nir Krakauer
nkrakauer@ccny.cuny.edu

[1] Laboratory of Applied Research in Hydraulics, University of Mustapha Ben Boulaid, Batna 2, Algeria

[2] Department of Civil Engineering, City College of New York, New York, NY 10031, USA

between rainfall and the ENSO (El Niño-Southern Oscillation) index. However, the effect of climate indices such as ENSO on the full probability distribution of precipitation in Algeria has not yet been comprehensively assessed. Therefore, there is room for new approaches to evaluate the response of precipitation to ENSO.

The comprehensive effects of climate variability/change on precipitation have been studied by using several non-stationary models. For instance, Micevski et al. (2006) examined the effect of Inter-decadal Pacific Oscillation (IPO) for mainly coastal, eastern Australian flood data and found that the IPO modulates the flood risk within parts of eastern Australia. Ouarda and El-Adlouni (2011) discussed nonstationary frequency analysis models in hydrology with a focus on the Bayesian approach and demonstrated that the Bayesian approach can be applied to more general and more complex models where parameters are expressed as nonlinear functions of covariates. Madsen et al. (2014) presented a review of trend analysis of extreme precipitation and floods and described that non-stationarity in extreme precipitation and flood characteristics due to climatic changes is high on the research agenda in Europe. Chen et al. (2014) developed a Hierarchical Bayesian approach for regional rainfall and streamflow forecast using appropriate climate indicators and described that this approach allows appropriate grouping of information in the region and explicit modeling of the covariance of the model errors and the regression coefficients to better represent the uncertainty in the model parameters and the final streamflow and rainfall forecasts. Sun et al. (2014) demonstrated that a Bayesian regional framework provides the opportunity to assess the value of regional information in better identifying the effect of climate variability on hydrometeorological extremes. Despite the results of these valuable studies, the reduction of the uncertainties and the identification of the effect of climate variability in these models across a relatively large area by regional pooling of information on model parameters raises questions. Some studies addressed this question but considered a relatively small homogeneous area over which information can be pooled (Aryal et al. 2009; Renard et al. 2008; Renard et al. 2013; Chen et al. 2014; Sun et al. 2014; Sun et al. 2015a). Moreover, some authors have applied Bayesian clustering models based on Dirichlet process or Expectation–Maximization (EM) algorithm (Xiong and Yeung 2004; Johnson et al. 2013; Nieto-Barajas and Contreras-Cristán, 2014) but these mixture models have been limited to a Gaussian mixture model. Sun and Lall (2015) and Sun et al. (2015b) developed a new Bayesian clustering approach for exploring homogeneity of response in large area datasets, through a multicomponent mixture model (or clusters). This approach allows the reduction of uncertainties through full pooling or partial pooling across automatically chosen

subsets of the data. In this study, we applied a similar Bayesian clustering approach with non-stationary GEV distribution for seasonal precipitation using ENSO index as covariate.

The current paper is structured as follows. Section 2 introduces the study area and data. Section 3 presents the Bayesian clustering approach. Section 4 describes the results and discussions, followed by Sect. 5 in which we concluded the main findings.

2 Study area and data

The Kebir Rhumel Basin (KRB) is located in northern Algeria and includes seven sub-basins, which covers an area of approximately 8815 km^2 (Table 1 and Fig. 1). The KRB is drained by two major rivers, Oued Rhumel in the southern part and Oued Endja in the western part. Beni-Haroun and Boussiaba are considered as the major dams in the basin, with a capacity of 960 hm^3 and 120 hm^3, respectively (Marouf et al. 2019). According to Mebarki (2005), of KRB climate is humid in northern part and semi-arid in the southern part.

The precipitation datasets for the period 1970–2013 used in this study were obtained from the National Agency for Hydraulic Resources and Office National Office Meteorology. In the current study, the mean value of monthly precipitation is derived for the four seasons winter (DJF), spring (MAM), summer (JJA) and autumn (SON). For each season, a station was selected only if it had at least 90% of the total precipitation data available during the period 1970–2013. Overall, 24 stations are utilized for the analysis. Figure 1 displays the spatial distribution of the selected stations in KRB.

In order to analyze the effect of climate variability on the seasonal precipitation, the El Niño Southern Oscillation (ENSO) was considered. Many indices have been developed to characterize aspects of ENSO evolution (Trenberth 1997; Trenberth and Stepaniak 2001). The seasonal mean of Southern Oscillation Index (SOI) is used as the measure of ENSO, and was obtained from the National Oceanic and Atmospheric Administration Climate Prediction Center (https://www.cpc.ncep.noaa.gov/data/indices/soi/).

3 Methodology

3.1 Nonstationary model structure

For the nonstationary modeling of the seasonal precipitation, it is very important to choose an appropriate distribution. In recent decades, many studies have successfully used the generalized extreme value (GEV) distribution to

Table 1 The surface area and number of stations for each Sub-Basins in KRB

Name	Code	Surface area (km^2)	Number of stations
Oued Dehamecha	1001	1061	1
Oued Kebir-Endja	1002	1098	3
Oued Rhumel Amont	1003	1320	4
Oued Rhumel Seguin	1004	1220	4
Oued Boumerzoug	1005	1162	4
Oued Rhumel Smendou	1006	1865	2
Oued Kebir Maritime	1007	1089	6

Fig. 1 Locations of the study area and the selected stations
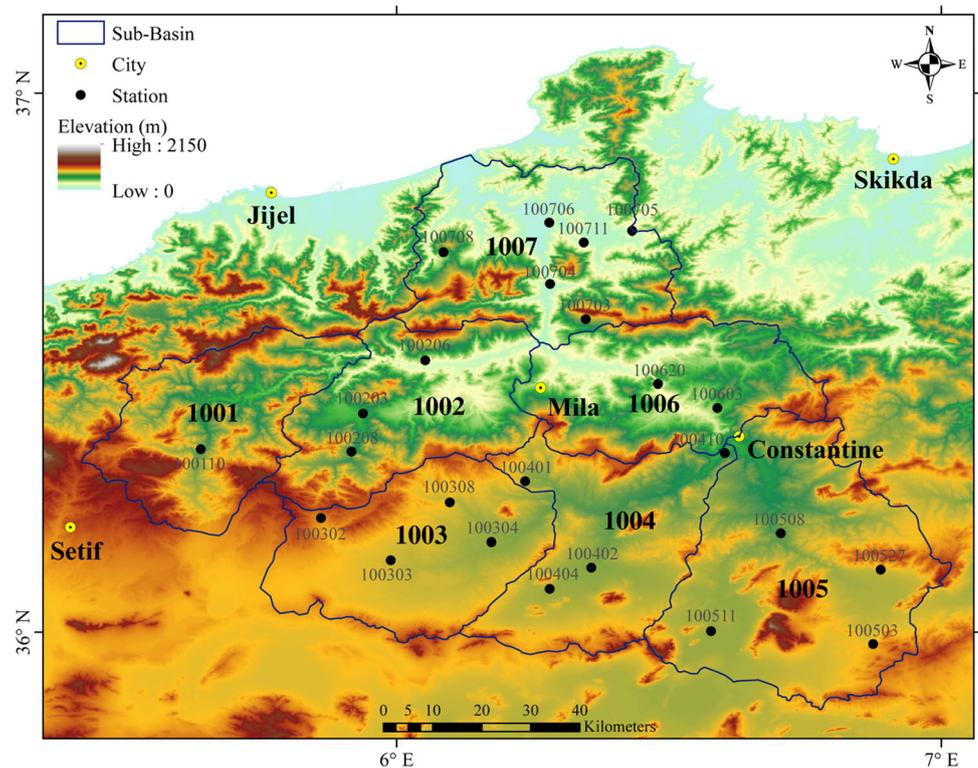
model the nonstationarity in precipitation events by including time-varying parameters (Du et al. 2014; Cheng and Aghakouchak 2014; Gao et al. 2016; Agilan and Umamahesh, 2017a, b; Su and Chen 2019). Steirou et al. (2019) identified links between seasonal flood probabilities and large-scale atmospheric indices for entire Europe by adopting a Bayesian framework with climate-informed (non-stationary) GEV distribution and compare it with the classical (stationary) GEV distribution with parameters invariant in time. They demonstrated that the climate-informed models were preferred over the classical GEV distribution for a high percentage of stations for most seasons and the seasonally averaged indices provided in most cases better fits compared with monthly values. Ossandón et al. (2022) developed a Bayesian Hierarchical Model (BHM) to project seasonal streamflow extremes for several lead times based on a Gaussian elliptical copula and

Generalized Extreme Value (GEV) margins with nonstationary parameters and demonstrated that the framework proposed could be useful for the early implementation of flood risk adaptation and preparedness strategies. Thus, a nonstationary generalized extreme value (GEV) distribution is considered to model the seasonal precipitation over the KRB at each station. Here, we assumed that the GEV location parameter was linked to the temporal climate covariate (ENSO index) using a linear regression model. In the preliminary analysis, we considered the effect of the climate index on both the location and scale parameter, but this did not provide very different results than those for a covariate on the location parameter only (not shown). The shape parameter is kept constant as its estimation includes large uncertainties, even under the assumption of stationarity (Coles et al. 2001; Papalexiou and Koutsoyiannis 2013; Silva et al. 2017; Steirou et al. 2017).

In this application, the three proposed model structures of no pooling, full pooling and partial pooling can be written as follows:

(a) No pooling model

$$Y(s,t) \sim GEV(\mu_0(s) + \mu_1(s) * ENSO(t), \sigma(s), \xi(s)) \tag{1}$$

(b) Full pooling model

$$Y(s,t) \sim GEV(\mu_0 + \mu_1(s) * ENSO(t), \sigma(s), \xi) \tag{2}$$

(c) Partial pooling model

$$Level1 : Y(s,t) \sim GEV(\mu_0 + \mu_1(s) * ENSO(t), \sigma(s), \xi)Level2 : \mu_1(s) \sim N(\mu_\mu, \sigma_\mu) \tag{3}$$

$Y(s,t)$ is the observation of the variable at station s and time t. $ENSO(t)$ is climate covariate at the time t. $\mu_0(s)/\mu_0$, $\mu_1(s)$, $\sigma(s)$ and $\xi(s)/\xi$ are model parameters, where the $\mu_0(s)/\mu_0$ is the intercept of the location parameter, and the $\mu_1(s)$ is the slope of the location parameter at station s. $\sigma(s)$ and $\xi(s)/\xi$ are the scale and shape parameters, respectively. $\mu_0(s)$, $\mu_1(s)$, $\sigma(s)$ and $\xi(s)$ are considered as a site-specific (local) parameter while μ_0 and ξ are a regional parameters. $N(.,.)$ denotes a normal distribution. μ_μ and σ_μ are the hyper-parameters in the second level model (Eq. 3).

In the current research, we use the same linear regression function on the location parameter in the proposed models to describe the ENSO temporal climate covariate, while these models differ in the settings of μ_0, μ_1 and ξ, which are either site-specific (no pooling), regional (full pooling), or have a second level (partial pooling). In the no pooling model (Eq. 1), the three GEV parameters are estimated locally (i.e., were site-specific). This model was used as a baseline. In the full pooling model (Eq. 2), all model parameters are allowed to be estimated independently for each station except that the intercept of the location parameter (μ_0) and the shape parameter (ξ) are regional and estimated by using all data. We considered these two parameters for clustering in order to standardize the seasonal precipitation data, and considering that the shape parameter requires more data to obtain a precise estimation due to large uncertainty (Coles, 2001, p.106). In the partial pooling model (Eq. 3), we allowed for pooling of information across stations for estimating the regression slope of the location parameter to reduce the associated uncertainty, but also for variability in this parameter between stations. In this model, the slope parameter is presumed to drawn from common hyper-distribution,

therefore the $\mu_1(s)$ is in turn described by a set of hyper-parameters μ_μ and σ_μ (i.e., second level model (Eq. 3)). Here, the hyper-distribution describes the second level of the hierarchical Bayesian model. The intercept μ_0 and the shape parameter ξ are still regional in this model.

3.2 Hierarchical Bayesian clustering model

The hierarchical Bayesian clustering model with a non-stationary GEV distribution is applied to model the seasonal precipitation. Sun and Lall (2015) and Sun et al. (2015b) demonstrated that this model can be applied directly to a heterogeneous area. In this mixture model, we assume that the stations can be classified into K clusters, and a hierarchical Bayesian model (H_k) for each cluster (k) is developed. Here, each station has a probability π_k to belong to a cluster k, which needs to be estimated.

The mixture distribution across all clusters can be given as follows:

$$Y(s) \sim \sum_{k=1}^{K} \pi_k f_{H_k}(s) \tag{4}$$

where $f_{H_k}(s)$ is the likelihood function of the hierarchical Bayesian model (H_k) at station s. $\pi = \{\pi_1, ..., \pi_k\}$ denotes the mixing probabilities (i.e., mixing coefficients or weights). In order to be valid probabilities, the probability π_k must satisfy:

$$0 \le \pi_k \le 1(k = 1,...,K), and \sum_{k=1}^{K} \pi_k = 1 \tag{5}$$

Therefore, the likelihood of the hierarchical Bayesian clustering model can be computed as follows:

$$L = \prod_{s=1}^{S} \sum_{k=1}^{K} \pi_k f_{H_k}(s) \tag{6}$$

where L is the likelihood function and S in the total number of stations. A schematic of hierarchical Bayesian clustering model is presented in Fig. 2.

In this study, full pooling Bayesian clustering (FPBC) and partial pooling Bayesian clustering (PPBC) models are considered for each cluster k. A summary of the two proposed models used in this research is shown in Table 2.

Based on the partial pooling model described in Eq. (3), we constructed this model for each cluster by setting the intercept of the location parameter (μ_0) and the shape parameter (ξ) to be fully pooled (regional) and the slope of the location parameter to be partially pooled. The hierarchical Bayesian model (H_k) for each cluster k can be written as follows:
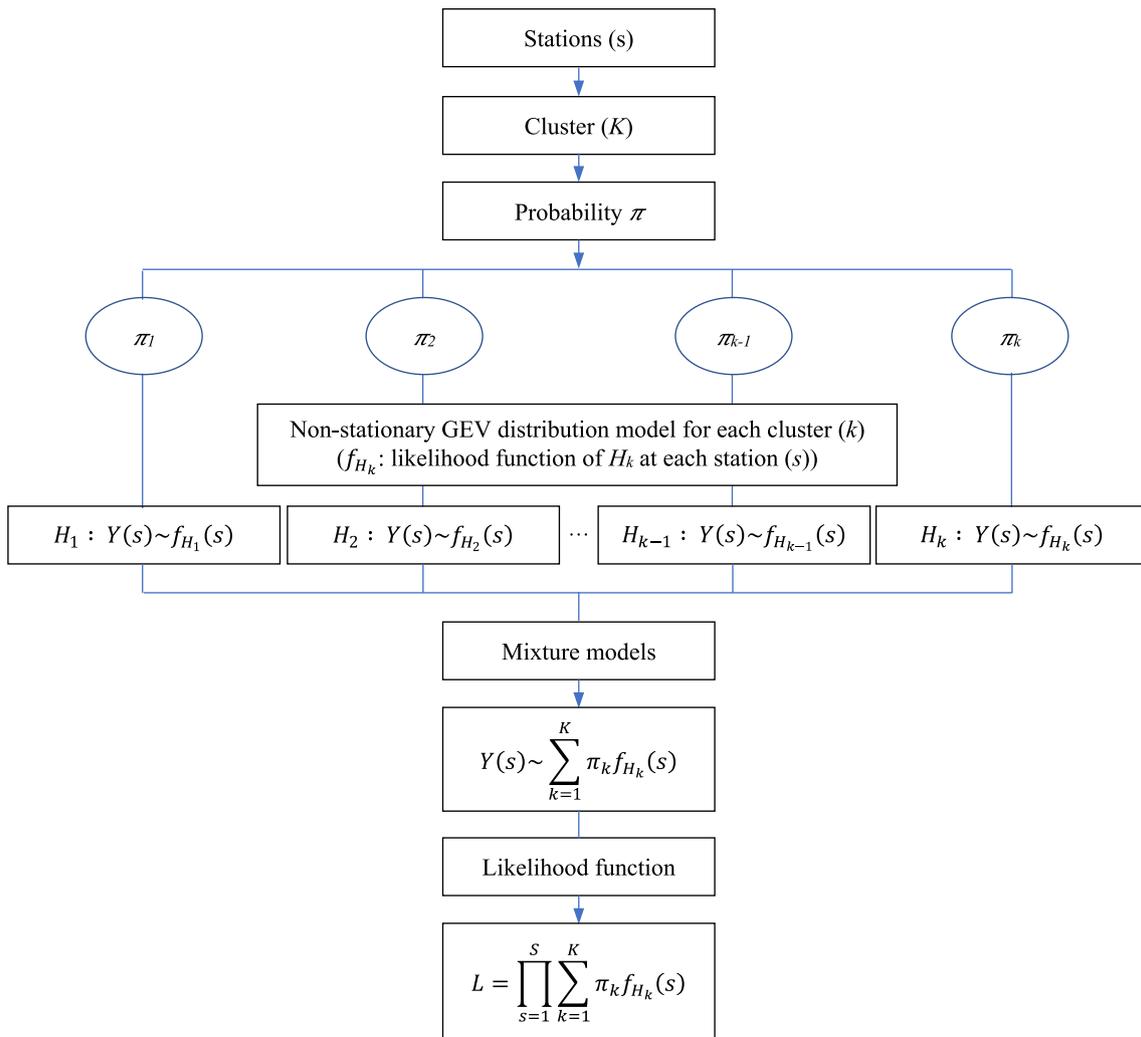
Fig. 2 Schematic diagram of hierarchical Bayesian clustering model

Table 2 Nonstationary GEV distribution models to fit the seasonal precipitation	Models	Parameters for clustering
	Full pooling Bayesian clustering model (FPBC) $Y(s,t) \sim GEV\big(\mu_{0,k} + \mu_1(s) * ENSO(t), \sigma(s), \xi_k\big)$	$\mu_{0,k}, \xi_k$
	Partial pooling Bayesian clustering model (PPBC) $Level1$: $Y(s,t) \sim GEV\big(\mu_{0,k} + \mu_1(s) * ENSO(t), \sigma(s), \xi_k\big)$ $Level2$: $\mu_1(s) \sim N\big(\mu_{\mu_k}, \sigma_{\mu_k}\big)$	$\mu_{0,k}, \xi_k, \mu_{\mu_k}, \sigma_{\mu_k}$

$$Level1 : Y(s,t) \sim GEV\big(\mu_{0,k} + \mu_1(s) * ENSO(t),$$
$$\sigma(s), \xi_k\big) Level2 : \mu_1(s) \sim N\big(\mu_{\mu_k}, \sigma_{\mu_k}\big) \tag{7}$$

where $\mu_{0,k}$, ξ_k, μ_{μ_k} and σ_{μ_k} are parameters that are associated with cluster k.

The likelihood function of H_k at station s can be calculated as follows:

$$f_{H_k}(s) = \prod_{t=1}^{T} f_{GEV}\big(Y(s,t)|\mu_{0,k}, \mu_1(s), \sigma(s), \xi_k, ENSO(t)\big)$$
$$X f_N\big(\mu_1(s)|\mu_{\mu_k}, \sigma_{\mu_k}\big) \tag{8}$$

The full likelihood function of the hierarchical Bayesian clustering model is obtained when integrating Eq. (8) into Eq. (4).

In this application, we used a Dirichlet distribution with identical parameters (a vector of 2 with length k) as a prior for π_k. For the other parameters, we used flat priors (normal or uniform distribution with large variance). For the initial values, we used the values that were obtained by fitting a nonstationary GEV distribution using maximum likelihood as the starting point for the intercept of the location parameter, scale and shape of each station. For the hyperparameters, the mean of slope parameter $\mu_1(s)$ is set to zero initially.

3.3 Implementation and model fitting

For each Bayesian clustering model, the posterior probability distribution of the model parameters is estimated using a No-U-Turn Hamiltonian Monte Carlo method (Hoffman and Gelman 2014). One chain of length 30,000 was run, with the first 15,000 iterations discarded as warmup. The convergence is evaluated by the potential scale reduction factor (Gelman and Rubin 1992), which should be smaller than 1.2 for each parameter. All the calculations are conducted using R and RStan (Stan Development Team 2022).

3.4 Selecting the optimal number of clusters

The selection of the optimal number of clusters is an important issue in mixture modeling. In a common trade-off in model selection problems, the mixture model with too many clusters may overfit the data, while a mixture with too few clusters may not be flexible enough to approximate the underlying model. Thus, it is important to adopt some statistical criteria to infer an optimal number of clusters (Deng and Han 2018). There are some statistical criteria, such as Akaike Information Criterion (AIC) (Akaike 1974), Bayesian Information Criterion (BIC) (Schwarz 1978), and Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002), can be used to select the models among different clusters. In the Bayesian framework, AIC and BIC can be applied to the integrated likelihood over the model parameters. However, AIC and BIC are not theoretically justified for mixture models and may not be the best way to determining the optimal number of clusters (Biernacki and Govaert 1997; McLachlan and Peel 2004). Alternately, the DIC is a model selection criterion that automatically considers parameter uncertainty by utilizing the posterior distribution. However, the definition and application of DIC to mixture models are not straightforward, and different definitions and adaptations have been proposed (Delorio and Roberst, 2002). To overcome the limitation of BIC, Biernacki et al. (2000) proposed an Integrated Completed Likelihood (ICL) criterion, which showed that it performs well both for selecting a mixture model and an optimal number of clusters. Also, ICL differs from other criteria in that the integrated likelihood of the complete data (observed data) are used to evaluate mixture models. The ICL criterion is defined by:

$$ICL\left(K, \widehat{\theta}\right) = -2\mathrm{log}f\left(Y, \widehat{Z}|K, \widehat{\theta}\right) + \upsilon_K log(S) \qquad (9)$$

where Y is the observed data. $\mathrm{log}f\left(Y, \widehat{Z}|K, \widehat{\theta}\right)$ is the log likelihood. $\widehat{Z}$ is a $m*K$ binary matrix, which refers to the estimated membership of each station, and $\widehat{Z}_{s,k} = 1$ if and only if station s belongs to the cluster k. $\widehat{\theta}$ is the collection of estimated parameters, υ_K is the number of parameters, and S is the number of stations. Among the different mixture models, the one with the lowest ICL value is preferred. In this study, we computed the ICL to select the optimal number of clusters for each Bayesian clustering model.

4 Results

4.1 Preliminary investigations

For the preliminary analysis, the no pooling model with a nonstationary GEV distribution was constructed using linear function of the ENSO-index time-covariate to fit the seasonal precipitation across the Kebir Rhumel Basin. The slope parameter (μ_1) characterizes the effect of the climate index (ENSO) on the location parameter. If the effect of the climate index is significant, the posterior distribution of μ_1 not have zero near its center.

The posterior probability distribution of the slope of location parameter (μ_1) estimated from the no-pooling model for all stations and four seasons as boxplots is illustrated in Fig. 3. The red and blue colors of the box represent the significant and non-significant effect of the climate index, respectively. As illustrated in Fig. 3, we can see that the most stations have a positive value of the median posterior distribution of the slope parameter during the winter and autumn seasons, However, most stations of them have a negative value of the median posterior distribution of μ_1 during spring and summer seasons, suggesting that in this basin, ENSO may have contrasting effects on precipitation in different seasons. In addition, the results show that 58% (14 out of 24) and 33% (8 out of 24) of the total stations have a significant effect of the climate index during the spring and autumn, respectively. By contrast, the significant effect of climate index during the
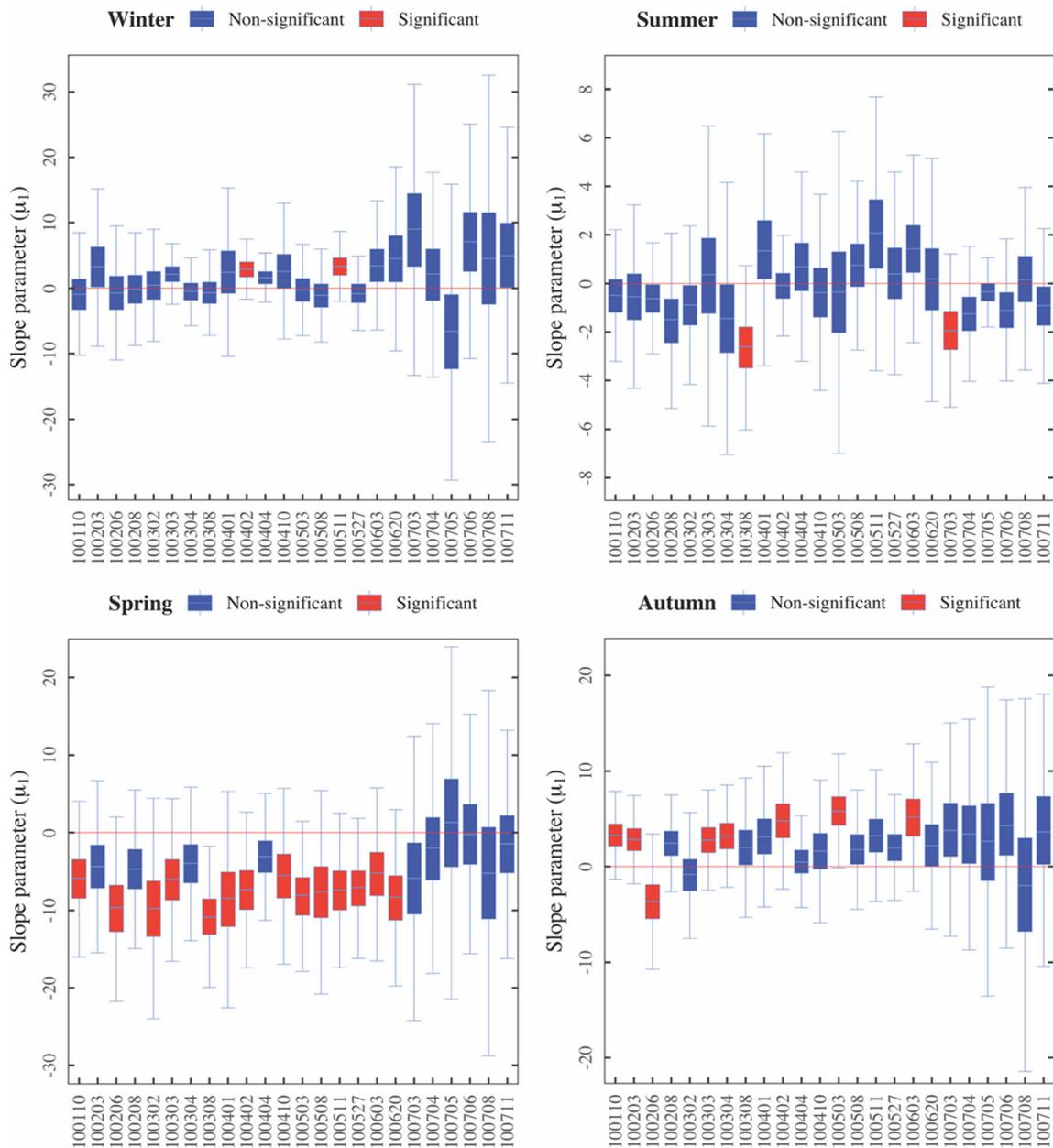
Fig. 3 Boxplot of the posterior distribution of the slope parameter $\mu_1(s)$ from the no pooling model for all station. Red boxes represent the stations with significant effect, while the blue boxes are for the stations with non-significant effect of the climate index

rest of the seasons is only detected in two stations. Next, we applied the full pooling Bayesian clustering and partial pooling Bayesian clustering models in order to understand more about the influence of the climate index on the seasonal precipitation and its spatial variability within the basin.

4.2 Bayesian clustering models

In the current study, we applied the full pooling Bayesian clustering (FPBC) and partial pooling Bayesian clustering

(PPBC) models using a nonstationary GEV distribution to fit the seasonal precipitation data by varying the number of clusters from 2 to 4.

4.2.1 Model selection

Model selection for the full pooling Bayesian clustering and partial pooling Bayesian clustering models includes the selection of the optimal number of clusters. Choosing the number of clusters (K) is a very important issue when using mixture models. In this study, the ICL is considered, and
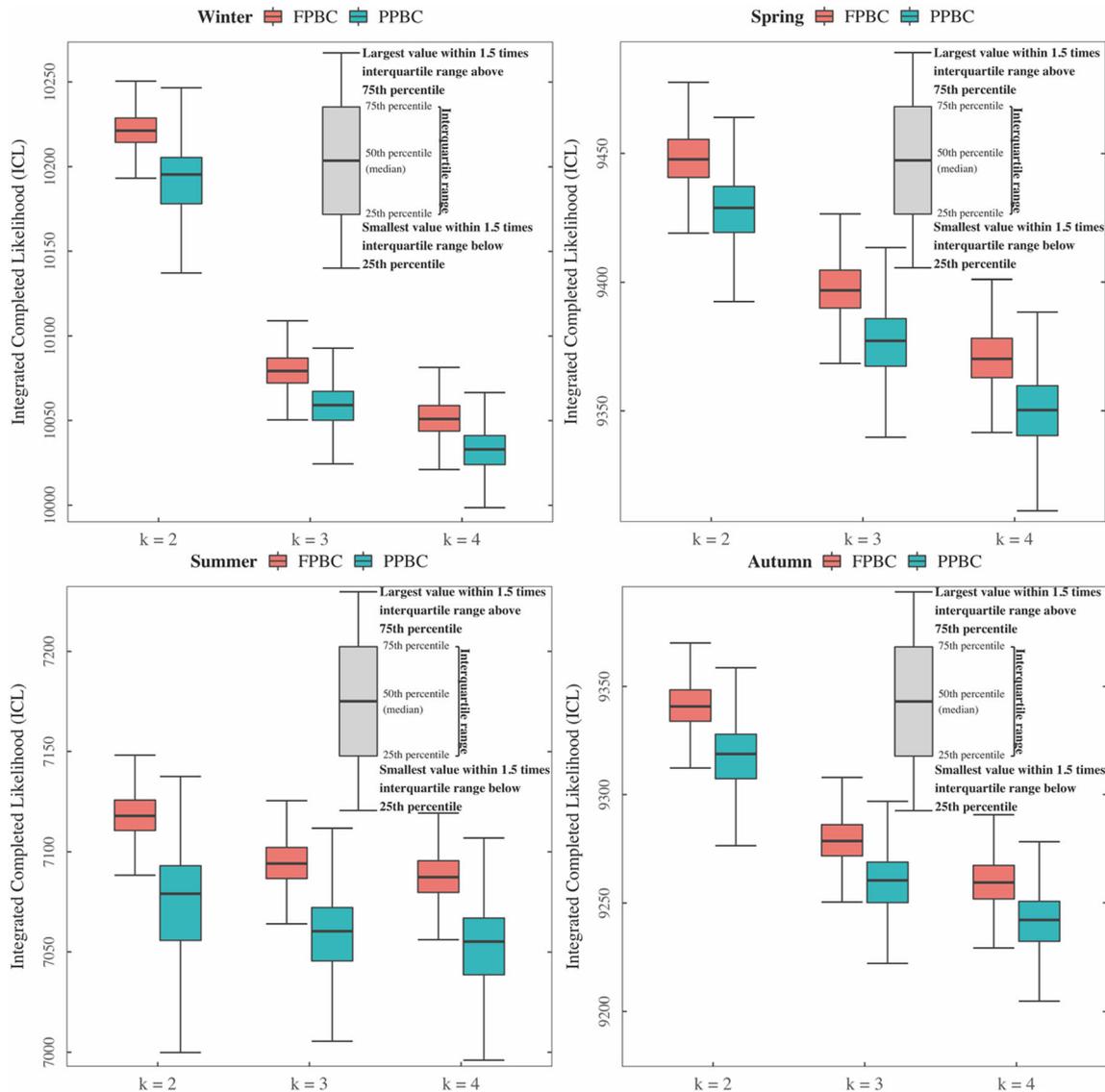
Fig. 4 Boxplots of the Integrated Completed Likelihood (ICL) of FPBC and PPBC models for each season

the best model is selected based on the lowest ICL value. Figure 4 shows the boxplots of the ICL values of FPBC and PPBC models. In all four seasons, the ICL values decrease from $K = 2$ to $K = 4$ for both mixture models, indicating that adding the number of clusters improves the fit of the mixture models. Thus, the optimal number of clusters is equal to 4. In other words, the mixture models with $K = 4$ provide better fit than a mixture models with $K = 3$ and 2. As shown in Fig. 4, the partial pooling Bayesian clustering model with K equal to 4 has the lowest values of ICL among all K values and thus provide the best fit for the seasonal precipitation data in the Kebir Rhumel Basin.

4.2.2 Assignment of stations to clusters

In order to determine the membership of each station in an appropriate cluster, we calculate the posterior probability (P_{post}) of each station (s) belonging to cluster k as follows:

$$P_{post}(s \in cluster k) = \frac{\pi_k f_{H_k}(s)}{\sum_{k=1}^{K} \pi_k f_{H_k}(s)} \quad (10)$$

A station is a member of the cluster k where its posterior probability is highest.

Figure 5 displays the membership of station s belonging to cluster k for the partial pooling Bayesian clustering model with the number of clusters equal to 4. The summary statistics of the posterior distribution of the model parameters for each cluster in PPBC model is presented in Table 3. From the results, we can see clearly that the PPBC

Fig. 5 Spatial distribution of the identified clusters by the PPBC model with $K = 4$ for each season. The up-pointing and down-pointing triangles denote the significant positive and negative effect of ENSO, respectively. Colored dots and triangles denote the identified clusters by the clustering model
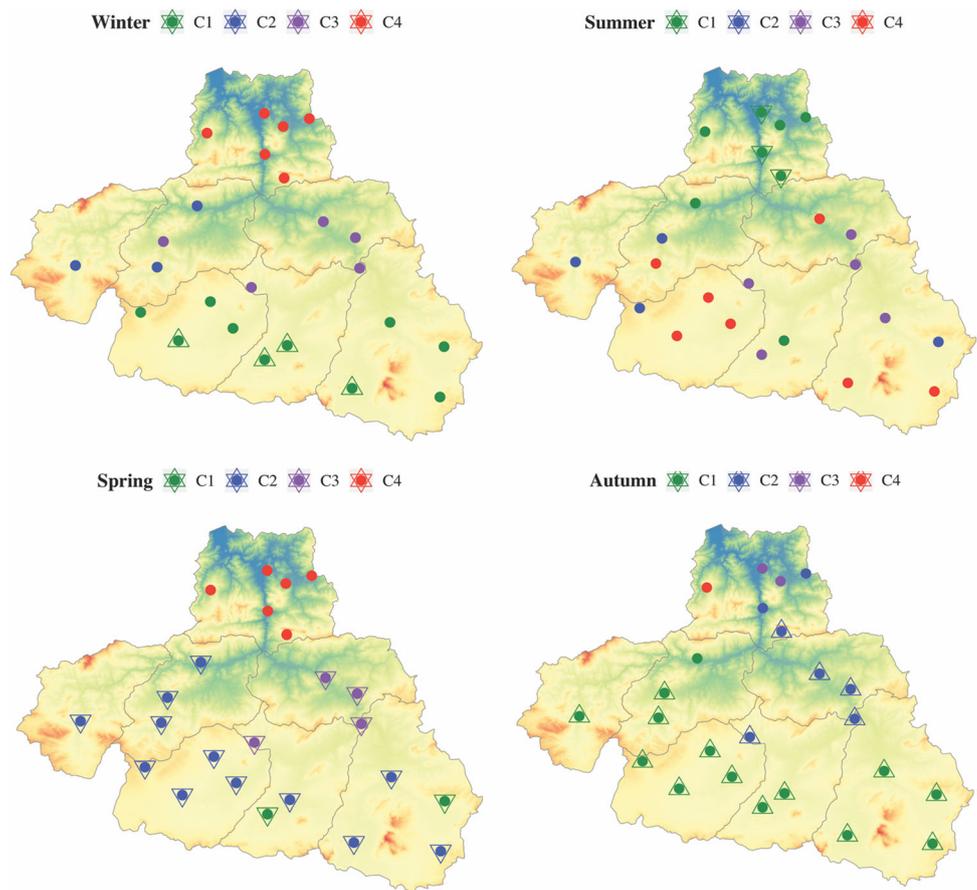
model identify four clusters during the four seasons. It can be seen in Table 3 that, about of 42%, 33% and 58% of the total stations belong to the first cluster in the season of winter, summer and autumn, respectively. Whereas, in the season of spring about of 50% of the total stations belong to the second cluster.

The estimated seasonal precipitation by the PPBC model with non-stationary GEV distribution for each cluster during the four seasons is displayed in Fig. 6. For the purpose of comparison, the empirical cumulative function of observed and fitted PPBC model with non-stationary GEV distribution to the seasonal precipitation is illustrated in Fig. 7 for each season. The figures show how the estimated theoretical distributions are adapted to the empirical distributions in each season.

As see in Figs. 6 and 7, we can observe that the values of the seasonal precipitation are changed from one season to another and from the first cluster to last one. Also, the cumulative distribution function (CDF) plot showed that PPBC with non-stationary GEV distribution is practically following the observed seasonal precipitation in all stations. Therefore, the results confirm that the PPBC model with $k = 4$ is the best model for seasonal precipitation which clearly explained the significant variation of the

precipitation between the seasons and clusters. The winter season consisted of the stations with the highest values of the seasonal precipitation while summer season had the smallest values. Moreover, the seasonal precipitation values are increased from the first cluster to the last cluster during all seasons.

In all four seasons, the values of the mean posterior distribution of the intercept μ_0 and scale σ parameters are increased from the first to the last cluster, indicating that the last cluster has the larger mean compared with the other clusters. In addition, we can clearly observe that the mean values of the posterior distribution of μ_0 and σ are increased form the south to the north of the Kebir Rhumel Basin in the seasons of winter, spring and autumn. Moreover, the vertical sorting of the clusters is related to the elevation and distance from the Mediterranean Sea. During the winter and spring, the stations belong to the last cluster are covered the sub-basin of Oued Kebir Maritime and situated near to the Mediterranean Sea. Thus, this indicated that the high values of the precipitation are observed in the sub-basin of Oued Kebir Maritime (1007). The mean posterior distribution of the shape parameter (ξ) in the most stations is negative during the seasons of winter, spring and autumn and positive during the summer season.

Table 3 Summary of the clustering results for PPBC model with $K = 4$

Clusters	%NBS	μ_0	$\overline{\mu_1}$	$\tilde{\sigma}$	ξ
Winter					
C1	41.67	27.00	0.91	14.73	− 0.160
C2	12.50	41.44	− 0.11	21.76	− 0.052
C3	20.83	57.47	2.63	27.94	0.016
C4	25.00	112.41	3.00	56.64	0.053
Spring					
C1	8.33	22.77	− 4.76	13.73	− 0.119
C2	50.00	33.10	− 6.91	15.91	− 0.109
C3	16.67	44.91	− 6.61	18.33	− 0.059
C4	25.00	64.79	− 2.15	30.45	− 0.046
Summer					
C1	33.33	6.20	− 0.59	4.21	0.096
C2	16.67	6.56	− 0.19	5.19	0.120
C3	20.83	6.81	0.00	5.78	0.143
C4	29.17	8.44	− 0.50	6.96	0.200
Autumn					
C1	58.33	23.57	2.40	12.91	− 0.030
C2	29.17	37.86	2.95	22.73	− 0.011
C3	8.33	60.89	3.88	29.99	0.020
C4	4.17	92.98	− 0.99	41.91	0.026

%NBS: Percentage of the number of stations

$\overline{\mu_1}$: The mean of the median posterior of μ_1.

$\tilde{\sigma}$: The mean of the median posterior of σ
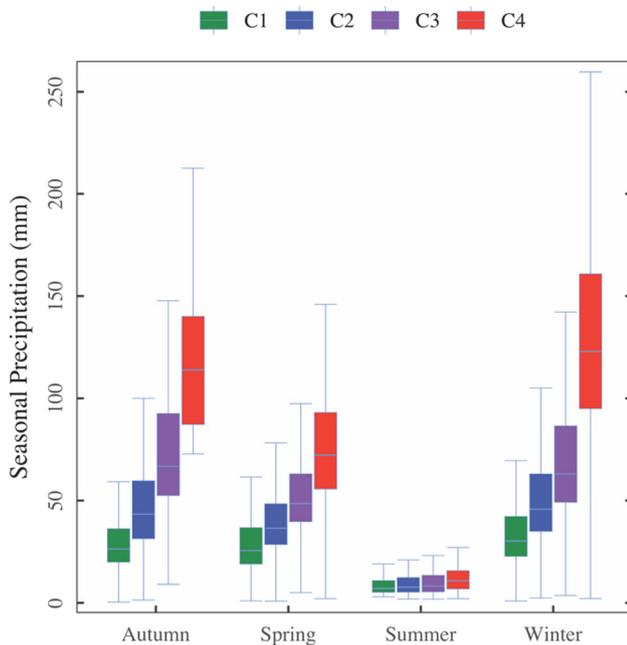
Fig. 6 Boxplots of the clusters obtained by PPBC model with k = 4 for each season

4.2.3 Identify the effect of the climate variability

In the current study, the effect of the El Niño Southern Oscillation (ENSO) index on the seasonal precipitation is characterized by slope of the location parameter (μ_1). For each season, the posterior distribution of the slope parameter μ_1 estimated by the partial pooling Bayesian clustering model for each station is displayed in Fig. 8. The colors of the boxes represent the identified clusters by clustering model with K equals to 4. The median posterior distribution of μ_1 for each cluster is presented in Table 3. From the results, we can see that most stations during the winter and autumn seasons have a positive value of the median posterior of the slope μ_1, while most stations in spring and summer seasons have a negative value of μ_1. This again indicates that ENSO (SOI) has a positive influence on the seasonal precipitation during winter and autumn seasons whereas a negative influence during spring and summer seasons.

In addition, to better understand the significant effect of ENSO on the seasonal precipitation, we examine the posterior distribution of the slope μ_1 of the location parameter for each season. For a significant effect of ENSO, the zero value is not included in the 90% posterior interval of the slope parameter μ_1. A significant positive (negative) effect is considered if the probability of the posterior distribution pdf of μ is larger (smaller) than zero at the 10% (90%) significance level. Table 4 presents the percentage of stations with a significant positive and negative effect of ENSO by PPBC model. Figure 5 shows the spatial distribution of the significant effect of ENSO for each season. From the results, we can observe that a significant effect of ENSO is found in precipitation at 17% (4 stations), 75% (18 stations), 12% (3 stations) and 75% (18 stations) of the total stations during winter, spring, summer and autumn seasons, respectively, indicating that the highest numbers of significant effect of ENSO are found in spring and autumn. Also, the significant positive and negative effects of ENSO on the seasonal precipitation differ from one season to another. In winter and summer, a significant effect of ENSO is only detected in the stations of the first cluster. Thus, a significant positive influence of ENSO during winter is observed in the southern part of the Kebir Rhumel Basin, but a significant negative influence during summer is detected in the northern part of the Kebir Rhumel Basin (sub-basin 1007). In spring and autumn, the stations of the first three and two clusters are negatively and positively influenced by ENSO, respectively. In addition, all stations of the sub-basins except those of Oued Kebir Maritime sub-basin are negatively and positively influenced by ENSO during spring and autumn, respectively. These results were roughly consistent with the findings in previous studies in Africa (Ropelewski and
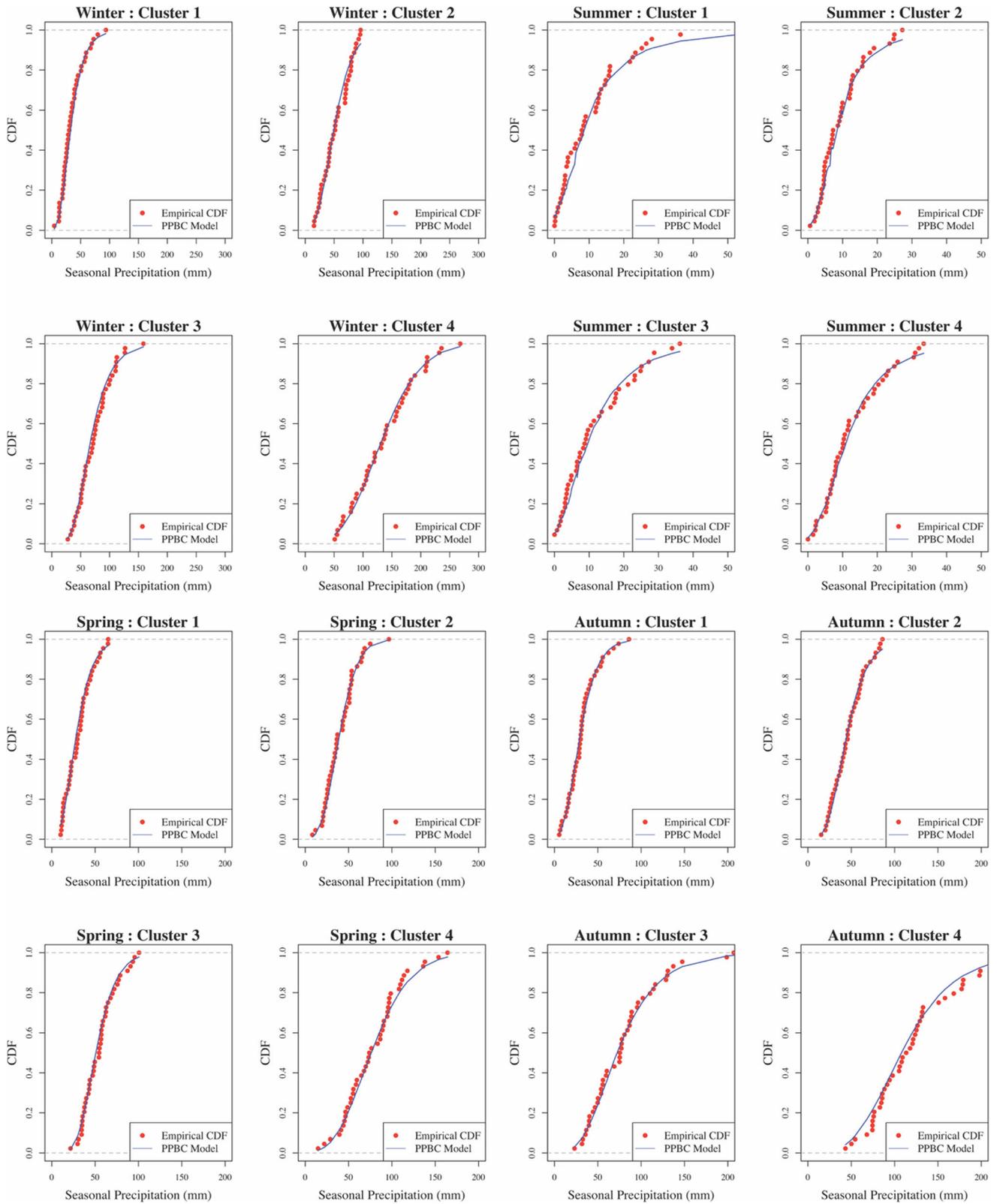
Fig. 7 Empirical distribution function for the seasonal precipitation of observed data (dots) and PPBC models for each cluster (dashed lines). The stations for each cluster are randomly selected
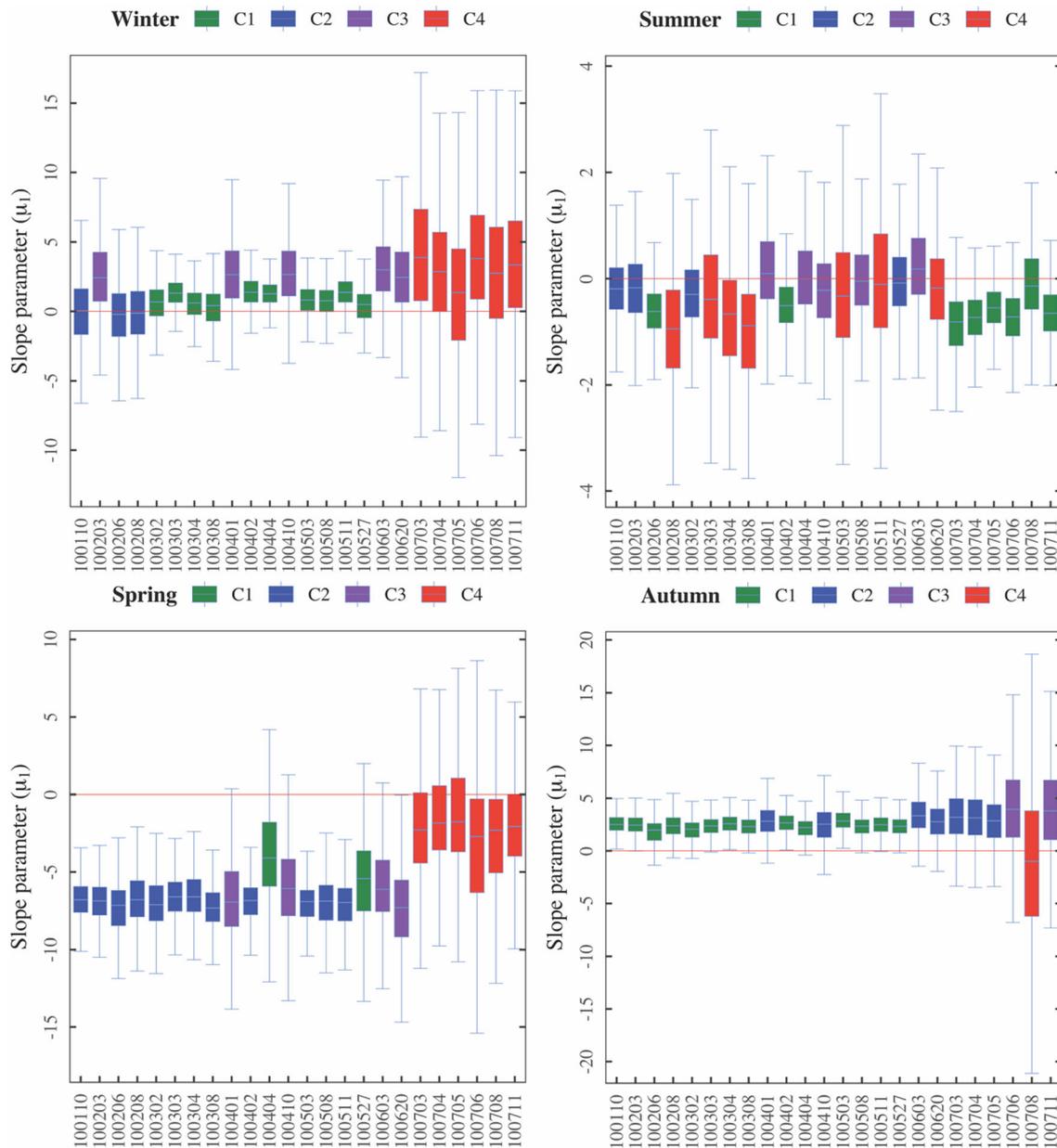
Fig. 8 Boxplot of the posterior distribution of the slope parameter $\mu_1(s)$ for each station by PPBC model with $K = 4$. Different colors present the identified clusters form the clustering models

Halpert 1987; Kiladis and Diaz 1989; Nicholson and Kim 1997; Mason and Goddard 2001; Lüdecke et al 2021) and Algeria (Meddi et al. 2010; Turki et al. 2016; Zeroual et al. 2016) in terms of the seasonal distribution of ENSO impacts on precipitation amount.

5 Summary and conclusions

This study aimed to analyze the effect of ENSO on the seasonal precipitation across the Kebir Rhumel Basin using Bayesian clustering approach. For each season, full pooling

Table 4 Percentage of stations with a significant effect of ENSO on seasonal precipitation in PPBC model

Clusters	Effect of ENSO		
	Significant	Positive	Negative
Winter			
Total	17	17	0
C1	17	17	0
C2	0	0	0
C3	0	0	0
C4	0	0	0
Spring			
Total	75	0	75
C1	8	0	8
C2	50	0	50
C3	17	0	17
C4	0	0	0
Summer			
Total	12	0	12
C1	12	0	12
C2	0	0	0
C3	0	0	0
Autumn			
Total	75	75	0
C1	54	54	0
C2	21	21	0
C3	0	0	0
C4	0	0	0

Bayesian clustering and partial pooling Bayesian clustering models with a nonstationary GEV distribution are applied. In these models, we assumed that the location parameter was linked to the temporal climate covariate using linear regression function. The intercept and the slope of the location parameter and the shape parameter were used for clustering. An advantage of the approach is that it allows the clustering and the model parameter estimation to proceed at the same time and reduce the uncertainty in the parameter estimation by transferring the information across stations with similar characteristics.

The main findings of this study are summarized as follows: (i) It was found that adding the number of clusters improves the fit of both Bayesian clustering models. (ii) For all four seasons, the partial pooling Bayesian clustering model with $K = 4$ provided the best fit for the seasonal precipitation data. (iii) ENSO significantly effects precipitation across large parts of the Kebir Rhumel Basin during spring and autumn seasons. (iv) In winter and autumn, 17% and 75% of the stations were found to be positively influenced by ENSO, respectively. On the contrary, 75% and 12% of the stations examined were negatively affected

by ENSO during spring and summer, respectively, indicating that the ENSO effect changed from one season to another. (v) The significant positive and negative influences of ENSO are observed in southern part and northern part of the Kebir Rhumel Basin during winter and summer, respectively. All stations except those found in Oued Kebir Maritime sub-basin are negatively and positively influenced by ENSO during spring and autumn, respectively.

An extension of the Bayesian clustering approach to consider several appropriate covariates at the same time is possible. In this study, we assumed a symmetric effect of the positive and negative phases of climate index, leading to a linear relationship between the ENSO measure SOI and the distribution location parameter. However, an asymmetric relation may better identify the influence of the climate variability on seasonal precipitation. As well, clustering models could also explicitly include spatial dependence across the stations. Sun et al. (2015b) demonstrated that considering spatial dependence in Hierarchical Bayesian clustering model can avoid under-estimating uncertainties. In future work, we expect to develop a strategy that can effectively model stations precipitation trends with adding various climate indices as a covariates using both symmetric and asymmetric analysis and considering the spatial dependence in a cluster.

Declarations

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical approval All the work is compliance with Ethical Standards.

Consent to participate The authors gave the consent to participate in this study.

References

Agilan V, Umamahesh NV (2017a) What are the best covariates for developing non-stationary rainfall intensity-duration-frequency relationship? Adv Water Resour 101:11–22

Agilan V, Umamahesh NV (2017b) Covariate and parameter uncertainty in non-stationary rainfall IDF curve. Int J Climatol 38:365–383

Akaike H (1974) New look at statistical-model identification. IEEE Trans Autom Control 19(6):716–723

Aryal SK, Bates BC, Campbell EP, Li Y, Palmer MJ, Viney NR (2009) Characterizing and modeling temporal and spatial trends in rainfall extremes. J Hydrometeorol 10(1):241–253

Belkhiri L, Kim TJ (2021) Individual influence of climate variability indices on annual maximum precipitation across the global scale. Water Resour Manage 35(9):2987–3003

Berghuijs WR, Larsen JR, van Emmerik THM, Woods RA (2017) A global assessment of runoff sensitivity to changes in precipitation, potential evaporation, and other factors. Water Resour Res 53(10):8475–8486

Biernacki C, Govaert G (1997) Using the classification likelihood to choose the number of clusters. Comput Sci Stat 29:451–457

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell 22(7):719–725

Chen X, Hao Z, Devineni N, Lall U (2014) Climate information based streamflow and rainfall forecasts for Huai River basin using hierarchical Bayesian modeling. Hydrol Earth Syst Sci 18(4):1539–1548

Cheng L, Aghakouchak A (2014) Nonstationary precipitation intensity-duration-fre- quency curves for infrastructure design in a changing climate. Sci Rep-UK 4:7093

Coles S, Bawa J, Trenner L, Dorazio P (2001) An introduction to statistical modeling of extreme values. Springer, London

Dai A, Fung IY, Del Genio AD (1997) Surface observed global land precipitation variations during 1900–88. J Clim 10(11):2943–2962

Delorio M, Robert CP (2002) Discussion of spiegelhalter. J Royal Statist Soc Ser B 64:629–630

Deng Hongbo, Han Jiawei (2018) Probabilistic models for clustering. In: Aggarwal CC, Reddy CK (eds) Data clustering: algorithms and applications. Chapman and Hall/CRC, Boca Raton, pp 61–86. https://doi.org/10.1201/9781315373515-3

Du H, Xia J, Zeng S, She D, Liu J (2014) Variations and statistical probability characteristic analysis of extreme precipitation events under climate change in Haihe River Basin. China Hydrol Process 28:913–925

Gao M, Mo D, Wu X (2016) Nonstationary modeling of extreme precipitation in China. Atmos Res 182:1–9

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457–472

Hoerling MP, Kumar A, Zhong M (1997) El Nino, La Nina, and the nonlinearity of their teleconnections. J Clim 10(8):1769–1786

Hoffman MD, Gelman A (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 15(1):1593–1623

IPCC CC (2021) The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge

Johnson DS, Ream RR, Towell RG, Williams MT, Leon Guerrero JD (2013) Bayesian clustering of animal abundance trends for inference and dimension reduction. J Agric Biol Environ Stat 18(3):299–313

Kiladis GN, Diaz HF (1989) Global climatic anomalies associated with extremes in the Southern oscillation. J Clim 2(9):1069–1090

Konapala G, Mishra AK, Wada Y, Mann ME (2020) Climate change will affect global water availability through compounding changes in seasonal precipitation and evaporation. Nat Commun 11(1):3044

Lüdecke HJ, Müller-Plath G, Wallace MG, Lüning S (2021) Decadal and multidecadal natural variability of African rainfall. J Hydrol Regional Stud 34:100795

Madsen H, Lawrence D, Lang M, Martinkova M, Kjeldsen TR (2014) Review of trend analysis and climate change projections of extreme precipitation and floods in Europe. J Hydrol 519:3634–3650

Marouf N, Remini BJJOW, Development L (2019) Impact study of Beni-Haroun dam on the environmental and socio-economic elements in Kébir-Rhumel basin. Algeria 43:120–132

Mason SJ, Goddard L (2001) Probabilistic precipitation anomalies associated with ENSO. Bull Am Meteorol Soc 82(4):619–638

McLachlan GJ, Peel D (2004) Finite mixture models. Wiley, New York

Mebarki A (2005) Thèse d'État: Hydrologie des bassins de l'Est algérien: ressources en eau, aménagement et environnement.

Meddi MM, Assani AA, Meddi H (2010) Temporal variability of annual rainfall in the Macta and Tafna catchments, northwestern Algeria. Water Resour Manage 24(14):3817–3833

Micevski T, Franks SW, Kuczera G (2006) Muttidecadal variability in coastal eastern Australian flood data. J Hydrol 327(1–2):219–225

Nicholson SE, Kim J (1997) The relationship of the El Nino-Southern oscillation to African rainfall. Int J Climatol 17(2):117–135

Nieto-Barajas LE, Contreras-Cristán A (2014) A Bayesian nonparametric approach for time series clustering. Bayesian Anal 9(1):147–170

Ossandón Á, Brunner MI, Rajagopalan B, Kleiber W (2022) A space–time Bayesian hierarchical modeling framework for projection of seasonal maximum streamflow. Hydrol Earth Syst Sci 26(1):149–166

Ouarda TBMJ, El-Adlouni S (2011) Bayesian nonstationary frequency analysis of hydrological variables. J Am Water Resour Assoc 47(3):496–505

Papalexiou SM, Koutsoyiannis D (2013) Battle of extreme value distributions: a global survey on extreme daily rainfall: survey on extreme daily rainfall. Water Resour Res 49(1):187–201. https://doi.org/10.1029/2012WR012557

Renard B, Sun X, Lang M (2013) Bayesian methods for nonstationary extreme value analysis. In: AghaKouchak A, Easterling D, Hsu K, Schubert S, Sorooshian S (eds) Extremes in a changing climate: detection, analysis and uncertainty. Springer Netherlands, Dordrecht, pp 39–95. https://doi.org/10.1007/978-94-007-4479-0_3

Renard B, Lang M, Bois P, Dupeyrat A, Mestre O, Niel H, Sauquet E, Prudhomme C, Parey S, Paquet E, Neppel L (2008) Regional methods for trend detection: assessing field significance and regional consistency. Water Resourc Res. 44(8)

Ropelewski CF, Halpert MS (1987) Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. Mon Weather Rev 115(8):1606–1626

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Silva AT, Portela MM, Naghettini M, Fernandes W (2017) A Bayesian peaks-over-threshold analysis of floods in the Itajaí-açu River under stationarity and nonstationarity. Stoch Env Res Risk Assess 31(1):185–204

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B 64(4):583–639

Stan Development Team (2022) Rstan: the R interface to Stan. R package version 2.26.4. https://CRAN.R-project.org/package=rstan

Steirou E, Gerlitz L, Apel H, Merz B (2017) Links between large-scale circulation patterns and streamflow in Central Europe: a review. J Hydrol 549:484–500

Steirou E, Gerlitz L, Apel H, Sun X, Merz B (2019) Climate influences on flood probabilities across Europe. Hydrol Earth Syst Sci 23(3):1305–1322

Su C, Chen X (2019) Covariates for nonstationary modeling of extreme precipitation in the Pearl River Basin. China Atmospheric Research 229:224–239

Sun X, Lall U (2015) Spatially coherent trends of annual maximum daily precipitation in the United States. Geophys Res Lett 42(22):9781–9789

Sun X, Thyer M, Renard B, Lang M (2014) A general regional frequency analysis framework for quantifying local-scale climate effects: a case study of ENSO effects on Southeast Queensland rainfall. J Hydrol 512:53–68

Sun X, Renard B, Thyer M, Westra S, Lang M (2015a) A global analysis of the asymmetric effect of ENSO on extreme precipitation. J Hydrol 530:51–65

Sun X, Lall U, Merz B, Dung NV (2015b) Hierarchical Bayesian clustering for nonstationary flood frequency analysis: application to trends of annual maximum flow in Germany. Water Resour Res 51(8):6586–6601

Trenberth KE (1997) The definition of El Niño. Bull Amer Meteor Soc 78:2771–2777

Trenberth KE, Stepaniak DP (2001) Indices of El Niño evolution. J Clim 14(8):1697–1701

Turki I, Laignel B, Massei N, Nouaceur Z, Benhamiche N, Madani K (2016) Hydrological variability of the Soummam watershed (Northeastern Algeria) and the possible links to climate fluctuations. Arab J Geosci. https://doi.org/10.1007/s12517-016-2448-0

Ward PJ, Eisner S, Flörke M, Dettinger MD, Kummu M (2014) Annual flood sensitivities to El Niño-Southern Oscillation at the global scale. Hydrol Earth Syst Sci 18(1):47–66

Xiong Y, Yeung DY (2004) Time series clustering with ARMA mixtures. Pattern Recogn 37(8):1675–1689

Zeroual A, Assani AA, Meddi M (2016) Combined analysis of temperature and rainfall variability as they relate to climate indices in Northern Algeria over the 1972–2013 period. Hydrol Res 48(2):584–595