Research Article

Amjad Hussein* and Safaa K. Kadhem

Spatial mixture modeling for analyzing a rainfall pattern: A case study in Ireland

https://doi.org/10.1515/eng-2022-0024
received August 23, 2021; accepted October 23, 2021

Abstract: This study investigates the spatial heterogeneity in the maximum monthly rainfall amounts reported by stations in Ireland from January 2018 to December 2020. The heterogeneity is modeled by the Bayesian normal mixture model with different ranks. The selection of the best model or the degree of heterogeneity is implemented using four criteria which are the modified Akaike information criterion, the modified Bayesian information criterion, the deviance information criterion, and the widely applicable information criterion. The estimation and model selection process is implemented using the Gibbs sampling. The results show that the maximum monthly rainfall amounts are accommodated in two and three components. The goodness of fit for the selected models is checked using the graphical plots including the probability density function and cumulative distribution function. This article also contributes via the spatial determination of return level or rainfall amounts at risk with different return periods using the prediction intervals constructed from the posterior predictive distribution.

Keywords: rainfall amounts, Bayesian mixture modeling, cumulative distribution, prediction interval, posterior predictive distribution

1 Introduction

Rainwater, also called precipitation, is a natural feature of the earth's weather system. Air currents in the atmosphere bring evaporated water from the ocean and the earth's surface into the sky. The evaporated liquid condenses in the cold air, forming moisture-filled rain clouds [1]. Rain water's most well-known and most important effect is providing water to drink. According to the United States Geological Survey, rainwater seeps into the ground in a process called infiltration. Some of the water seeps deep beneath the top layers of soil where it fills up the space between subsurface rocks and becomes groundwater, also called the water table. Less than 2% of the earth's water is groundwater, but it provides 30% of our freshwater. Without rain water's continued replenishment of the water table, potable water would become scarcer than it already is in ref. [2]. Furthermore, many researchers demonstrated the impact of heavy rainfall on floods [3–5]. Therefore, the analysis of precipitation quantities on a certain area aims to give an overall visualization or prior information to evaluate the risk of some natural disasters such as droughts, floods, landslides, and so in ref. [6].

Many researchers have used different analysis methods to analyze the rainfall trend. Meneghini et al. [7] used different statistical methods such as the area-time integral (ATI) method to estimate the average rainfall over a large space. Arvind et al. [8] used different analysis methods on the Annual and Monthly rainfall for Musiri Region; they concluded that Gumbel distribution is the best type. Panda and Sahu [9] used the Mann–Kendall test as a statistical method and Sen's slope estimator to examine and analyze the seasonal rainfall over the state of Odisha in India. Their results showed a relatively maximum amount of rainfall in monsoonal months. Nyatuame et al. [10] used linear regression analysis as a statistical method for annual and monthly rainfall. They stated an insignificant increasing trend in annual mean rainfall data among the Volta Region and a significant trend in mean monthly rainfall. Asfaw et al. [11] inspected the change of rainfall and temperature in north-central Ethiopia

* **Corresponding author: Amjad Hussein,** Department of Civil Engineering, College of Engineering, Al-Muthanna University Al-Muthanna, Samawah, Iraq; Civil Engineering Research Group, School of Computing, Science and Engineering, University of Salford, Salford, Greater Manchester M5 4WT, United Kingdom, e-mail: amjad.muhamad@mu.edu.iq, tel: +96-47807921448
Safaa K. Kadhem: Department of Mathematics and Computer Applications, College of Science, Al-Muthanna University, Samawah, Iraq

using gridded monthly precipitation data. The Mann–Kendall was used to detect the time series trend. Praveen et al. [12] analyzed and forecasted rainfall changed in India. They used Pettitt and Mann–Kendall tests as analysis methods. However, this group of studies did not take into account the variation in data.

This study assumes that the maximum rainfall quantities follow the normal distribution assuming that rains are distributed equally throughout Ireland. However, this assumption is invalid due to variation or fluctuation in rainfall amounts, leading to a phenomenon called the heterogeneity of the data. Identifying spatial heterogeneity of rainfall can give a valuable indicator for the analytical studies and government planning to detect the factors linked to various causes to the low or height in the rain falling. For this reason, the statistical modeling method called the mixture model with a finite number of components was proposed to accommodate the heterogeneity in data. The task of finite mixture models is to capture unobserved heterogeneity in the population by assuming that the population consists of K homogeneous subgroups [13]. However, identifying the number of homogeneous subgroups K or the rank of model forms is more challenging. Several criteria have been addressed in the literature to determine the model's rank under both the frequentist and the Bayesian settings. As in this article, the Bayesian framework focused on estimating the model parameters, four well-known model selection criteria derived under the Bayesian principle. These criteria are the modified Akaike information criterion (AIC) [14], the modified Bayesian information criterion (BIC) [15], the deviance information criterion (DIC) [16], and the widely applicable information criterion (WAIC) [17]. An approach was followed to fit a set of candidate models to the data and select the best one. In other words, in this research, the assumption that the number of the model components is fixed and unknown and the best model is determined by one of our four proposed criteria via fitting several candidate mixture models with different components. Despite that, another approach called the reversible jump Markov chain Monte Carlo sampling [18] can be applied to select the appropriate number of components. However, this latter approach has drawbacks when the Markov chain moves between mixture models with different classes [13].

This article also determines the spatial return level or rainfall amounts at risk with different return periods using the prediction intervals constructed from the posterior predictive distribution. This latter can be considered as a newly developed alternative approach to the confidence intervals adopted by several kinds of literature to identify the rainfall amounts at risk [4,6].

This article is classified as follows. Section 2 introduces the article's methodology, including the model's construction, estimation, and election. Section 3 includes a description of the data under study. The results and discussion are shown in Section 4. Finally, Section 5 summarizes the important conclusions of this article.

2 Methodology

This section introduced the building of the model and estimation of the model parameters under the Bayesian principle. After that, the best model to fit the maximum was selected. Each station's monthly rainfall quantity is reported under study using model selection criteria such as AIC, BIC, DIC, and WAIC. In addition, the goodness of fit was also checked.

2.1 Model construction and Bayesian analysis

Let us assume that the study region is divided into m stations, and let y_i represent the maximum monthly rainfall quantity reported by ith station, $i = 1, 2, \ldots, n$. To take into account the heterogeneity in the data of rainfall quantities, those data follow a mixture of univariate Gaussian distribution:

$$\Pr(\mathbf{y}|k, \boldsymbol{w}, \mu, \sigma^2) = \sum_{j=1}^{k} w_j N_j(\mathbf{y}|\mu_j, \sigma_j^2), \tag{1}$$

where k is the number of components of the model (that can be viewed as levels of monthly rainfall quantity), $\boldsymbol{w} = (w_1, w_2, \ldots, w_k)$ is the vector of probabilities associated with the components of the model, with $\sum_{j=1}^{k} w_j = 1$ and $w_j \geq 0$, and $N(y_i|.)$ represent the mixed Gaussian probability density function (PDF) which is defined as

$$N_j(\mathbf{y}|\mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{1}{2\sigma_j^2}(\mathbf{y} - \mu_j)^2\right\}, \tag{2}$$

where μ_j and σ_j^2 are the mean and variance of jth mixture of the Gaussian distribution, respectively. The mixture models can be analytically easier by including latent variables in their formation as this latter makes it more useful for the purpose of interpretation and numerical computations. For a mixture with a certain number of components, the model can be described by inserting n independent discrete variables, $z_1, z_2, \ldots, z_n$, with the multinomial

distribution $\Pr(z_i = j|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}, K) = w_j$, for $j = 1, 2, \ldots, K$. Given $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, equation (1) can be written as

$$\Pr(\mathbf{y}, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}, K) = \prod_{i=1}^{n}\prod_{j=1}^{K}\{w_j N_j(y_i|\mu_j, \sigma_j^2)\}^{z_{ij}}, \quad (3)$$

which is called the complete-data likelihood function. The role of latent variable, z_i, is to assign the observation y_i to one of the mixture components. By taking the logarithm for equation (3), we obtain:

$$\begin{aligned}\ell(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}|\mathbf{y}, \mathbf{z}) &= \log\left[\prod_{i=1}^{n}\prod_{j=1}^{K}\{w_j N_j(y_i|\mu_j, \sigma_j^2)\}^{z_{ij}}\right], \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K}[\log w_{z_{ij}} N_j(y_i|\mu_{z_{ij}}, \sigma_{z_{ij}}^2)].\end{aligned} \quad (4)$$

The log-likelihood function in equation (4) can be approximated over the posterior distribution. For example, given $(\ell^{(0)}, \ell^{(1)}, \ldots, \ell^{(M)})$ computed over a full Monte Carlo Markov chain (MCMC) run, obtaining the estimated log-likelihood by post-processing the posterior outcome:

$$\begin{aligned}&\hat{\ell}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}|\mathbf{y}, \mathbf{z}) \\ &= \frac{1}{M}\sum_{m=1}^{M}\left(\sum_{i=1}^{n}\sum_{j=1}^{K}[\log w_{z_{ij}^{(m)}}^{(m)} N(y_i|\mu_{z_{ij}^{(m)}}^{(m)}, \sigma_{z_{ij}^{(m)}}^{2(m)})]\right).\end{aligned} \quad (5)$$

To complete the Bayesian analysis for the model, we have to define the prior and posterior distributions for all the model parameters. For this purpose, Algorithm 1, given by ref. [19], is used to implement the sampling process using one of the MCMC approaches that is called Gibbs sampler.

Algorithm 1: Gibbs sampler for K-component normal mixture model

1. Initialization: Choose $w_j^{(0)}$, $\mu_j^{(0)}$ and $\sigma_j^{2(0)}$, $j = 1, 2, \ldots, K$.
2. Iteration: for $m = 1, 2, \ldots, M$
   (a) Generate $z_t^{(m)}$; $t = 1, \ldots, T$ from ($j = 1, 2, \ldots, K$)
   $$\Pr(z_t^{(m)} = j) = \propto \frac{w_j^{(m-1)}}{\sigma_j^{(m-1)}} \exp\left(-\frac{(y_t - \mu_j^{(m-1)})^2}{2(\sigma_j^2)^{(m-1)}}\right),$$
   and compute: $n_k^{(m)} = \sum_{l=1}^{n} \mathbb{I}_{z_l^{(m)}=j}$
   and $s_j^{y(m)} = \sum_{l=1}^{n} \mathbb{I}_{z_l^{(m)}=j} y_l$.
   (b) Update $w_j^{(m)}$ from
   $\mathrm{Dir}(\delta_1 + n_1^{(m)}, \delta_2 + n_2^{(m)}, \ldots, \delta_K + n_K^{(m)})$,
   for $j = 1, 2, \ldots, K$
   (c) Generate $\mu_j^{(m)}$; $j = 1, 2, \ldots, K$
   from $N\left(\frac{\eta_j\zeta_j + s_j^{y(m)}}{\zeta_j + (n_j)^{(m)}}, \frac{\sigma_j^{2(m-1)}}{\zeta_j + (n_j)^{(m)}}\right)$,
   and compute: $s_j^{y(m)} = \sum_{t=1}^{n} \mathbb{I}_{z_l^{(m)}=j}(y_t - \mu_j^{(m)})^2$.
   (d) Generate $\sigma_j^{2(m)}$; $j = 1, 2, \ldots, K$ from
   InvGamma
   $(a_j + 0.5(n_j^{(m)} + 1)$,
   $b_j + 0.5\zeta_j(\mu_j^{(m)} - \eta_j)^2 + 0.5(s_j^{y(m)}))$.

where η_j, ζ_j, a_j, b_j, and δ_j are known hyper-parameters, $j = 1, 2, \ldots, K$, and they are commonly given non-informative hyper-priors or flat values [20]. For instance, the inverse Gamma with parameters $a = 0.001$ and $b = 0.001$ and thus a mean of $a/b = 1$ and a variance of $a/b^2 = 1{,}000$ can give diffuse values of this form. The prior of the mean parameter can be assigned flat values from a Normal distribution with a shape parameter, $\eta = 0$, and a scale parameter, $\zeta = 0.001$, which has a large variance equal to 1,000. The weight parameter, π, is given a Dirichlet prior with non-informative value, $\delta_j = 1$, $j = 1, 2, \ldots, K$.

2.2 Model selection criteria

These first three sections introduce four criteria for choosing the number of components in Gaussian mixture models in Bayesian settings. In the last section, a graphitic display method to evaluate the goodness of fitness of the model is shown.

2.2.1 Akaike information and Bayesian information criteria

Two well-known criteria modified were introduced under the Bayesian principle, which are the AIC and BIC proposed by Kadhem et al. [21]. These criteria depend on the deviance and penality term. From equation (5), the deviance can be defined as twice the negative log likelihood:

$$D(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}) = -2\{\hat{\ell}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}|\mathbf{y}, \mathbf{z})\}, \quad (6)$$

where the deviance above is approximated over MCMC samples as explained in equation (5). The penality term is computed based on the free parameters of the model as: $h = 3\ \mathrm{K}^{-1}$ [22]. The AIC and BIC take the deviance as a measure of model fit and penalizes it for the number of parameters in the model. Then, AIC and BIC are given as follows:

$$\begin{aligned}\mathrm{AIC} &= D(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}) + 2h, \\ \mathrm{BIC} &= D(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}) + h\log(n),\end{aligned} \quad (7)$$

where n is the sample size.

2.2.2 DIC

Another criterion proposed in this article is the DIC. Eight versions of this criterion were introduced by Celeux et al. [23]. They recommended the version that is based on the complete-data likelihood. In this article, we apply this version which is given by:

$$\begin{aligned}\mathrm{DIC} &= -4E_{\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{w},\mathbf{z}}[\log \Pr(\mathbf{y}, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}, K)] \\ &\quad + 2E_{\mathbf{z}}[\log \Pr(\mathbf{y}, \mathbf{z}|\hat{\boldsymbol{\mu}}(\mathbf{z}), \hat{\boldsymbol{\sigma}}^2(\mathbf{z}), \hat{\boldsymbol{w}}(\mathbf{z}))],\end{aligned} \tag{8}$$

with its effect number of parameters, p_{DIC}, defined as follows:

$$\begin{aligned}p_{\mathrm{DIC}} &= -2E_{\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{w},\mathbf{z}}[\log \Pr(\mathbf{y}, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}, K)] \\ &\quad + 2E_{\mathbf{z}}[\log \Pr(\mathbf{y}, \mathbf{z}|\hat{\boldsymbol{\mu}}(\mathbf{z}), \hat{\boldsymbol{\sigma}}^2(\mathbf{z}), \hat{\boldsymbol{w}}(\mathbf{z}))],\end{aligned} \tag{9}$$

where $\hat{\boldsymbol{\mu}}(\mathbf{z})$, $\hat{\boldsymbol{\sigma}}^2(\mathbf{z})$, and $\hat{\boldsymbol{w}}(\mathbf{z})$ are the complete-data posterior modes of the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$, and $\boldsymbol{w}$, respectively, which are computed for each samples from the posterior $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w})$.

2.2.3 WAIC

The last criterion proposed in this article is the WAIC. This criterion is fully Bayesian and it is computed based on the so-called integrated pointwise predictive density (ilppd). For a Gaussian mixture distribution, the ilppd can be defined as follows:

$$\begin{aligned}\mathrm{ilppd}_{\mathbf{y}} &= \log \prod_{i=1}^{n} \mathrm{Gaussian}_{\mathrm{post}}(y_i), \\ &= \sum_{i=1}^{n} \log E_{\{\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{w},\mathbf{z}\}}[\mathrm{Gaussian}(y_i|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}, \mathbf{z})|\mathbf{y}], \\ &= \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} \log \mathrm{Gaussian}(y_i|\mu_{z_i^{(m)}}^{(m)}, \sigma_{z_i^{(m)}}^{2(m)}, w_{z_i^{(m)}}^{(m)}),\end{aligned} \tag{10}$$

where $\mu_{z_{it}^{(m)}}^{(m)}$, $\sigma_{z_i^{(m)}}^{2(m)}$, and $w_{z_i^{(m)}}^{(m)}$ represent the mth sample drawn in Gibbs sampler. To complete the definition of the WAIC, Gelman et al. [20] proposed adding a correction term or the so-called effect number of parameters p_{WAIC}, to avoid the bias. This number is based on computing the variance of individual terms in the ilppd, which is defined as follows:

$$\begin{aligned}p_{\mathrm{WAIC}} &= \sum_{i=1}^{n} V_{\{\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{w},\mathbf{z}\}}[\log \mathrm{Gaussian}(y_i|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}, \mathbf{z})|\mathbf{y}], \\ &= \sum_{m=1}^{M} \sum_{i=1}^{n} V_{\{\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{w},\mathbf{z}\}}[\log \mathrm{Gaussian} \\ &\quad \times (y_i|\mu_{z_i^{(m)}}^{(m)}, \sigma_{z_i^{(m)}}^{2(m)}, w_{z_i^{(m)}}^{(m)})].\end{aligned} \tag{11}$$

The WAIC then is constructed as follows:

$$\mathrm{WAIC} = -2\widehat{\mathrm{ilppd}}_{\mathbf{Y}} + 2p_{\mathrm{WAIC}}. \tag{12}$$

2.2.4 Graphical display method

In this research, the cumulative distribution function (CDF) is used as one of the graphical display methods to reinforce the correct model chosen by the model selection criteria above. The CDF plot is implemented to visualize the fitness of model distributions where it is monotonically increasing between the limits from 0 to 1. The CDF of a Gaussian mixture with K components can be given as

$$F(x) = w_1F_1(x) + w_2F_2(x) + \ldots + w_KF_K(x). \tag{13}$$

2.3 Analysis of rainfall amounts at risk

In this section, the so-called prediction intervals that are being constructed from the posterior predictive distribution were introduced [20] to analyze the extreme amounts of rainfall. The predicted values can be used as a goodness of fit approach to prediction accuracy of a statistical model. The limits of prediction interval can be constructed by the lower prediction limit, LPI (y^*), and upper prediction limit, UPI (y^*), where y^* represent the predicted data. In such cases, the interval [LPI (y^*), UPI (y^*)] is termed as the prediction interval and has a prediction coefficient of $(1 - p)100\%$. By introducing the prediction interval, the range that T-year rainfall takes can be theoretically estimated and also becomes possible to estimate the swing of T-year probability hydrological quantities in flood control measures for a T-year probability scale. This makes it possible to interpret the record-breaking heavy rainfall mentioned above as a phenomenon within the prediction interval. In other words, the prediction interval can be a tool to evaluate the return period of heavy rainfall.

In this research, the prediction interval is constructed as follows. Given the estimation of the model parameters sampling through an MCMC, $(w_j^{(m)}, \mu_j^{(m)}, \sigma_j^{2(m)}; m = 1, 2, \ldots, M)$ and observed infections data, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, for each region i at the time t, the PPD for predicted infections, $\mathbf{y}_i^*$; $i = 1, 2, \ldots, n$ of the normal mixture model can be defined as

$$\begin{aligned}\Pr(\mathrm{y}_i^*|\mathbf{y}) = \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\sigma}^2} \int_{\mathbf{z}} &\mathrm{Normal}(\mathrm{y}_{it}^*|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}, \mathbf{z}) \\ &\times \mathrm{Normal}_{\mathrm{post}}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}, \mathbf{z}|\mathbf{y})\mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\sigma}^2,\end{aligned} \tag{14}$$

where $\text{Normal}_{\text{post}}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{w}, \mathbf{z}|\mathbf{y})$ represents the joint complete posterior distribution. Given samples of the relative risk parameter, $\lambda_{jt}^{(m)}$, and latent variables, $\mathbf{z}^{(m)}$, drawn from an MCMC run, the predictive data of a Poisson mixture model can be approximated as

$$\mathbf{y}_i^{*(m)} \sim \text{Normal}\left(\mu_{z_{ijt}^{(m)}}^{(m)}, \sigma_{z_{ij}^{(m)}}^{2(m)}\right); \quad i = 1, 2, \ldots, n; \; j = 1, 2, \ldots, K. \tag{15}$$

Hence, the prediction interval can be formulated as follows:

$$\mathbf{y}_{i+h}^{*} \pm c\hat{\sigma}_h, \tag{16}$$

where $\hat{\sigma}_h$ is an estimate of the standard deviation of the h-step forecast distribution and c is the multiplier that includes a range of coverage probabilities assuming a normal forecast distribution.

Given that the return period can be calculated as follows. Let us assume that X is the variable that equals to or greater than an incident of magnitude x_T occurring once in T years. In a given year, the probability of occurrence of incident X, $\Pr(X \geq x)$, is expressed as:

$$\Pr(X \geq x) = \frac{1}{T}, \tag{17}$$

$$T = \frac{1}{1 - \Pr(X \geq x)}. \tag{18}$$

Wilks [6] pointed out that the amounts of maximum monthly rainfall with the 50-year or 100-year return periods cannot be directly calculated from the data set used here, but have to be extrapolated from the 98th and 99th percentiles of the fitted distribution, respectively, i.e., $[1 - 0.98^{-\text{year}}]^{-1} = 50$ years and $[1 - 0.99^{-\text{year}}]^{-1} = 100$ years. On this basis, we can compute the return periods based on the prediction intervals.

3 Data description

Ireland is an island in North western Europe in the North Atlantic Ocean with 84,421 km^2 (land; 98.2%, water; 1.8%). The coordinates are 53°20.65′N6°16.06′W. It lies on the European continental shelf, part of the Eurasian Plate. Low lowlands and low mountainous beaches characterize it. Corran Tuathail, at 1,041 m above sea level, is the highest summit. The western coast has a rough shoreline with many islands, peninsulas, headlands, and bays. The Shannon River is Ireland's longest river, flowing south from County Cavan in Ulster to the Atlantic Ocean. Along Ireland's rivers, there are several big lakes, the greatest of which being Lough Neagh: 392 km^2. More than 134,600 ha

Table 1: Geographic characteristics of the stations under study

No.	Stations	Latitude (N)	Longitude (E)	Elevation (m)
1	Ballyhaise	52.69	−8.918	15
2	Shannon Airport	51.847	−8.486	155
3	Cork Airport	51.793	−8.244	40
4	Roches Point	51.576	−9.428	21
5	Sherkin Island	52.164	−8.264	46
6	Moore Park	54.494	−8.243	33
7	Finner Camp	55.372	−7.339	20
8	Malin Head	53.248	−6.241	71
9	Dublin Airport	53.364	−6.35	48
10	Phoenix Park	53.289	−8.786	40
11	Athenry	53.326	−9.901	21
12	Mace Head	53.306	−6.439	91
13	Casement	51.948	−10.241	24
14	Valentia Observatory	54.051	−7.31	78
15	Belmullet	54.228	−10.007	9
16	Claremorris	53.711	−8.993	68
17	Knock Airport	53.906	−8.817	201
18	Dunsany	53.516	−6.66	83
19	Newport Furnace	53.922	−9.572	22
20	Gurteen	53.053	−8.009	75
21	Markree Castle	54.175	−8.456	34
22	Mullingar	53.537	−7.362	101
23	Johnstown Castle	52.298	−6.497	62
24	Oak Park	52.861	−6.915	62
25	Mount Dillon	53.727	−7.981	39

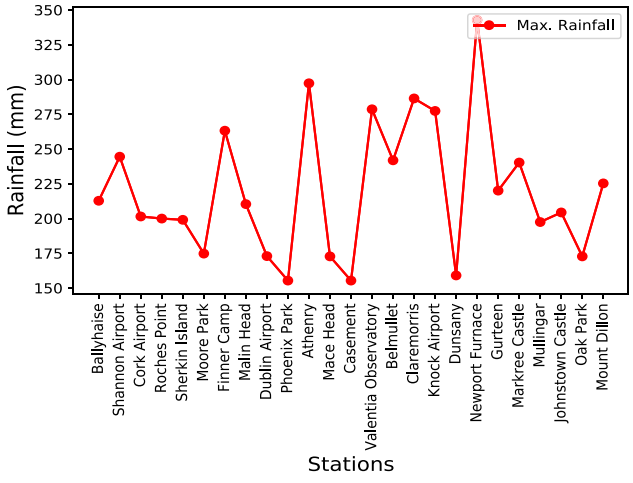

Figure 1: The maximum values of rainfall amounts for all stations over period from January 2018 to December 2020.

(19.5% of the total area) were devoted to growing crops; 6 and 1.5% of the agricultural area were used to grow cereals, and root and green crops, respectively. Over half of the agricultural production is exported. The income was increased from 314£ million in 1972 to 1920£ million in 1995 and 1.1£ billion in 2001. Ireland's climate is mild, moist, and changeable, with abundant rainfall and a lack of temperature extremes. Ireland's climate is defined as a temperate oceanic climate. In general, Iceland has warm summers and pleasant winters, as compared to, say, Newfoundland, which is significantly warmer at the same latitude and located downwind of the Atlantic Ocean. It is also hotter than marine climates around the same latitude, such as the Pacific Northwest, due to the heat released by the Atlantic overturning circulation, which includes the North Atlantic Current and Gulf Stream. In comparison, Dublin is 9° warmer in the winter than St. John's, Newfoundland, and 4° warmer than Seattle, Washington [23]. Ireland's climate is not vulnerable to extreme weather phenomena, such as tornadoes, and storms are uncommon. Throughout the winter, the North Atlantic Current keeps the Irish coast clear of ice. Ireland, on the other hand, is vulnerable to storms heading eastward from the North Atlantic. The prevailing wind is from the southwest, and it breaks on the west coast's steep mountains. As a result, off the west coast of County Kerry, Valentia Island receives nearly twice as much rain as Dublin, on the east coast (1,400 vs 762) mm, demonstrating the importance of rainfall in Western Ireland. The coldest months are January and February, with average daily air temperatures ranging from four to 7 °C. July and August are the hottest months, with average daily air temperatures of 14 to 16 °C. In July and August, daily maximum temperatures range from 17 to 18° along the coast to 19 to 20° inland. May and June are the months with the highest sunshine, with an average of 5–7 h per day. Extreme weather occurrences do occur, notwithstanding their rarity in comparison to other European countries. Atlantic depressions can bring gusts of up to 160 km/h to Western coastal regions in December, January, and February. Thunderstorms are common during the summer months, especially in late July and early August. This article investigates the rainfall pattern via analyzing the monthly maximum rainfall quantities reported by the stations on Island. The data on the rainfall

Table 2: Results of the best model selected by all or most selection criteria for every station

		No. of components						Best model
No.	**Stations**	**1**	**2**	**3**	**4**	**5**	**6**	**(Selection %)**
1	Ballyhaise	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
2	Shannon Airport	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
3	Cork Airport	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
4	Roches Point	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
5	Sherkin Island	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
6	Moore Park	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
7	Finner Camp	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
8	Malin Head	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
9	Dublin Airport	—	AIC	BIC, DIC, WAIC	—	—	—	$K = 3$ (75%)
10	Phoenix Park	—	AIC	BIC, DIC, WAIC	—	—	—	$K = 3$ (75%)
11	Athenry	—	AIC	BIC, DIC, WAIC	—	—	—	$K = 3$ (75%)
12	Mace Head	—	AIC	BIC, DIC, WAIC	—	—	—	$K = 3$ (75%)
13	Casement	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
14	Valentia Observatory	—	AIC, BIC, DIC	WAIC	—	—	—	$K = 2$ (75%)
15	Belmullet	—	AIC, BIC, DIC	WAIC	—	—	—	$K = 2$ (75%)
16	Claremorris	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
17	Knock Airport	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
18	Dunsany	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
19	Newport Furnace	—	AIC	BIC, DIC, WAIC	—	—	—	$K = 3$ (75%)
20	Gurteen	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
21	Markree Castle	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)
22	Mullingar	—	AIC	BIC, DIC, WAIC	—	—	—	$K = 3$ (75%)
23	Johnstown Castle	—	AIC	BIC, DIC, WAIC	—	—	—	$K = 3$ (75%)
24	Oak Park	—	AIC, BIC, DIC	WAIC	—	—	—	$K = 2$ (75%)
25	Mount Dillon	—	AIC, BIC, DIC, WAIC	—	—	—	—	$K = 2$ (100%)

are obtained from Met' Eireann, Ireland's National Meteorological Service, via web site (https://www.met.ie/climate/available-data/monthly-data#top) as shown in Table 1. The maximum values of rainfall amounts for all stations are presented from January 2018 to December 2020, as shown in Figure 1.

4 Results and discussion

For every station, the aim is to select the best normal mixture fitted to the monthly maximum rainfall amounts for return periods of 50 and 100 years expressed with prediction interval periods derived from a posterior predictive distribution. Table 2 shows the results of model selection, where the maximum rainfall amounts reported by the stations follow models either with two or three components as produced by the proposed criteria. Note that model with $K = 1$ (standard normal distribution) is not selected by all model selection criteria suggesting that the data suffer from the heterogeneity. The model estimation results for the models selected by the criteria are shown in Table 3, where the estimated weights, means, and variances represent the estimated parameters of the models selected by our proposed criteria. In addition, we provide the CDF and PDF plots for the model selection, in Figures 2 and 3, respectively, which appear as adequate goodness of fit. It can be noted from Table 2 that most of the rainfall data reported by stations (18 stations) are modeled by models with $K = 2$, except seven stations such as Dublin Airport, Phoenix Park, Athenry, Mace Head, Newport Furnace, Mullingar, and Johnstown Castle are modeled by mixture model with $K = 3$.

For the results of determination of the return periods, Figures 2 and 3 show the estimation of risk levels for a certain incident, with including a plot of the actual maximum monthly rainfall amounts in red color (as threshold), which are represented by the heights of rainfall levels of 50 and 100 year return periods with 95% prediction intervals. It can see from Figure 4 that there is a stable in rainfall amounts over 50 years return period (all values of mid-point of prediction intervals under the actual data), but there is a notable increase in the rainfall

Table 3: Results of the best model selected by the model selection criteria

	Weights			Means			Variances		
Station	w_1	w_2	w_3	μ_1	μ_2	μ_3	σ_1^2	σ_2^2	σ_3^2
Ballyhaise	0.54	0.45	—	61.81	121.91	—	21.76	41.31	—
Shannon Airport	0.78	0.21	—	70.90	170.42	—	29.79	39.32	—
Cork Airport	0.71	0.28	—	82.57	177.62	—	33.55	17.44	—
Roches Point	0.55	0.44	—	60.97	133.86	—	23.21	35.791	—
Sherkin Island	0.55	0.44	—	68.29	141.36	—	26.44	33.82	—
Moore Park	0.56	0.43	—	57.26	134.74	—	20.07	26.65	—
Finner Camp	0.63	0.36	—	78.24	164.70	—	27.02	43.46	—
Malin Head	0.51	0.48	—	73.48	126.48	—	26.05	39.82	—
Dublin Airport	0.49	0.41	0.09	34.51	84.73	140.68	15.49	11.53	23.37
Phoenix Park	0.47	0.44	0.08	33.27	80.30	139.25	15.74	12.67	14.69
Athenry	0.62	0.32	0.05	76.90	148.94	272.88	22.99	24.85	24.48
Mace Head	0.31	0.19	0.49	46.24	160.97	97.68	16.11	10.36	16.74
Casement	0.36	0.63	—	30.60	84.88	—	12.52	29.65	—
Valentia Observatory	0.57	0.42	—	101.25	216.01	—	37.68	30.20	—
Belmullet	0.24	0.75	—	47.43	133.60	—	16.93	42.71	—
Claremorris	0.75	0.24	—	98.21	182.56	—	39.09	54.85	—
Knock Airport	0.64	0.35	—	103.33	186.36	—	36.37	39.62	—
Dunsany	0.79	0.20	—	58.02	139.22	—	25.10	17.04	—
Newport Furnace	0.264	0.43	0.30	152.88	244.76	73.52	14.69	44.09	25.04
Gurteen	0.61	0.38	—	61.67	106.88	—	26.86	48.08	—
Markree Castle	0.83	0.16	—	89.19	202.55	—	30.27	24.39	—
Mullingar	0.47	0.29	0.24	47.01	134.91	86.86	15.53	28.73	11.59
Johnstown Castle	0.36	0.42	0.22	41.83	100.52	166.24	15.48	18.02	19.36
Oak Park	0.56	0.44	—	44.25	108.62	—	20.91	29.38	—
Mount Dillon	0.68	0.32	—	76.63	132.65	—	27.12	52.12	—

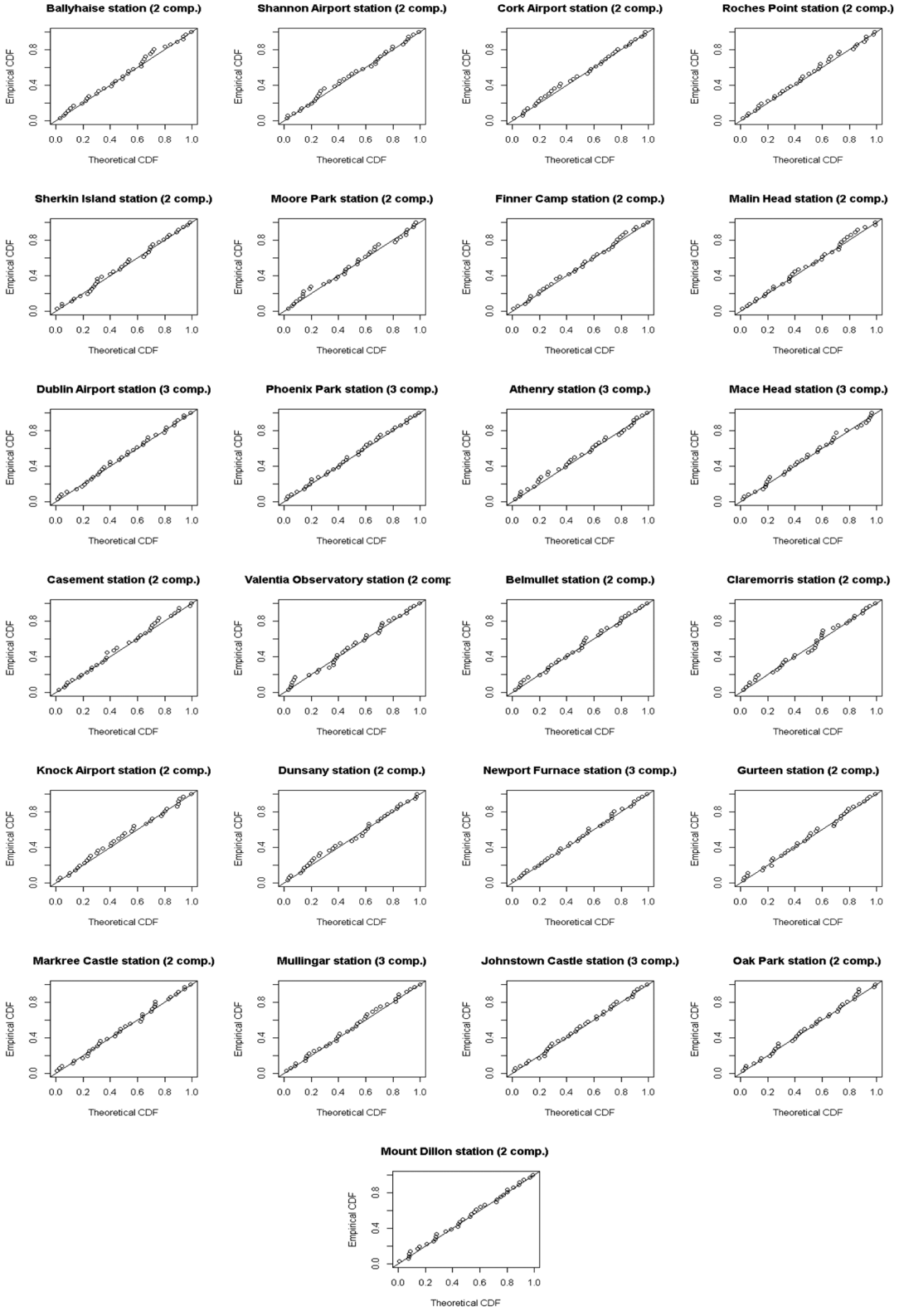

Figure 2: The CDF plots for the model selected by the proposed model selection criteria for each station.

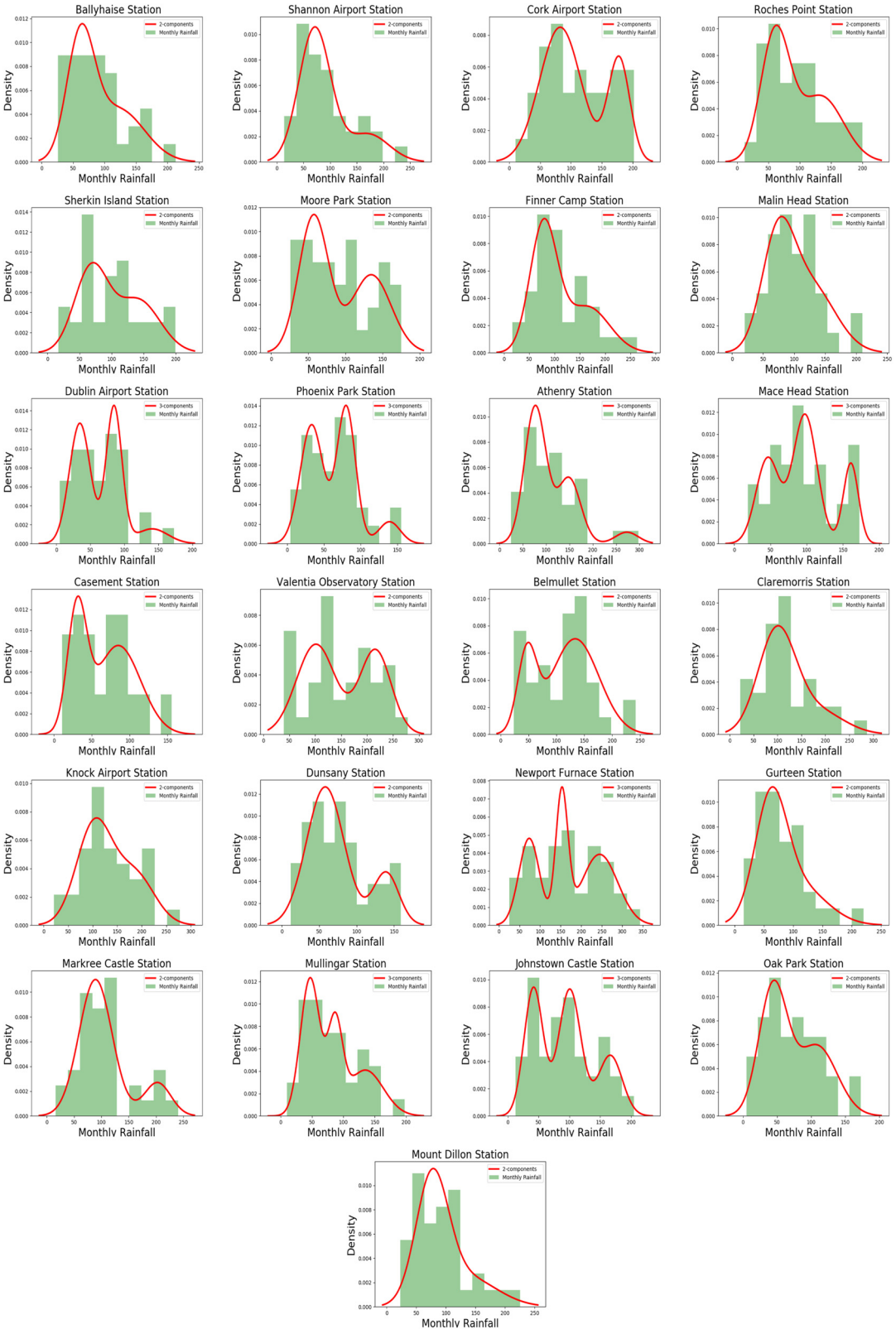

Figure 3: The best-fitting model for each station.

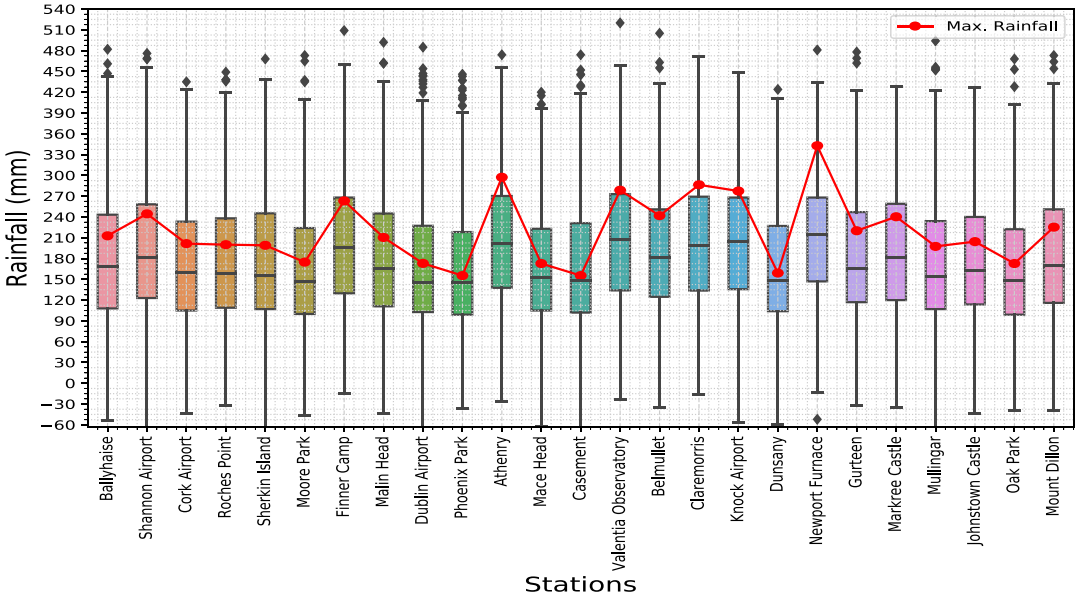

Figure 4: The maximum monthly rainfall amounts (measured in millimeter (mm)) for each station *vs* the predicted values represented by 95% prediction intervals for return period 50 years.

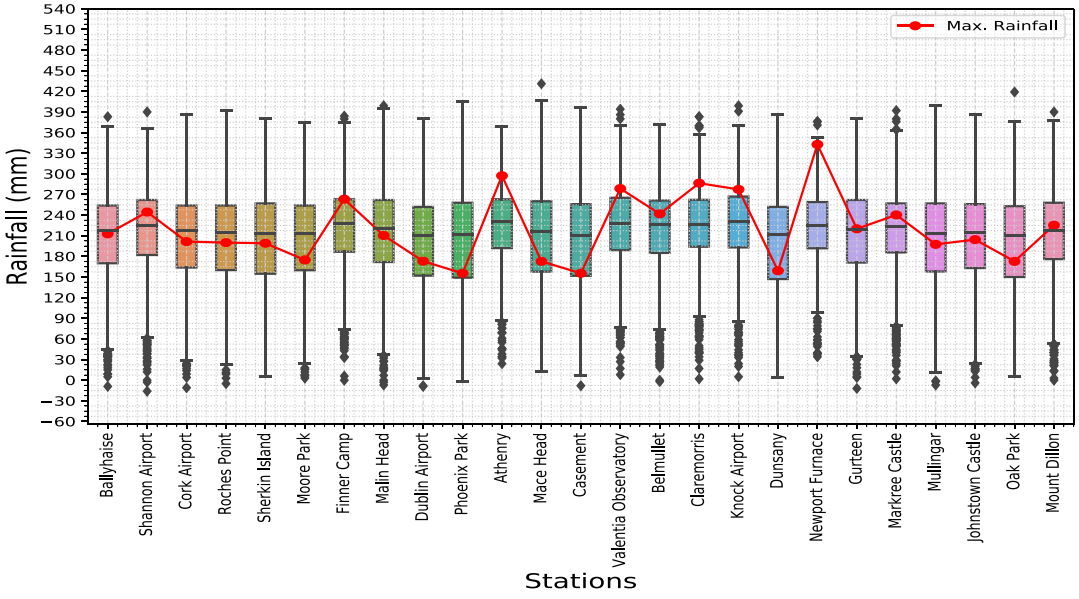

Figure 5: The maximum monthly rainfall amounts (measured in millimeter (mm)) for each station *vs* the predicted values represented by 95% prediction intervals for return period 100 years.

amounts at the risk over 100 years return period as shown in Figure 5. More specifically, the stations: Cork Airport (215 mm), Roches Point (214 mm), Sherkin Island (214 mm), Morre Park (213 mm), Malin Head (214 mm), Dublin Airport (209 mm), Phoenix Park (210 mm), Mace Head (212 mm), Casement (210 mm), Dunsany (210 mm), Mullingar (214 mm), Johnstown Castle (212 mm) and Oak Park (215 mm), show high rainfall amounts on the long term.

5 Conclusion

In this study, we used a finite mixture of Gaussian distributions to model the heterogeneity in monthly rainfall quantities in Ireland based on databases taken from 25 stations. Several tools were used to assess the Bayesian normal mixture models fitted to the data under study. The model selection criteria such as AIC, BIC, DIC, and WAIC were used to select the best model fit to the data. In

addition, we used graphical approaches such as the CDF and PDF to assess both the selected models. Most of the data reported by stations were modeled under a normal mixture model with two components, while the rest stations were modeled with only three components. We conclude that data that have been modeled by two components are more homogeneous than those modeled by three components. This article also shows the computation of return periods, for 50 and 100 years, for each station using the prediction intervals deriving from the posterior predictive distribution of the models selected by the model selection criteria. The diagnostic of high rainfall rates in the long term can help the related systems to put the plans that save lives and possessions. The advantage of the methodology used in this article is to reveal the heterogeneity in data by modeling it over several homogeneous groups. On the other hand, the disadvantage of this methodology is that it could not taken into account the hidden trend in increasing and decreasing in the maximum rates of rainfall. This latter can be modeled by so-called the hidden Markov mixture model which represents our future interest to study this phenomenon.

Acknowledgement: The authors would like to thank the engineering and science colleges as well as all anonymous reviewers in advance for their constructive comments.

Conflict of interest: Authors state no conflict of interest.

References

[1] Brutsaert W. Evaporation into the atmosphere: theory, history and applications. Vol. 1. Amsterdam: Springer Science and Business Media; 2013.
[2] Fletcher TD, Andrieu H, Hamel P. Understanding, management and modeling of urban hydrology and its consequences for receiving waters: a state of the art. Adv Water Res. 2013;51:261–79.
[3] Rojas R, Feyen L, Bianchi A, Dosio A. Assessment of future flood hazard in Europe using a large ensemble of bias-corrected regional climate simulations. J Geophys Res Atmospheres. 2012;117(D17):1–22.
[4] Alfieri L, Burek P, Feyen L, Forzieri G. Global warming increases the frequency of river floods in Europe. Hydrol Earth Sys Sci. 2015;19(5):2247–60.
[5] Tabari H. Climate change impact on flood and extreme precipitation increases with water availability. Scientific Reports. 2020;10(1):1–10.
[6] Wilks DS. Comparison of three-?parameter probability distributions for representing annual extreme and partial duration precipitation series. Water Resour Res. 1993;29(10):3543–9.
[7] Meneghini R, Jones JA, Iguchi T, Okamoto K, Kwiatkowski J. Statistical methods of estimating average rainfall over large space-timescales using data from the TRMM precipitation radar. J Appl Meteorol. 2001;40(3):568–85.
[8] Arvind G, Kumar PA, Karthi SG, Suribabu CR. Statistical analysis of 30 years rainfall data: a case study. IOP Confer Ser Earth Environ Sci. 2017;80(1):012067.
[9] Panda A, Sahu N. Trend analysis of seasonal rainfall and temperature pattern in Kalahandi, Bolangir and Koraput districts of Odisha, India. Atmospheric Sci Lett. 2019;20(10):e932.
[10] Nyatuame M, Owusu-Gyimah V, Ampiaw F. Statistical analysis of rainfall trend for Volta Region in Ghana. Int J Atmospheric Sci. 2014;2014:1–11.
[11] Asfaw A, Simane B, Hassen A, Bantider A. Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: a case study in Woleka sub-basin. Weather Climate Extremes. 2018;19:29–41.
[12] Praveen B, Talukdar S, Mahato S, Mondal J, Sharma P, Islam ARMT, et al., Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. Scientific Reports. 2020;10(1):1–21.
[13] Fruhwirth-Schnatter S, Celeux G, Robert CP. Handbook of mixture analysis. Boca Raton, FL: CRC Press, 2019.
[14] Parzen E, Tanabe K, Kitagawa G. Selected papers of Hirotugu Akaike. USA: Springer Science and Business Media; 2012.
[15] Schwarz G. Estimating the dimension of a model. Annals Statistics. 1978;6(2):461–4.
[16] Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc B (Statist Methodol). 2002;64(4):583–639.
[17] Watanabe S. A widely applicable Bayesian information criterion. J Mach Learn Res. 2013;14(Mar):867–97.
[18] Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995;82(4):711–32.
[19] Marin JM, Robert CP. Bayesian essentials with R. New York: Springer; 2014.
[20] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. Boca Raton, FL: CRC Press; 2013.
[21] Kadhem SK, Hewson P, Kaimi I. Using hidden Markov models to model spatial dependence in a network. Aust N Z J Stat. 2018;60(4):423–46.
[22] Celeux G, Forbes F, Robert CP, Titterington DM. Deviance information criteria for missing data models. Bayesian Anal. 2006;1(4):651–73.
[23] Leahy PG, Foley AM. Wind generation output during cold weather-driven electricity demand peaks in Ireland. Energy 2012;39(1):48–53.