AI in Civil Engineering

ORIGINAL ARTICLE

Open Access

# Generation of rainfall data series by using the Markov Chain model in three selected sites in the Kurdistan Region, Iraq

Evan Hajani[1*] and Gaheen Sarma[1]

**Abstract**

Rainfall forecasting can play a significant role in the planning and management of water resource systems. This study employs a Markov chain model to examine the patterns, distributions and forecast of annual maximum rainfall (AMR) data collected at three selected stations in the Kurdistan Region of Iraq using 32 years of 1990 to 2021 rainfall data. A stochastic process is used to formulate three states (i.e., decrease—"d"; stability—"s"; and increase—"i") in a given year for estimating quantitatively the probability of making a transition to any other one of the three states in the following year(s) and in the long run. In addition, the Markov model is also used to forecast the AMR data for the upcoming five years (i.e., 2022–2026). The results indicate that in the upcoming 5 years, the probability of the annual maximum rainfall becoming decreased is 44%, that becoming stable is 16%, and that becoming increased is 40%. Furthermore, it is shown that for the AMR data series, the probabilities will drop slowly from 0.433 to 0.409 in about 11 years, as indicated by the average data of the three stations. This study reveals that the Markov model can be used as an appropriate tool to forecast future rainfalls in such semi-arid areas as the Kurdistan Region of Iraq.

**Keywords** Time series, Rainfall, Markov chain, Forecast, Transition Probability

## 1 Introduction

Rainfall Prediction is one of the most difficult research topics globally (Oswal, 2019), but it is extremely important in water resource engineering for proper management of floods and droughts (Gao et al., 2020). Modeling and predicting rainfall is crucial for generating data as well as for providing information, which can then be used in a variety of applications, such as water resource management, hydrology, and agriculture (He et al., 2022). The prediction of rainfall and other climate conditions requires various models, depending on the time and spatial scales involved (Yusuf et al., 2014). For analyzing, simulating and forecasting hydrological variables, a variety of models, methods and techniques are found in literature review, e.g., Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Network (ANN), Nearest-Neighbors (NN), Fuzzy System, Numerical Weather Prediction model, and Holt's method (Holt, 1957). These methods, models and techniques can be selected typically based on the goals of research, the accessibility of input data, the quality of models, and certain predefined assumptions (Makridakis et al., 1998).

To guarantee a high degree of accuracy, researchers have evaluated the properties and characteristics of different models, so as to determine whether a certain model is appropriate for application in a given real-world situation. As a result, model selection becomes one of the major factors that influence the precision of the prediction data series. For example, Brath et al. (2002) used three models – ANN, ARIMA and NN – to enhance the prediction of the occurrence of floods caused by rainfall.

*Correspondence:
Evan Hajani
evan.hajani@uod.ac
[1] Water Resource Engineering Department, College of Engineering, University of Duhok, 38 Zakho Street, Kurdistan Region 1006 AJ, Duhok 42001, Iraq

With six hours' worth of rainfall data, it was found that ANN was the best at predicting rainfall. In a study by Kottegoda et al. (2004), a first-order Markov chain model successfully fitted the observed rainfall data in Italy; the model was built under the presumption that daily rainfall was dependent on the amount of rain that had fallen down the day before. Also using a Markov chain model, Barkotulla (2010) generated the daily rainfall occurrence based on transitional probability matrices; the model's parameters were acquired from historical daily rainfall records from 1980 to 2009. The results revealed that the model could successfully generate rainfall data.

Moreover, Liu et al. (2011) applied a Markov chain model to predict the daily rainfall series in 2004 based on the daily rainfall data collected in 2002 and 2003 in Tianjin, China. Because the predicted results met practical requirements, the model could be used as a weather generator to produce rainfall data for future periods. Chung et al. (2016) also used a Markov chain model for hourly rainfall data collected from 1985 to 2014 in Korea, finding that the rainfall occurrence was successfully fitted with the results of the Markov chain model. In addition, Fadhil et al. (2016) developed a stochastic rainfall generator model based on a first-order Markov chain model, by employing the data of a rainfall time series collected from 1976 to 2006 in Northwest Perak, Malaysia, with a two-state model (i.e., for dry and wet conditions) taken into account. In a word, first-order Markov models are generally deemed to be satisfactory, so it is justifiable to use such models to produce future rainfall series under various climate change scenarios.

Moreover, Gui and Shao (2017) and Zhou et al. (2017) applied certain Markov chain models to predict annual rainfall data series in Dangshan County and Shandong Province, China, respectively, finding that their proposed models had good prediction accuracy. Mahanta et al. (2018) applied a Markov chain model to the daily rainfall data collected at two stations (Dhaka and Chittagong) in Bangladesh. Examination of the behavior of the then-current daily rainfall data revealed that 56% of the days from May to October in Dhaka station are rainy, while 58% of the days in Chittagong are sunny. Malakoutian et al. (2021) used the rainfall data in six meteorological regions of North Cyprus from 1975 to 2014 to predict each meteorological region's yearly rainfall 5 years ahead. Three different models (i.e., Markov, ARIMA, and Holt-Winter) were adopted. The selected model for each region was then used to predict the rainfall for the five successive hydrologic years from 2014–2015 to 2018–2019.

Although most of the earlier rainfall forecasting studies have noted the effectiveness of the Markov model in forecasting rainfall, few studies have compared the outcomes of the prediction of Markov probability matrixes using various rainfall states with the outcomes of Markov chain models for forecasting rainfall in future periods. To address the gaps in previous research, this study sets the followings goals: (1) To provide additional insights into the changes in rainfall patterns; (2) To calculate the length of time needed for finding the steady-state probabilities in forecasting rains; (3) To predict and forecast rainfall in future periods; (4) To show how to use the first-order Markov chain model to create annual rainfall data for future times. This study makes the following research contributions: (1) A novel approach for creating prediction models is proposed using different rainfall states based on the Markov model; (2) The ability of the Markov model to predict and generate time series data is proved. The methodology used in this study can be applied to other regions of Iraq as well as to other nations.

## 2 Area and data of this study

Located in northern Iraq, and bordering Iran, Syria and Turkey, the Kurdistan Region comprises three main governorates: Duhok, Erbil and Sulaymaniya (Danilovich, 2016). With a Mediterranean climate (Hajani & Klari, 2022), the Kurdistan Region has hot and dry summers, but its temperature is mild in winter, with a very attractive spring season (WCG, 2019). Its average annual temperature is 32 °C, with about 71 mm of precipitation in a year (WCK, 2021). The daily average maximum temperature in Kurdistan goes up to 45 °C in the summer (in July) and the minimum goes down to 11 °C in January (Aziz et al., 2022). It is dry for 315 days across a year, with an average humidity of 26% and a UV index of 7 (Hajani et al., 2022). This study covers the historical daily rainfall data series for 32 years in the period of 1990–2021 at three rainfall stations in the Kurdistan region (see Fig. 1). The daily rainfall data are analyzed to identify the maximum precipitation in a year (i.e., 365 days), so as to use in the analysis of this study.

Table 1 shows the descriptive statistics of the three adopted stations, including their variance, standard deviation (Std. D.), coefficient of variation (C.V.), variance, coefficient of skewness (SK), and kurtosis. The mean AMR varies between 48.179 mm in the northern part of the study area (i.e., the Duhok station) and 62.998 mm in the southwestern part (i.e., the Sulaymaniya station). The minimum (Min) and maximum (Max) in Table 1 show that the lowest amount of rainfall (23.9 mm in 1996) occurred at the Erbil station and an extremely high amount occurred at the Duhok station (150 mm in 1993). The highest values of Std. D., C.V., Variance and Kurtosis are discovered at the Duhok station, indicating a significant variation in the rainfall data; and the highest Kurtosis value in the data set tends to have a clear peak close to the mean. The Sulaymaniya station has the highest

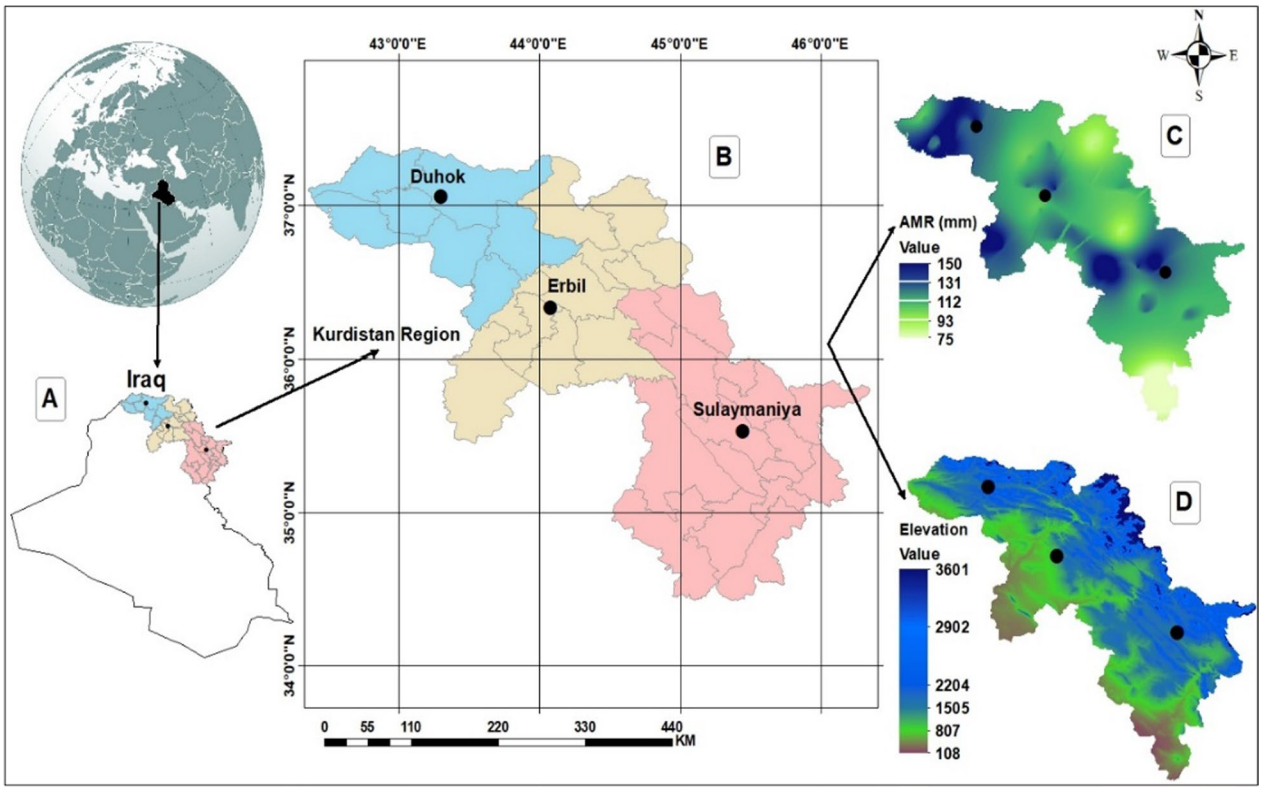


**Fig. 1** The location map of the study area

**Table 1** The statistical description of the Duhok, Erbil and Sulaymaniya stations

| Station | Mean (mm) | Min. (mm) | Max. (mm) | Std. D. (mm) | C.V | Variance ($mm^2$) | SK | Kurtosis |
|---|---|---|---|---|---|---|---|---|
| Duhok | 59.622 | 24.200 | 150.000 | 26.822 | 44.987 | 719.435 | 1.511 | 3.072 |
| Erbil | 48.179 | 23.900 | 103.900 | 17.212 | 35.726 | 296.269 | 1.183 | 1.982 |
| Sulaymaniya | 62.998 | 36.800 | 131.800 | 23.306 | 36.995 | 543.153 | 1.663 | 2.901 |

value of SK, which clearly shows that this station is highly skewed, and its asymmetric tail also extends to the right of the mean value.

## 3 Methodologies

### 3.1 Markov probability model

Markov analysis is a scientific method for studying and analyzing a phenomenon in the current period, so as to predict its behavior in the future (Kenton, 2021). The Markov process is a type of stochastic process that can, in essence, predict a random variable based solely on the current circumstances surrounding the variable, without the need to know the past (Rykov et al., 2010). The Markov chain contributes to situation forecasting by identifying a phenomenon from one period to the next with a Markov Matrix, which is known as Transition Probabilities Matrix (TPM) from the prior case (Jimoh & Webster, 1996). According to Eq. (1), the conditional prediction of any future state ($X_{n+1}$) by using a Markov chain model is independent of the past state ($X_0$, …, $X_{n-1}$), but depend only on the present state ($X_n$) (Yusuf et al., 2014).

$$P\left(X_{n+1}=j \mid X_n=i, X_{n-1}=i_{n-1}, \ldots, X_1=i_1, X_0=i_0\right)=P\left(X_{n+1}=j \mid X_n=i\right)=P_{ij}. \quad (1)$$

In this study, the first-order Markov chain model was used. The three states adopted include: decrease—"d"; stability—"s"; and increase—"i". The flowchart

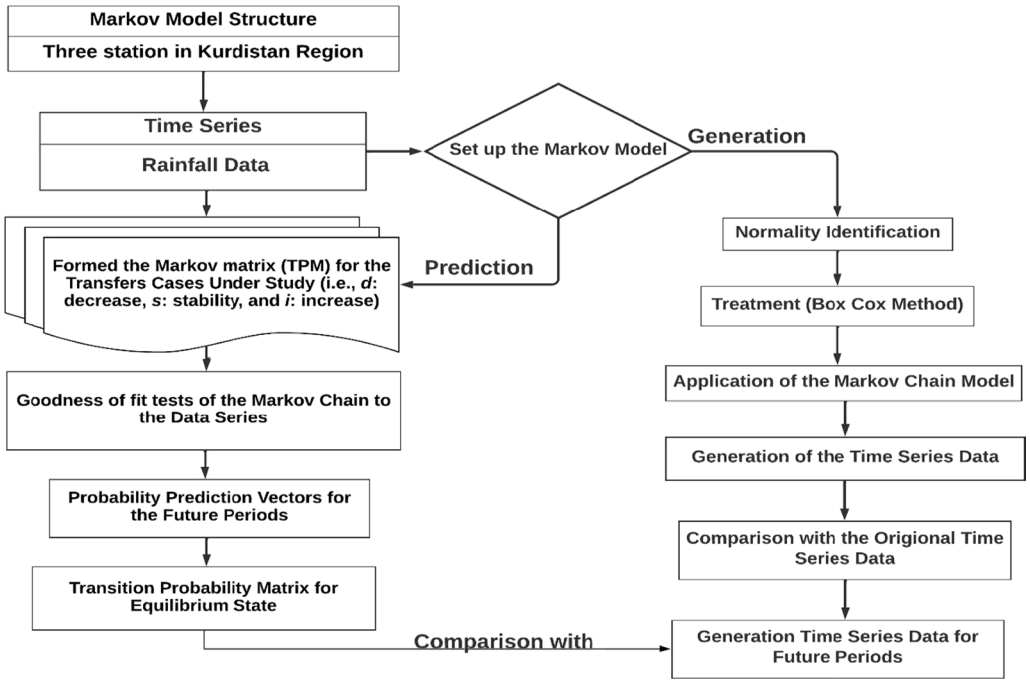


**Fig. 2** The procedure of applying the Markov model

of Markov modeling adopted in this study is given in Fig. 2. The transition probability matrix (Markov matrix) is defined as $P_{ij} \equiv P(j \mid i)$, where $i, j \in$ the states. The Markov matrix is formed through observing the time series and identifying the number of transitions from one state to another for the three states under study ("d", "s", and "i"). In this study, depending on the AMR data for the three adopted stations, the transition forms of the Markov matrix are modeled as below:

1. State decrease—"d": the AMR data are below 50 mm.
2. State stability—"s": the AMR data are between 50 mm and 60 mm.
3. State increase—"i": the AMR data are above 60 mm.

The transition between the states is described by the transition diagram and probability matrix ($P$) below:

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}.$$

The Maximum Likelihood Estimator (MLE) is used for estimating the transition probabilities in the Markov matrix (Jale et al., 2019; Zhang et al., 2014). The sum of the probabilities of the $P$ matrix in each row must equal one. The MLE process involves dividing the elements in a row of the $P$ matrix by the sum of all the elements in that row, so as to estimate each element in the Markov matrix.

### 3.2 Test for goodness of fit

In order to predict the AMR data series by using the Markov chain model, the validity of the proposed three states (i.e., decrease, stability and increase of rainfall) used in the Markov chain approach was tested in the following way: The null hypothesis $H_0$: rainfall occurrences in consecutive years are independent; vs. The alternative hypothesis $H_1$: rainfall occurrences in consecutive years are not independent (Garg & Singh, 2010). For the three suggested states in the Markov chain, two tests for goodness of fit, namely the WS test and the Chi-Squared ($\chi^2$) test (Wang & Martiz, 1990; Preacher, 2001; Jale et al., 2019), were made, as given in Eqs. (2) and (3) below:

$$\mathrm{WS} = \frac{A + B - 1}{\sqrt{V(A + B - 1)}} \quad N(0,1), \tag{2}$$

where $A = P_{dd} + P_{ss} + P_{ii}$; $B = P_{id}P_{di} + P_{si}P_{is} + P_{ds}P_{sd} - P_{dd}P_{ss} - P_{dd}P_{ii} - P_{ss}P_{ii}$. The variance (V) of $(A + B - 1)$ in Eq. (1) is given by $V(A + B - 1) = 2p_1p_2p_3\left(\frac{1}{n_dn_s} + \frac{1}{n_sn_i} + \frac{1}{n_in_d}\right)$, where $n_d, n_s$ and $, n_r$ are the numbers of states used in this study (i.e., "d", "s", and "i" states), while $p_1, p_2$ and $p_3$ indicate the stationary probabilities, specifically: $p_1 = [(1 + p) + \frac{(1+s)p}{q}]^{-1}$; $p_2 = \left[r + \frac{ps}{q}\right]p_1$; $p_3 = [p/q]p_1$.

$p = \left[P_{di} + \frac{P_{si}(1-P_{dd})}{P_{sd}}\right]\left(\frac{1}{1-P_{ii}}\right)$; $q = 1 + \left[\frac{P_{si}P_{id}}{P_{sd}(1-P_{ii})}\right]$; $r = \left(\frac{P_{ds}}{1-P_{ss}}\right)$; $q = 1 + \left[\frac{P_{si}P_{id}}{P_{sd}(1-P_{ii})}\right]$; $r = \left(\frac{P_{ds}}{1-P_{ss}}\right)$; $s = \left(\frac{P_{is}}{1-P_{ss}}\right)$. The probability ($P$) transforming from a state of "d" into another state of "d" is represented by$P_{dd}$; that from a state of "s" into another "s" is represented by$P_{ss}$; that from a state of "i" into another "i" is represented by$P_{ii}$; that from a state of "d" into a state of "i" is represented by$P_{di}$; that from a state of "i" into a state of "d" is represented by$P_{id}$; that from a state of "s" into a state of "i" is represented by$P_{si}$; that from a state of "i" into a state of "s" is represented by$P_{is}$; that from a state of "d" into a state of "s" is represented by$P_{ds}$; and that from a state of "s" into a state of "d" is represented by$P_{sd}$.$P_{dd}$, $P_{ss}$, $P_{ii}$, $P_{di}$ ,$P_{id}$, $P_{si}$, $P_{is}$, $P_{ds}$ and $P_{sd}$ are the elements of the TPM.

According to the test procedure, the critical region is $|\mathrm{WS}|_c \geq Z_\alpha$ at the $\alpha$ level of significance, i.e., the null hypothesis is rejected if $|\mathrm{WS}| \geq Z_\alpha$, where $Z_\alpha$ indicates the $100(1-\alpha)$ lower percentage points of a standard normal distribution (Garg & Singh, 2010).

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - E_i)^2}{E_i}, \tag{3}$$

where $X_i$ is the observed frequency of the data sample, $E_i$ is the expected frequency of the data sample as calculated by $E_i = F(X_2) - F(X_1)$, and $F$ is the cumulative distribution function of the probability distribution being tested. The test was conducted at a 5% significant level.

### 3.3 The initial state vector and the *n*-step transition state vectors

The initial state vector ($\pi^0$) of the Markov chain is estimated from the sum elements of the rows of the TPM divided by the sum of all the elements of the TPM (Howard, 1971; Jain, 1986) as follows:

$$\pi^0 = \frac{\sum_{i=1}^{n} P_i}{\sum P}. \tag{4}$$

In this study, the n-step transition probability state vectors ($\pi^n$) of the Markov chain are following Cox and Miller (1984) and Yusuf et al. (2014). In the equation above, $P_i^n$ is the probability that the annual maximum rainfall is in the $i$th state at the $n$th observation. In particular, $\pi^{\mathrm{n}}$ is the state vector of the Markov chain at the $n$th year (Jain, 1986; Yusuf et al., 2014), and it can be estimated by using the following equation:

$$\pi^{(n+1)} = \pi^{(n)} P, \tag{5}$$

where $P$ is the TPM and $\pi^{n+1}$ is the state vector at the $(n+1)$th data observation. After a different number of iterations, the n-step state vectors will be estimated as:

$$\pi^n = \pi^0 P^n. \tag{6}$$

### 3.4 Equilibrium probabilities

The equilibrium probabilities $\pi_{d,}\pi_s$ and $\pi_i$ (of a "d", "s" and "i", respectively) are determined by resolving the stationary matrix equation (Garg & Singh, 2010; Jale et al., 2019). As a result, there is no change in the probability of being in any state over time, i.e., there is a limit of $\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j > 0$, where $\pi_j$ satisfies only the following stable state equation: $\pi_j = \sum_{i=1}^{m} \pi_i P_{ij}$ to $j = 1, \ldots, m$. Therefore, the linear system of equations (i.e., $\pi_d = \pi_d P_{dd} + \pi_s P_{sd} + \pi_i P_{id}, \pi_s = \pi_d P_{ds} + \pi_s P_{ss} + \pi_i P_{is}$ and $\pi_i = \pi_d P_{di} + \pi_s P_{si} + \pi_i P_{ii}$) can be solved to obtain the estimators of the long-term equilibrium probabilities, together with the probability normalization condition $\pi_d + \pi_s + \pi_i = 1$.

## 4 Using a Markov model in forecasting

The formula of a Markov model for generating a data series is based on the following procedures (Gupta, 1989).

i. Transform data to the normal distribution

Before starting the steps of building a random number generation model, it must be first confirmed that the time series used in the model is subject to the normal distribution. In this study, Anderson–Darling (AD) and Lilliefors (LT) tests for normality were applied (Anderson & Darling, 1954; Abdi & Molin, 2007). If the normal distribution cannot be achieved in a data series, the Box-Cox method will be applied for transformation (Box & Cox, 1964; Sakia, 1992). The forms of the Box-Cox transformation are given by:

$$X(\lambda) = \frac{\left(X^\lambda - 1\right)}{\lambda} \quad \text{for } \lambda \neq 0. \tag{7}$$

$$X(\lambda) = \ln(X) \quad \text{for } \lambda = 0. \tag{8}$$

The lambda (λ) value ranges from − 5 to 5, and the best λ value for the data is chosen after taking all possible values into account. The optimal value of λ appears where a normal distribution curve is most closely approximated (Chedded, 2020).

ii. Make descriptive statistics

Three important parameters were used during the analysis of the Markov model: the mean, standard devi-

ation, and correlation coefficient, as determined by the following equations:

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{9}$$

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)^2}, \tag{10}$$

$$r_k = \frac{n\sum_{i=1}^{n-k}(X_i - \mu_X)(X_{i+k} - \mu_X)}{(n-k)\sigma_X^2}; \quad k = 1. \tag{11}$$

iii. Generate random numbers

Random numbers ($t$) were generated in Microsoft Excel (Kuo, 2016; Maass et al., 1962) with the RAND () command. After a random variable $t$ with an arithmetic mean of zero and unit variance was acquired, i.e., $N$ (0, 1), it was then converted to follow the standard normal distribution according to the following equation (inverse error function).

$$\mathrm{erf}^{-1}(z) = \frac{1}{2}\sqrt{\pi}\left(z + \frac{\pi}{12}z^3 + \frac{7\pi^2}{480}z^5 + \frac{127\pi^3}{40320}z^7 + \frac{4369\pi^4}{5806080}z^9 + \cdots\right), \tag{12}$$

where $z$ value can be found through the cumulative function distribution (CDF) function of the normal logarithm, as follows:

$$\mathrm{CDF} = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left|\frac{\ln x - \mu}{\sqrt{2\sigma^2}}\right| = \mathrm{RAND}(), \tag{13}$$

$$\mathrm{erf}(z) = \mathrm{erf}\left|\frac{\ln x - \mu}{\sqrt{2\sigma^2}}\right| = [\mathrm{RAND}() - 0.5] * 2, \tag{14}$$

$$z = 2[\mathrm{RAND}() - 0.5]. \tag{15}$$

As the normal logarithm of random numbers has properties of μ = 1 and σ = 1, therefore:

$$\left|\frac{\ln x - 1}{\sqrt{2}}\right| = z \rightarrow \mathrm{erf}^{-1}(z) = \frac{\ln x - 1}{\sqrt{2}} \rightarrow \sqrt{2}\mathrm{erf}^{-1}(z) = \ln x - 1$$

Then the random number formula is:

$$t = \ln x = \sqrt{2}\mathrm{erf}^{-1}(z) + 1. \tag{16}$$

iv. Build the Markov model for the generation

The Markov model's generalized form, as developed by Thomas and Fiering (Bin Muhammad, 2012), is represented by the following equation:

$$X_{i+1} = \mu_X + r_i(X_i - \mu_X) + \sigma_x t_i\sqrt{(1 - r_i^2)}, \tag{17}$$

where $X_i$ is the rainfall value in time $i$, $\mu_X$ is the mean value of the data series, $r_i$ is the correlation coefficient, $\sigma_x$ is the standard normal deviation, and $t_i$ is the random number.

### 4.1 Measurements of the reliability of the forecasting

To assess the forecasting model in this study, the relative error (RE), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) were measured. Additionally, the Willmott (1981) index of agreement ($D$) was measured, which can be interpreted as the relative prediction error between the actual data and the fitted data from the Markov chain model. $D = 1$ indicates a perfect match, while $D = 0$ indicates no agreement at all. The errors mentioned above are expressed in the following equations:

$$\mathrm{RE} = \sum_{t=1}^{n}\frac{X_t - \hat{X}_t}{X_t}, \tag{18}$$

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \hat{X}_t\right)^2}, \tag{19}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|X_t - \hat{X}_t\right|, \tag{20}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|X_t - \hat{X}_t\right|}{X_t} * 100, \tag{21}$$

$$D = 1 - \frac{\sum_{i=1}^{n}\left(X_t - \hat{X}_t\right)^2}{\sum_{i=1}^{n}\left(\left|\hat{X}_t - X\right| + \left|\bar{X}_t - \bar{X}\right|\right)^2}, \quad 0 \le D \le 1, \tag{22}$$

where $X_t$ is the actual AMR value (mm), $\hat{X}_t$ is the estimated (forecasted) AMR value (mm), and $\bar{X}$ is the average actual AMR (mm).

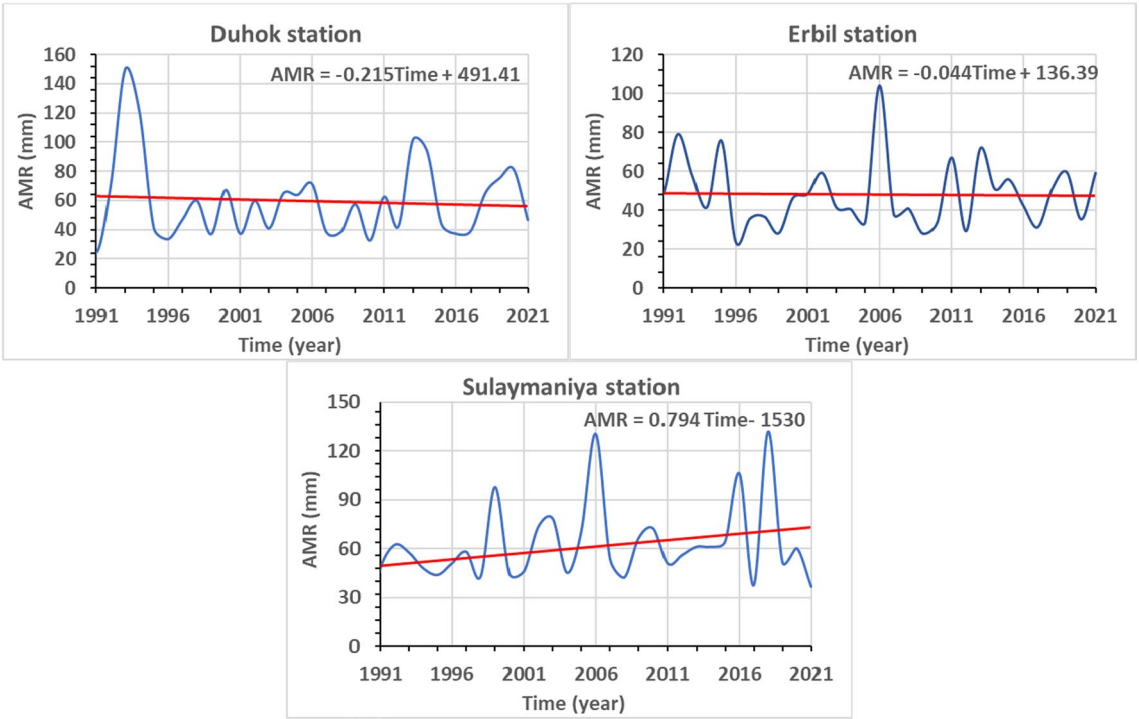


**Fig. 3** The AMR data series for the three adopted stations

## 5 Results and discussion

### 5.1 The AMR data series

The plots for the time series of the AMR data at the three adopted stations during the period of 1990–2021 are presented in Fig. 3. As can be seen, the trend line moves downward at a rate of 0.215 mm/year at the Duhok station and at a rate of 0.044 mm/year at the Erbil station; for the Sulaymaniya station, however, the trend line moves upward at a rate of 0.794 mm/year, although the trend slopes of decreasing and increasing are quite small in magnitude.

### 5.2 The transition probability matrix (TPM)

The data of AMR time series were used in this study to form the Markov matrix (TPM), with three states taken into account (i.e., "d", "s" and "i"). For these states, the TPM was sorted by the number of transitions from one state to another. To obtain the TPM, the elements of each row of the TPM were divided by the sum of the rows. The number of transfers in the states being examined is shown in Table 2. The Initial State Vector ($\pi^0$) was estimated for each station based on the sum of the rows in Table 2. The TPM for the three adopted stations is shown

**Table 2** The transition count and initial state vector for the AMR behavior between "d", "s" and "i" states

| **Transition count** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Duhok** | | | **Erbil** | | | **Sulaymaniya** | | |
| **"d" to "d"** | **"d" to "s"** | **"d" to "i"** | **"d" to "d"** | **"d" to "s"** | **"d" to "i"** | **"d" to "d"** | **"d" to "s"** | **"d" to "i"** |
| 5 | 2 | 8 | 5 | 1 | 8 | 3 | 1 | 8 |
| **"s" to "d"** | **"s" to "s"** | **"s" to "i"** | **"s" to "d"** | **"s" to "s"** | **"s" to "i"** | **"s" to "d"** | **"s" to "s"** | **"s" to "i"** |
| 2 | 2 | 0 | 1 | 1 | 3 | 1 | 2 | 1 |
| **"i" to "d"** | **"i" to "s"** | **"i" to "i"** | **"i" to "d"** | **"i" to "s"** | **"i" to "i"** | **"i" to "d"** | **"i" to "s"** | **"i" to "i"** |
| 8 | 1 | 4 | 8 | 2 | 3 | 9 | 2 | 5 |
| Initial state vector ($\pi^0$) | | | | | | | | |
| [ 0.469 0.125 0.406 ] | | | [ 0.438 0.156 0.406 ] | | | [ 0.375 0.125 0.500 ] | | |

**Table 3** The TPM (*P*) of the AMR behavior for the three adopted stations

| Stations | Duhok | Erbil | Sulaymaniya |
|---|---|---|---|
| *P* | $\begin{bmatrix} 0.333 & 0.133 & 0.534 \\ 0.500 & 0.500 & 0 \\ 0.615 & 0.077 & 0.308 \end{bmatrix}$ | $\begin{bmatrix} 0.357 & 0.071 & 0.572 \\ 0.200 & 0.200 & 0.600 \\ 0.615 & 0.154 & 0.231 \end{bmatrix}$ | $\begin{bmatrix} 0.250 & 0.083 & 0.667 \\ 0.250 & 0.500 & 0.250 \\ 0.563 & 0.125 & 0.312 \end{bmatrix}$ |

in Table 3. The state-transition diagrams of the system for the three stations are shown in Fig. 4. The results of the TPM for the three adopted stations show that the probabilities of transitions from the state "i" to "d" and from the sate "d" to "i" are higher than for other states at the Duhok and Erbil stations; but at the Sulaymaniya station, the probability of transition from the state "i" to "d" is higher than for other states. The lowest value of the probability of transition is found for transition from the state "s" to "i" at the Duhok station.

### 5.3 The results of goodness-of-fit tests

The statistics of WS and $\chi^2$ tests (as described in Sect. 3) were used to measure the goodness of fit of the Markov chain to the AMR data for the three adopted stations (Duhok, Erbil, and Sulaymaniya). The estimated values of the WS and $\chi^2$ statistics as well as their p-values are presented in Table 4. As seen in Table 4, the results of these two statistics indicate that the Markov chain is a model appropriate to the AMR data series for the three adopted stations, and it is proven that the both tests satisfy a major property of the Markov chain model.

**Table 4** Estimated values of the WS and $\chi^2$ statistics and their *p*-values

| Stations | Duhok | Erbil | Sulaymaniya |
|---|---|---|---|
| WS | 3.141 ($p<0.001$) | 4.280 ($p<0.001$) | 5.914 ($p<0.001$) |
| $\chi^2$ | 0.889 ($p=0.828$) | 1.126 ($p=0.890$) | 1.100 ($p=0.777$) |
| Reject? | No | No | No |

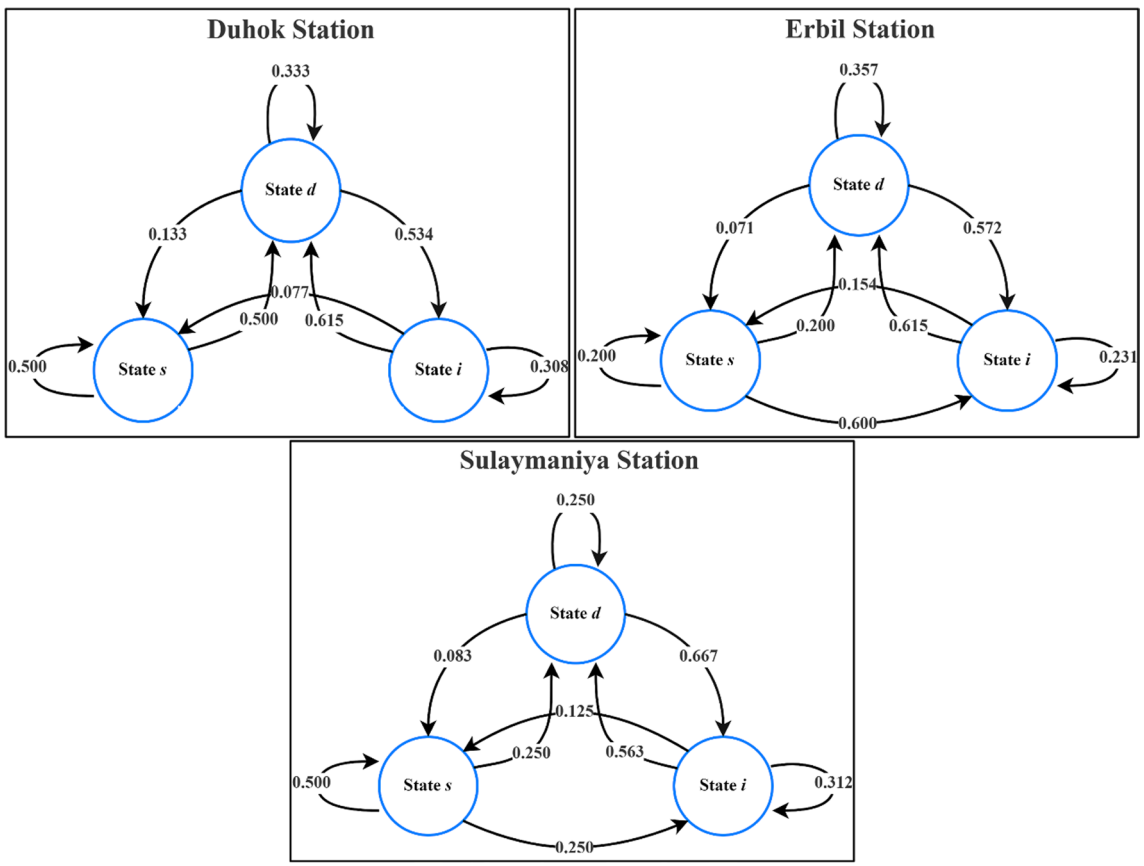


**Fig. 4** Diagrams of state transitions for the Duhok, Erbil and Sulaymaniya stations

### 5.4 The transition probability for future periods

Depending on the initial state vector shown in Table 2, the average probability of moving from all transition count states to the state "d" for all the three stations is 0.427; that for the state "s" is 0.135; and that for the state "i" is 0.438. The probability prediction of all the states under study for a future period (e.g., 5 years) was estimated based on the initial state vector and Markov matrix, as described in Sect. 3 (i.e., $\pi^1 = \pi^0 * P$, $\pi^2 = \pi^1 * P$, ..., $\pi^5 = \pi^4 * P$). The result is shown in Table 5.

### 5.5 The transition probability matrix for the *n*-steps (equilibrium state)

The equilibrium and stability state can be reached based on Formula (4) in iteration, as described in Sect. 3. In this study, MATLAB software was used to obtain accurate results of n-steps of the TPM. It should be noted that the TPM was corrected to 3 decimal places. As observed in Additional file 1: Table S1, when $n$ increases, the results will become more similar. According to the results in Additional file 1: Table S1, the TPM will reach an equilibrium and stability state in 12 years for the Duhok station, in 11 years for the Erbil station, and in 9 years for the Sulaymaniya station. These probability matrices continue to infinity (if the current climate conditions are kept). After the process reaches a steady state as mentioned before, the limiting state vectors (by using Eq. (5)) are: for the Duhok station: $\pi_{\text{Duhok}}^{(n\geq 12)} = [\,0.464 \;\; 0.178 \;\; 0358\,]$; for the Erbil station: $\pi_{\text{Erbil}}^{(n\geq 11)} = [\,0.448 \;\; 0.123 \;\; 0.429]$; and for the Sulaymaniya station: $\pi_{\text{Sulaymaniya}}^{(n\geq 9)} = [\,0.387 \;\; 0.174 \;\; 0.439\,]$. Therefore, the results of limiting state vectors show that the probabilities would drop slowly from 0.433 to 0.409 in about 11 years on average for the three adopted stations.

### 5.6 Steps of applying the Markov model in forecasting the AMR data

i. The normality distribution of the data

To find out whether the AMR data series follows a normal distribution or not, the normality tests (i.e., AD and LN) was conducted, as mentioned in Sect. 3, with the results demonstrated in Table 6. Whereas the null hypothesis for the tests is that the data were normally distributed, the alternate hypothesis is that the data did not come from a normal distribution. The $p$ values of the both tests in Table 6 are less than 0.05, so the tests reject the hypothesis of normality. Hence, it can be concluded that the AMR data series for three adopted stations do not follow a normal distribution. The Box-Cox transformation method (Sect. 3) was used to transform non-normal AMR data, so as to meet a normal distribution, as shown in Table 7.

ii. The statistical measures and features of the Markov model

**Table 6** Estimated values of the normality tests of AMR data series

| Stations | Duhok | Erbil | Sulaymaniya |
|---|---|---|---|
| AD statistic (*p*-values) | 1.330 (0.002) | 0.740 (0.041) | 2.001 (< 0.001) |
| LT statistic (*p*-values) | 1.103 (0.008) | 0.732 (0.028) | 1.108 (0.003) |
| Reject? | Yes | Yes | Yes |

**Table 7** The transformed function depends on the Lambda value (λ)

| Station | λ: Transformed function |
|---|---|
| Duhok | 0: Log (AMR) |
| Erbil | − 0.5: 1/Sqrt (AMR) |
| Sulaymaniya | − 1: 1/(AMR) |

**Table 8** Descriptive statistical measures of the transformed AMR data series

| Stations | Duhok | Erbil | Sulaymaniya |
|---|---|---|---|
| Min | 1.384 | 0.098 | 0.008 |
| Max | 2.176 | 0.208 | 0.027 |
| Mean | 1.738 | 0.151 | 0.018 |
| SD | 0.177 | 0.025 | 0.005 |
| r | 0.135 | − 0.148 | − 0.292 |

**Table 5** The probability prediction vectors for a certain future period in the Duhok, Erbil, and Sulaymaniya stations

| Period | $\pi$ | Duhok | Erbil | Sulaymaniya |
|---|---|---|---|---|
| 2022 | $\pi^1$ | [ 0.468 0.156 0.376 ] | [ 0.437 0.125 0.438 ] | [ 0.407 0.156 0.437 ] |
| 2023 | $\pi^2$ | [ 0.465 0.169 0. 366] | [ 0.450 0.123 0.427 ] | [ 0.387 0.166 0.447 ] |
| 2024 | $\pi^3$ | [ 0.464 0.175 0.361 ] | [ 0.448 0.122 0.430 ] | [ 0.390 0.171 0. 439] |
| 2025 | $\pi^4$ | [ 0.464 0.177 0.359 ] | [ 0.449 0.123 0.428 ] | [ 0.387 0.173 0.440 ] |
| 2026 | $\pi^5$ | [ 0.463 0.178 0.359 ] | [ 0.448 0.122 0.429 ] | [ 0.388 0.174 0.438 ] |

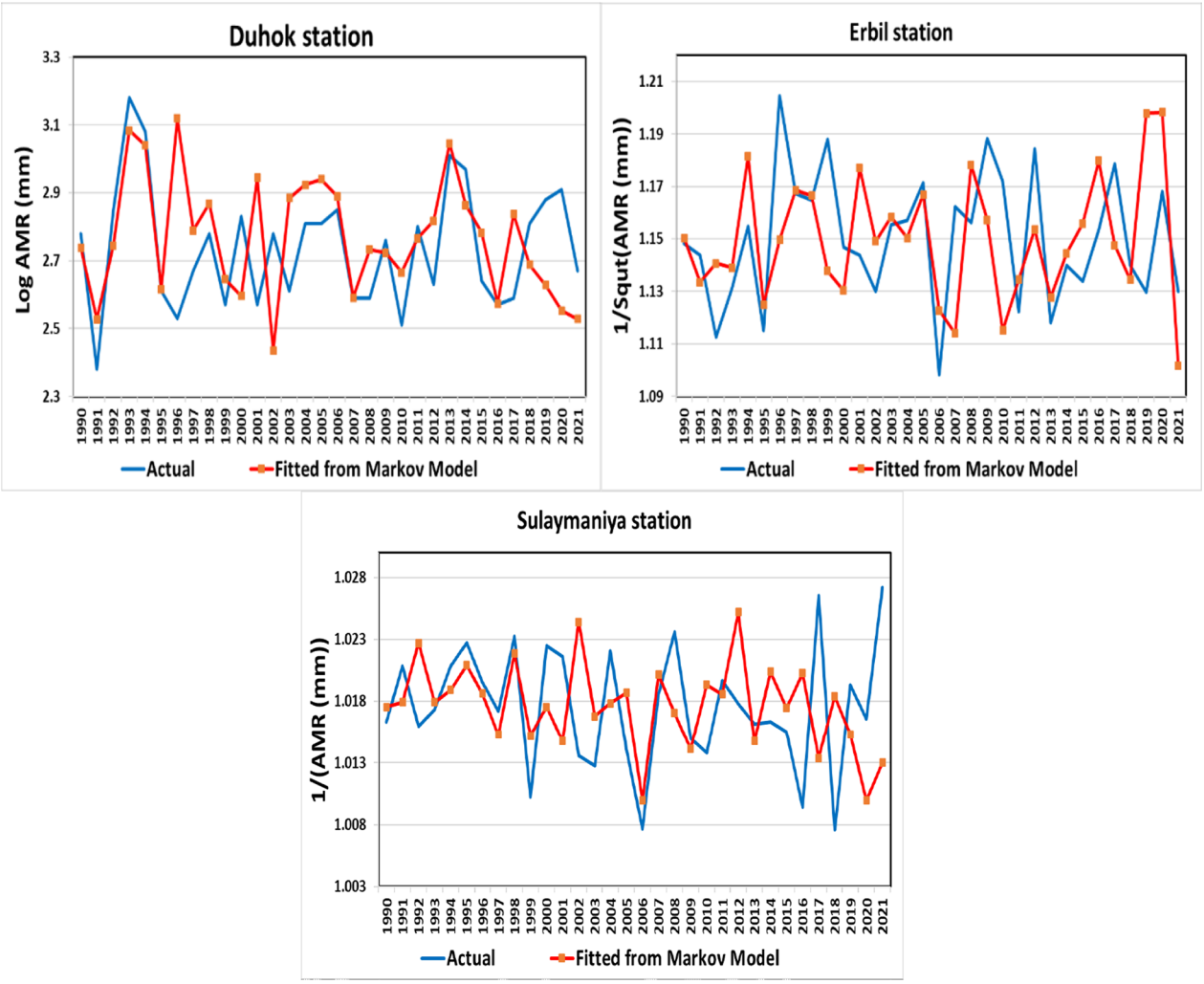


**Fig. 5** Time series data was generated by using the Markov model for the Duhok, Erbil and Sulaymaniya stations. The actual data are given in blue color, and the fitted values are in red line color

Table 8 presents the summary of statistical results of the transformed AMR data for the three adopted stations. These statistical results are essential for generating the AMR data in the Markov model, as described in Sect. 3.

iii. Generation of random numbers

At this stage, random numbers were generated in Microsoft Excel through its RAND () command, as mentioned in Sect. 3. For example, the procedures for generating random numbers for the years 1990–2000 are shown in Additional file 1: Table S2.

iv. The application of the Markov Model for generating the AMR data

The Markov model consists of two parts: the first part is deterministic and considers the effect of the previous statistical values on the model (in Subsection ii); and the second part is for random numbers and represents the random part of the model (in Subsection iii). By combining and adding these two parts, a Markov generation model is applied according to Eq. (17) in Sect. 3. It should be noted that for the purposes of using Eq. (17), the first-generation value was assumed to be the mean (μ) of the data series. The plots between the observed and predicted values in Fig. 5 for the Markov model on the transformed AMR data series for the three adopted stations indicates that the predicted values follow the observed data closely enough.

In addition, five other statistical tests (i.e., RE, MAE, MAPE, RMSE, and D) were conducted to evaluate the closeness between the actual AMR data and the fitted AMR data in the Markov chain model, as shown in Table 9. The lower values of RE, MAE, MSE and RMSE

**Table 9** The performance of the Markov model

| Stations | RE (%) | MAE (mm) | MAPE (%) | RMSE (mm) | *D* |
|---|---|---|---|---|---|
| Duhok | 2.838 | 0.159 | 2.406 | 0.116 | 0.834 |
| Erbil | 1.513 | 0.044 | 0.247 | 0.064 | 0.829 |
| Sulaymaniya | 1.319 | 0.024 | 0.137 | 0.007 | 0.871 |

**Table 10** The AMR time series (mm) forecasted by using the Markov model for the period from 2022 to 2026

| Stations | 2022 | 2023 | 2024 | 2025 | 2026 |
|---|---|---|---|---|---|
| Duhok | 53.424 | 36.355 | 45.193 | 49.405 | 42.105 |
| Erbil | 42.011 | 50.267 | 36.268 | 42.149 | 38.354 |
| Sulaymaniya | 65.462 | 64.911 | 55.798 | 54.569 | 68.231 |

indicate a higher accuracy of the Markov chain model for each adopted station. The value of $D$ is sensitive to the differences between the observed data and the generated data, and is also sensitive to certain changes in proportionality (Willmott et al., 1981). The results in Table 9 show that the $D$ value between the observed data and the generated AMR from the Markov model for the three stations are greater than 80%, indicating that in the three adopted stations, the magnitudes of $D$ were consistent with the independent and intuitive evaluations.

With a $D$ value of 0.871, the fitted value from the Markov model for the Sulaymaniya station is identified as a slightly "better" estimator of the observed variables than for the Duhok ($D=0.834$) and Erbil ($D=0.829$) stations. The results of the Willmott index ($D$) in Table 9 show a good match between the actual data and the generated AMR from the Markov chain model. Therefore, the generated AMR is acceptable at large percentages for the three adopted stations. The actual AMR data and the fitted AMR data by the Markov chain model are plotted in a scatterplot, as shown in Additional file 1: Fig. S1. A visual examination of the AMR time series scatterplot (Additional file 1: Fig. S1) confirms the results of the Willmott index.

The Markov chain model was used to generate the AMR data for a future period from 2022 to 2026 (five years), as shown in Table 10. It is found that, in general, the results of the generated AMR data for the future period confirm the future probability prediction vectors for the three stations as made in Sub Sect. 4.4. For example, for the Duhok station, it is found that the AMR data (Table 10) would decrease for the future period. This consists of the probability prediction vectors for the Duhok station (Table 5, and Sub Sect. 4.4), showcasing a decrease in the probability for the state "d" for most of the adopted future period more than for the "s" and "i" states.

## 6 Conclusions

In this study, a Markov chain model was adopted to examine the pattern, distribution and forecast of the annual maximum rainfall (AMR) data at three selected stations (Duhok, Erbil, and Sulaymaniya) in the Kurdistan region of Iraq based on the rainfall data collected there within 32 years of 1990–2021. A Markov matrix (TPM) was formed based on the AMR time series data and sorted by the number of transitions from one state to another among the three states under study (i.e., decrease—"d"; stability—"s"; and increase—"i"). The Markov model was used to forecast the AMR data for several upcoming years (i.e., 2022–2026). To evaluate how well the Markov chain model fits the data, the Chi-square and WS tests were conducted. Additionally, the Lilliefors and Anderson–Darling tests for normality were used, while the Box-Cox transformation method was used to transform the non-normal AMR data to meet a normal distribution. To assess the Markov chain generation model, the following tests were conducted: the relative error (RE), the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the Willmott index. The results of two goodness-of-fit tests (i.e., WS and $\chi^2$) indicate that the Markov chain is an appropriate model for the AMR data series from the three adopted stations. As demonstrated, in the years of 2022–2026, the probability of the annual rainfall data decreasing will be 44%, with 16% of the annual rainfall keeping stable and 40% of the annual rainfall increasing. The TPM will reach an equilibrium and stability state after 12 years at the Duhok station, after 11 years at the Erbil station, and after 9 years at the Sulaymaniya station. In addition, it is shown that the probabilities will drop slowly from 0.433 to 0.409 for the AMR data series in about 11 years as an average for the three adopted stations. The predicted AMR data from the Markov model were compared with the observed AMR data, so as to determine the prediction precision, revealing that the RE, RMSE, MAE and MAPE test results between the observed data and the predicted data are less than 4%, which satisfies the criteria for forecast accuracy. In addition, the Willmott index shows a good match between the actual data and the generated AMR data from the Markov chain model. Therefore, the upcoming AMR data can be forecasted in this Markov model. The results of the generated AMR data for future periods were found able to confirm the probability prediction vectors for future periods at the three stations.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s43503-023-00014-2.

**Additional file 1: Table S1.** The n-step of TPM for the Duhok, Erbil, and Sulaymaniya stations. **Table S2.** Generating random numbers for 10 years (1990–2000). **Figure S1.** Scatterplots of actual vs. generated AMR (Fitted from Markov Model) for three adopt stations (1990–2000).

**Acknowledgements**
The authors thank the Directorate of Meteorology and Seismology in Erbil and Duhok governorates in the Kurdistan region, Iraq for providing rainfall data that was used in this study.

**Author contributions**
EH designed and prepared the paper and contributed to the analysis and writing of this manuscript. GS contributed to the analysis and writing of the manuscript. Both authors read and approved the final manuscript.

**Funding**
This research is not supported by any funding agency.

**Availability of data and materials**
The rainfall data can be obtained from the Directorate of Meteorology and Seismology in Erbil and Duhok governorates in the Kurdistan region, Iraq All data, models, or code generated or used during the study are proprietary or confidential in nature and may only be provided with restrictions.

## Declarations

**Competing interests**
The author of this paper declares no competing interests.

Received: 1 December 2022 Revised: 15 April 2023 Accepted: 21 May 2023
Published online: 07 June 2023

**References**

Abdi, H., & Molin, P. (2007). Lilliefors/Van Soest's Test of Normality. Encyclopedia of Measurement and Statistics, pp. 540–544.
Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association, 49*, 765–769.
Aziz, F.H., Omar, R, & Ahmed, Q. (2022). Historical Overview of Air Temperature of Kurdistan Region-Iraq from 1973 to 2017. Journal of University of Garmian, the 10th Scientific Conference: Drought and Water Scarcity Management.
Barkotulla M. A. B. (2010). Stochastic Generation of the Occurrence and Amount of Daily Rainfall. *Pakistan Journal of Statistics and Operation Research*, 61–74.
Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (methodological), 26*(2), 211–243.
Brath, A., Montanari, A., & Toth, E. (2002). Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models. *Hydrology and Earth System Sciences, 6*(4), 627–639.
Cheddad, A. (2020). On box-cox transformation for image normality and pattern classification. *Institute of Electrical and Electronics Engineering (IEEE), 8*, 154975–154983.
Chung, G., Sim, K. B., Jo, D. J., & Kim, E. S. (2016). Hourly precipitation simulation characteristic analysis using Markov chain model. *Journal of the Korean Society of Hazard Mitigation, 16*(3), 351–357.
Danilovich A. (2016). New horizons: Iraqi federalism. In: *Iraqi Kurdistan in Middle Eastern Politics*. Routledge, pp. 49–70.
Fadhil, R. M., Rowshon, M. K., Ahmad, D., Fikri, A., & Aimrun, W. (2016). A stochastic rainfall generator model for simulation of daily rainfall events in Kurau catchment: model testing. *III International Conference on Agricultural and Food Engineering, 1152*, 1–10.
Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., & Lin, Q. (2020). Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. *Journal of Hydrology, 589*, 125188.
Garg, V. K., & Singh, J. B. (2010). Three-state Markov chain approach on the behaviour of rainfall. *New York Science Journal, 3*(12), 76–81.
Gui, Y., & Shao, J. (2017). Prediction of precipitation based on weighted markov chain in Dangshan. In *Proceedings of the International Conference on High Performance Compilation, Computing and Communications*, pp. 81–85.
Gupta, R. S. (1989). *Hydrology and hydraulic systems* (pp. 343–350). Prentice Hall.
Hajani, E., & Klari, Z. (2022). Trends analysis in rainfall data series in Duhok City, Kurdistan Region, Iraq. *Modeling Earth Systems and Environment., 8*, 4177.
Hajani, E., Shajee, K., Kaled, F., & Abdulhaq, H. (2022). Characteristics of changes in rainfall data in the Kurdistan Region Iraq. *Arabian Journal of Geosciences, 15*(509), 1–21.
He, R., Zhang, L., & Chew, A. W. (2022). Modeling and predicting rainfall time series using seasonal-trend decomposition and machine learning. *Knowledge-Based Systems, 251*, 109125.
Holt, C.C. (1957). Forecasting Seasonals and Trends by Exponentially Weighted Averages (O.N.R. Memorandum No. 52). Carnegie Institute of Technology, Pittsburgh USA.
Howard, R. A. (1971). *Dynamic probabilistic systems, 1–2*. Wiley.
Jain, S. (1986). A Markov chain model and its application. *Computers and Biomedical Research, 19*, 374–378.
Jale, J. S., Xavier Jr, S. F. A., Xavier, E. F. M., Stosic, T., Stosic, B., & Ferreira, T. A. E. (2019). Application of Markov chain on daily rainfall data in Paraíba-Brazil from 1995–2015. *Acta Scientiarum Technology, 41*(1), e37186.
Jimoh, O. D., & Webster, P. (1996). The optimum order of a Markov chain model for daily rainfall in Nigeria. *Journal of Hydrology, 185*(1–4), 45–69.
Kenton W. (2021). Markov Analysis: What is Markov Analysis. https://www.investopedia.com/terms/m/markov-analysis.asp
Kottegoda, N. T., Natale, L., & Raiteri, E. (2004). Some considerations of periodicity and persistence in daily rainfalls. *Journal of Hydrology, 296*, 23–37.
Kuo M. (2016). Mbd Excel: How to Create a Normally Distributed Set of Random Numbers in Excel. http://www.mbaexcel.com/excel/
Liu, C., Tian, Y. M., & Wang, X. H. (2011). Study of rainfall prediction model based on GM (1, 1)-Markov Chain. In *2011 International Symposium on Water Resource and Environmental Protection* (vol. 1, pp. 744–747). Institute of Electrical and Electronics Engineering (IEEE).
Maass, A., Hufschmidt, M. M., Dorfman, R., Thomas, H. A., Marglin, S. A., & Fair, G. M. (1962). *The design of water-resource systems* (p. 467). Cambridge: Harvard University Press.
Mahanta, J., Dey, S., & Khosro, P. (2018). Analyzing rainfall condition of Bangladesh: an application of Markov chain. *Thailand Statistician, 16*(2), 203–212.
Makridakis, S. G., Wheelwright, S., & Hyndman, R. (1998). *Forecasting: Methods and applications* (3rd ed.). New York: Wiley.
Malakoutian, M. M. A., Malakoutian, Y., Mostafapoor, P. & Abed, S. Z. D. (2021). Prediction for monthly rainfall of six meteorological regions and TRNC (Case Study: North Cyprus). *Computational Research Progress in Applied Science and Engineering*. H&T Publication hal-03228691.
Muhammad M. K. I. B. (2012). Time Series Modelling Using Markov and ARIMA Models. Doctoral dissertation, MSc. Thesis. Faculty of Civil Engineering, University of Technology Malaysia.
Oswal N. (2019). Predicting Rainfall Using Machine Learning Techniques. ARXIV preprint arXiv:1910.13827.
Preacher K. J. (2001). Calculation for the Chi-Square Test. An Interactive Calculation Tool for Chi-Square Tests of Goodness of Fit and Independence. http://quantpsy.org.
Rykov, V. V., Balakrishnan, N., & Nikulin, M. S. (2010). *Mathematical and statistical models and methods in reliability: Applications to medicine, finance, and quality control*. Springer Science and Business Media.
Sakia, R. M. (1992). The Box-Cox transformation technique: A review. *Journal of the Royal Statistical Society: Series D (The Statistician), 41*(2), 169.
Wang, D. Q., & Martiz, J. S. (1990). Note on testing a three state Markov chain for independence. *Journal Statistics Computation and Simulation, 37*(1–2), 61–68.
Weather and Climate Kurdistan, WCK (2021). Best Time to Visit Kurdistan, Iraq. https://www.besttimetovisit.com.pk/iraq/kurdistan-2330031/.

World Climate Guide, WCG (2019). Climate-Iraq. https://www.climatestotravel.com/climate.

Yusuf, A. U., Adamu, L., & Abdullahi, M. (2014). Markov chain model and its application to annual rainfall distribution for crop production. *American Journal of Theoretical and Applied Statistics, 3*(2), 39–43.

Zhang, S., Wang, H., & Zhang, X. (2014). Estimation of channel state transition probabilities based on Markov chains in cognitive radio. *Journal of Communications, 9*(6), 468–474.

Zhou, X., Wang, Y., & Zhou, X. (2017). Precipitation estimation based on weighted Markov chain model. In *2017 Seventh International Conference on Information Science and Technology (ICIST), Institute of Electrical and Electronics Engineering (IEEE)*, pp. 64–68.

## Publisher's Note