


Investigating the impact of climatic components on daily rainfall simulation (Case study: Khorramabad station)

Fereshteh Ahmadi^a, Mohammad Nazeri Tahroudi^{a*} 

^a Department of Water Engineering, Faculty of Agriculture, Lorestan University, Khorramabad, Iran

ABSTRACT

This study investigates the performance of three models Random Forest, Gaussian Process Regression and Contemporaneous Autoregressive Moving Average in simulating rainfall values at a rain gauge station, Khorramabad, Iran base on CanESM2 predictors. The models were evaluated using Root Mean Square Error and Nash-Sutcliffe Efficiency statistics to determine their predictive accuracy and efficiency. In the training phase, RF model exhibited an RMSE of 3.98 mm and an NSE of 0.32, indicating moderate predictive accuracy and efficiency. GPR showed improved performance with an RMSE of 2.55 mm and an NSE of 0.67, reflecting better predictive accuracy and a higher level of efficiency than RF. CARMA model demonstrated the best performance, achieving an RMSE of 1.2 mm and an NSE of 0.94, signifying high predictive accuracy and excellent efficiency. In the testing phase; the progressive improvement in RMSE values from 4.8 mm (GPR) and 4.1 mm (RF) to 1.3 mm (CARMA) across the models highlights the increasing accuracy in rainfall simulation. Similarly, the NSE values, ranging from 0.15 (GPR) and 0.2 (RF) to 0.93 (CARMA), underscore the enhanced efficiency of the models. The results of a graphical examination of different models in rainfall simulating values at the studied station also showed that the values simulated by the CARMA model are much more similar in terms of dispersion to the observed values. Among the three, CARMA model stands out as the most reliable and effective model for simulating rainfall values, making it a valuable tool for hydrological studies and water resource management.

ARTICLE INFO

Keywords:

CARMA Model
Contemporaneous Simulation
Gaussian Process
Random Forest

Article history:

Received: 02 Nov 2024
Accepted: 02 Jan 2025

*Corresponding author

E-mail address:
nazeri.mh@lu.ac.ir
(M. Nazeri Tahroudi)

Citation:

Ahmadi, F. & Nazeri Tahroudi, M., (2025). Investigating the impact of climatic components on daily rainfall simulation (Case study: Khorramabad station), *Sustainable Earth Trends*: 5(3), (36-51).

DOI: 10.48308/set.2025.238066.1099

1. Introduction

Climate change is defined by the accelerated misuse of natural resources by humans. Significant changes in weather patterns, including global warming, altered rainfall patterns, rising sea levels, and other atmospheric changes, are being used by nature to highlight this misuse. Human activities, such as greenhouse gas emissions from fossil fuels and deforestation, play a crucial role in these changes (Khalili et al., 2016b). The detection, simulation, and prediction of climate change and its components such as climate predictors of CanESM2 and CanESM5 have become essential for the optimal management of water resources in recent years. Various models have

been examined to simulate and predict precipitation under the influence of climate factors, and the impact of climate change on water resources and agricultural production has been studied. The simulation of rainfall and its pattern is a critical aspect of climate modeling, especially in the context of climate change (Tabatabaei et al., 2025). Accurate rainfall predictions are essential for effective water resource management, agriculture, and disaster preparedness. Recent advancements in machine learning techniques, particularly Random Forest (RF), and some useful models such as Gaussian Process Regression (GPR) and Contemporaneous Autoregressive models



(CARMA), have shown promise in improving the precision of rainfall simulations (Khalili et al., 2016b; Ahmad et al., 2019).

Random Forest is a powerful ensemble learning method that has been successfully utilized in various environmental modeling scenarios, including rainfall prediction (Ali et al., 2020). Its ability to handle complex data structures makes it particularly suitable for modeling nonlinear relationships between climate predictors and rainfall outcomes (Nazeri Tahroudi et al., 2023; Bageri et al., 2023). For instance, a study demonstrated that RF is effective in managing the interactions among multiple climate predictors, thus outperforming traditional linear models in predicting rainfall. This adaptability is crucial in climate studies, where datasets are often incomplete or exhibit nonlinear characteristics (Lee et al., 2017).

In addition, RF has been employed to generate multi-model ensemble (MME) precipitation and temperature data. The findings indicate that the MME generated using RF significantly enhances the accuracy of climate simulations compared to simpler methods (Lee et al., 2017). The model's ability to capture complex relationships between climate predictors and precipitation is a notable advantage, making RF a valuable asset in rainfall simulation studies.

Research has also highlighted the application of RF in related fields, such as predicting flood susceptibility. A study utilizing RF and boosted-tree models in Seoul revealed high accuracy in predicting flood-prone areas, achieving validation accuracies of 78.78% for regression and 79.18% for classification (Dou et al., 2019). This demonstrates the model's robustness in environmental applications and its capacity to integrate various climate predictors, such as topography and land use, to simulate rainfall-related events. Moreover, RF has shown effectiveness in handling high-dimensional data, providing insights into variable importance—which is critical for understanding factors contributing to flood risks (Dou et al., 2017). The implications of these findings extend to rainfall simulation, suggesting that RF can enhance the reliability of predictions in changing climate conditions.

In the field of climate change research and its impact on meteorological and hydrological parameters, various researchers have demonstrated the effects of climate change concerning precipitation and temperature changes. For instance, Kouhi et al. (2012)

analyzed extreme precipitation events in the Kashaf River basin using statistical downscaling methods and provided future scenarios. Their findings indicated that significant changes in precipitation patterns are expected in the near future, with an increase in the frequency and intensity of extreme precipitation events.

Mohammadi et al. (2017) examined trends in daily extreme precipitation indices in Iran using 24-hour data from 47 synoptic stations over the period from 1982 to 2012. Their results showed that annual precipitation decreased at 92% of the stations, with western stations and the Zagros foothills experiencing a more pronounced declining trend.

Askarizadeh et al. (2018) studied fluctuations in extreme precipitation indices in Mashhad using climate models and downscaling techniques. Data from the HadCM3, NCCCSM, and CNCM3 models under the B1A and A2 scenarios were extracted for two future periods (2018-2021 and 2100-2102). Their research indicated that climate change could lead to increased frequency and intensity of extreme precipitation in Mashhad, particularly in urban and peripheral areas. Zarei et al. (2019) predicted climatic elements of temperature and precipitation for the synoptic station in Gorgan using the SDSM downscaling model and outputs from the CanESM climate change model under various scenarios. The results suggest that in the near and mid-term future, precipitation will decrease from February to August, with a projected reduction of up to 47.51 mm during the distant period from December to August. Chengcheng et al. (2024) predicted changes in precipitation and temperature in the Beijing-Tianjin-Hebei region using ten models under RCP8.5 and RCP4.5 scenarios, showing that the MIROC-ESM-CHEM model is more consistent in predicting future precipitation capacity. Additionally, RCP8.5 indicates a greater increase in temperature compared to RCP4.5.

Vijayakumar et al. (2024) forecasted future climate changes in the state of Odisha, India, under RCP 4.5 and RCP 8.5 scenarios, indicating that temperatures will rise significantly in the coming centuries, particularly in western Odisha, with a notable increase in annual precipitation.

Bagora et al. (2024) examined precipitation variability patterns in the upper Chambal River basin up to the Gandhi Sagar Dam, utilizing the

SDSM model to refine results across three projected timeframes (2006–2036, 2037–2067, and 2068–2098) under RCP scenarios 2.6, 4.5, and 8.5. The results indicate that annual precipitation is expected to increase by 0.22%, 11.21%, and 5.65% during these timeframes, respectively.

Based on literature review, it appears that various models have been employed in climate change studies worldwide, depending on their specific nature (Ali et al., 2020; Mianabadi et al., 2023; Vijayakumar and Ramaraj, 2024; Nazeri Tahroudi and Mirabbasi, 2025). The present study aims to evaluate different statistical models regarding their performance, error rates, and efficiency, with the goal of identifying the best model for simulating and downscaling precipitation values at stations based on the fifth assessment report scenarios.

2. Material and methods

Lorestan Province is located in western Iran and has diverse climatic features. Considering its position in the Central Zagros,

this region includes high mountains and deep valleys. The average annual temperature ranges from 15 to 30 degrees Celsius, and the mean annual rainfall is about 600 to 800 millimeters. This province is situated between 50 to 55 degrees north longitude and 30 to 35 degrees south latitude. The presence of major rivers such as the Karkheh and Seymareh, which play an important role in irrigation and water supply for the province, increases the significance of studying climatic changes. The objective of this research is to simulate rainfall amounts at the Khorramabad rain gauge station based on the climatic parameters of CanESM2 model. In this study, daily rainfall data from the Khorramabad weather station, as well as climate predictors derived from the fifth Intergovernmental Panel on Climate Change (IPCC) report, have been utilized (<https://climate-scenarios.canada.ca/?page=main>). The geographical location of the Khorramabad weather station is illustrated in Fig. 1, and the changes in rainfall at this station during the statistical period from 1982 to 2023 are shown in Fig. 2. The statistical characteristics of the rainfall amounts are also presented in Table 1.

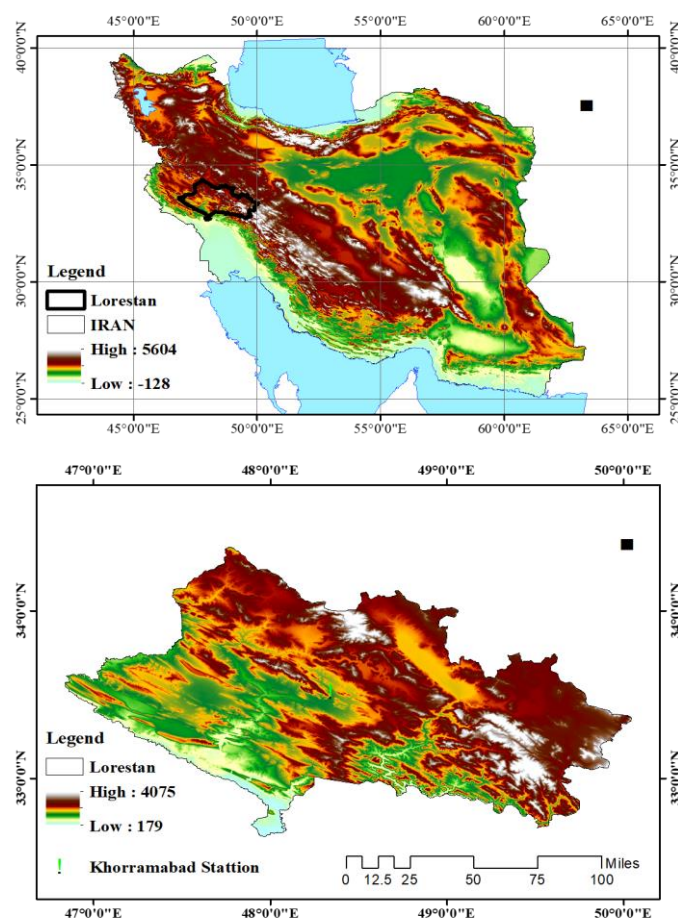


Fig. 1. Geographical location of the Khorramabad rain gauge station in Iran and Lorestan province.

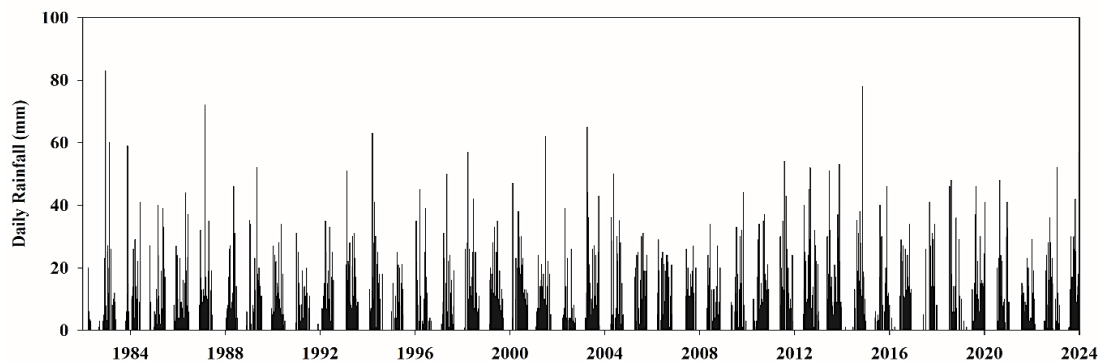


Fig. 2. Changes in rainfall amounts at the Khorramabad rain gauge station during the statistical period from 1982 to 2023 on a daily scale.

Table 1. Statistical characteristics of rainfall components at the studied station.

SKEW	KURT	VAR	STDEV	AVE (mm)	Max (mm)
6.22	50.77	25.26	5.03	1.28	83

2.1. Fifth IPCC report

The fifth IPCC report includes the Representative Concentration Pathways (RCPs), known as RCP 126, RCP 245, and RCP 585 (Amirabadizadeh et al., 2019). These scenarios are part of the climate change projections developed by the IPCC for modeling and predicting the effects of climate change and greenhouse gas reduction policies. Each of these scenarios represents different levels of greenhouse gas emissions and their impacts on the global climate (Swart et al., 2019; Javaherian et al., 2021).

2.1.1. RCP 126

This scenario is recognized as the “low-emission” scenario. The primary goal is to limit the increase in temperature to less than 2 degrees Celsius compared to pre-industrial levels. Achieving this goal requires significant and rapid reductions in greenhouse gas emissions (Chylek et al., 2011). This scenario can help preserve ecosystems and reduce the risks associated with climate change.

2.1.1.1. Strategies for RCP 126

- Extensive use of renewable energy sources (such as solar, wind, and hydrogen).
- Implementation of carbon capture and storage (CCS) technologies.
- Changes in energy consumption patterns and optimization of consumption.

It can mitigate negative impacts such as rising sea levels, droughts, and severe storms.

2.1.2. RCP 245

This scenario is considered a “medium-term” scenario. In this scenario, greenhouse gas emissions initially increase and then gradually decrease. The goal is to limit the temperature increase to around 2.5 degrees Celsius. A higher likelihood of more serious issues compared to RCP 126, but with measures to mitigate negative impacts.

2.1.2.1. Strategies for RCP 245

- A combination of renewable energy and fossil fuels with energy optimization technologies.
- Gradual implementation of emission reduction policies.

Increased temperatures and their effects on climate patterns and ecosystems.

2.1.3. RCP 585

This scenario is known as the “high-emission” scenario and indicates a lack of serious action to reduce greenhouse gas emissions. In this scenario, temperature increases are expected to exceed 4 degrees Celsius. This scenario could lead to severe and irreversible consequences, including significant sea level rise, drastic changes in climate patterns, and environmental and humanitarian crises. There are greater risks to public health, agriculture, and biodiversity.

2.1.3.1. Strategies for RCP 585

- Continued use of fossil fuels without significant changes in energy policies.

- Neglect of emission reduction technologies and consumption optimization.

2.2. Mann-Kendall Trend Test

The Mann-Kendall Test is a non-parametric statistical test used to identify trends in time series data. This test is particularly useful in environmental and climatic sciences for analyzing changes in variables such as temperature, precipitation, and other climatic parameters. The Mann-Kendall Test has been widely used in hydrology, climatology, and meteorology. It was developed by Mann (1945) and Kendall (1975). In this test, the statistic is calculated as follows (Khalili et al., 2016) (Eq. 1):

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(x_j - x_i) \quad (1)$$

Where x_j are sequential data values, n is the length of the dataset, and $\text{sgn}(\theta)$ is the sign function are defined as follows (Eq. 2):

$$\text{sgn}(\theta) = \begin{cases} 1 & \text{if } \theta > 0 \\ 0 & \text{if } \theta = 0 \\ -1 & \text{if } \theta < 0 \end{cases} \quad (2)$$

Where t_i is the number of identical data points in i^{th} group, m is equal to the number of paired groups. The standardized test statistic (Z) is calculated as follows (Eq. 3):

$$Z = \begin{cases} \frac{S - I}{\sqrt{\text{Var}(S)}} & S > 0 \\ 0 & S = 0 \\ \frac{S + I}{\sqrt{\text{Var}(S)}} & S < 0 \end{cases} \quad (3)$$

$$\text{Var}(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^m t_i(t_i-1)(2t_i+5)}{18}$$

The standardized test statistic MK1 (Z) follows a standard normal distribution with a mean of zero and a variance of one. The null hypothesis is accepted under the condition $-Z_{1-\alpha/2} \leq Z \leq Z_{1-\alpha/2}$. A very useful indicator in the Mann-Kendall test is the slope of the trend line, or the Sen's slope, which indicates the magnitude of the monotonic trend. The trend slope is estimated using the method

presented by Theil (1950) and Sen (1968) with the following relationship (Eq. 4):

$$\beta = \text{Median} \left(\frac{x_j - x_i}{j - i} \right) \quad \forall i < j \quad (4)$$

Where β is the estimator of the trend slope, x_i and x_j are the i -th and j -th observational values, respectively. Positive values indicate an increasing trend, while negative values indicate a decreasing trend. This method has been widely used in hydrological studies. It is important to note that estimating the Sen's slope is necessary for the calculations of the modified Mann-Kendall test. The modified Mann-Kendall method was developed by Kumar et al. (2009) by removing the effects of internal autocorrelation. In this method, the first-order autocorrelation coefficient is calculated, and if it is significant, it is removed from the data series. This method has also been applied by Burn et al. (2004), Lu et al. (2008), and Bandopadhyay et al. (2009).

2.3. Random forest algorithm

The Random Forest algorithm was introduced by Breiman (2001) as an ensemble learning method for regression and clustering problems based on the development of decision trees. A random forest is a collection of unpruned trees, where each tree is generated using a recursive partitioning algorithm. In other words, a random forest is a combination of several decision trees, each built using multiple self-organizing samples from the data (Friedman et al., 2001). To create a regression tree, recursive partitioning and multiple regressions are utilized. The decision process at each internal node of the root node is repeated according to the tree rule until a previously determined stopping criterion is met. In the RF method, a random vector X_n , which is independent of the random vectors $X_1, X_2, X_3, \dots, X_{n-1}$, is generated for the n training dataset and X_n , producing a set of trees equal to n as described below (Eqs. 5 and 6) (Breiman, 2001):

$$X_n = \{h_1(x), h_2(x), \dots, h_n(x)\} \quad (5)$$

$$h_n = h(x, X_n), x = \{x_1, x_2, \dots, x_p\} \quad (6)$$

The above p -dimensional vector forms a forest, and the outputs for each tree are provided as follows (Eq. 7):

$$y_1 = h_1(x), y_2 = h_2(x), \dots, y_n = h_n(x) \quad (7)$$

In the above equation, \bar{Y}_1 is the output of the n -th tree. To obtain the final output, the average of all the predictions from the trees is calculated (Breiman, 2001).

2.4. Gaussian Process Regression (GPR)

Gaussian regression operates based on a Bayesian approach by determining a probability distribution over all possible values available as inputs to the system (Kim and Gu, 2004). The Bayesian approach to regression establishes a probability distribution over all possible values for each parameter in a proposed function rather than providing precise values, based on classical approaches (Birge, 2004). To achieve this, Bayes' theorem is utilized (Sain et al., 1994; Burt et al., 2019; Yu et al., 2019).

2.5. Contemporaneous Autoregressive Moving Average (CARMA)

Various methods are required for analyzing and modeling hydrological time series. A characteristic feature of simultaneous models is the diagonal parameter matrix, where the estimation of parameters is independent of univariate models. Among the multivariate linear models, we can mention the multivariate autoregressive model, the simultaneous autoregressive moving average model defined as CARMA(p,q), the combined simultaneous and moving average model defined as CSM-CARMA(p,q), and the seasonal multivariate autoregressive model defined as MPAR(p).

Modeling multivariate hydrological processes based on the complete multivariate ARMA model often leads to challenges in parameter estimation. The CARMA model (simultaneous autoregressive moving average model) was proposed as a simpler alternative to the complete multivariate ARMA model (Salas et al., 1980). In the CARMA(p,q) model, the parameter matrices for both the autoregressive and moving average models are assumed to be diagonal, allowing a multivariate model to be considered independently of the univariate ARMA model. Thus, instead of jointly estimating the model parameters, they can be estimated independently for each univariate ARMA station, which aids in identifying the best univariate ARMA model. Therefore, if a

complete multivariate ARMA model is employed, the different temporal dependency structures can be modeled as a similar dependent structure over time for all stations rather than for each station separately.

The CARMA(p,q) model for n stations can be represented as follows (Eq. 8):

$$Y_t = \sum_{i=1}^p \varphi_i Y_{t-i} + \varepsilon_t - \sum_{j=1}^q \theta_j \varepsilon_{t-j} \quad (8)$$

Where Y_t is an $n \times 1$ column matrix of observational series with a normal distribution and zero mean, representing different stations $k=1,2,\dots,n$. $\varphi_1, \varphi_2, \dots, \varphi_p$ is an $n \times n$ diagonal matrix of autoregressive (self-correlated) model parameters, and $\theta_1, \theta_2, \dots, \theta_q$ is an $n \times n$ diagonal matrix of moving average model parameters. Additionally, ε_t is an $n \times 1$ matrix of normally distributed random data with zero mean and covariance-variance g (Eq. 9).

$$\begin{bmatrix} Y_t^{(1)} \\ Y_t^{(2)} \\ \vdots \\ Y_t^{(n)} \end{bmatrix} = \begin{bmatrix} \varphi^{11} & \varphi^{12} & \dots & \varphi^{1n} \\ \varphi^{21} & \varphi^{22} & \dots & \varphi^{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi^{n1} & \varphi^{n2} & \dots & \varphi^{nn} \end{bmatrix} \begin{bmatrix} Y_{t-1}^{(1)} \\ Y_{t-1}^{(2)} \\ \vdots \\ Y_{t-1}^{(n)} \end{bmatrix} + \begin{bmatrix} \theta^{11} & \theta^{12} & \dots & \theta^{1n} \\ \theta^{21} & \theta^{22} & \dots & \theta^{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta^{n1} & \theta^{n2} & \dots & \theta^{nn} \end{bmatrix} \begin{bmatrix} \varepsilon_t^{(1)} \\ \varepsilon_t^{(2)} \\ \vdots \\ \varepsilon_t^{(n)} \end{bmatrix} \quad (9)$$

The CARMA model is capable of preserving the zero-lag cross-correlation in the space between different stations. Additionally, the temporal dependency structure for each station is defined by the parameters p and q (Salas et al., 1980).

2.5.1. Parameter Estimation of the CARMA Model

Considering N years of data at each station $Y_t^{(i)}$ with observational data, where $i=1,2,3,\dots,n$, the general model matrix Y_t is described as follows (Eq. 10):

$$Y_t = \mu + \sigma Z_t \quad (10)$$

where μ and σ represent the mean and variance of Y_t , respectively. The standardization of the variables is calculated using the following (Eq. 11):

$$Z_t^{(i)} = (y_t^{(i)} - \mu_t^{(i)}) / \sigma_t^{(i)}, i=1,2,3,\dots,n \quad (11)$$

Parameters of the CARMA(p(i), q(i)) model are determined in a manner similar to the parameters of the ARMA model. The residual time series of the model is independent of time but is dependent on each other (spatially dependent). This cross-dependency can be modeled using the following relationship (Eqs. 12 and 13):

$$\varepsilon_t^{(i)} = \frac{\varepsilon_t^{(i)}}{\sigma_t^{(i)}} \quad (12)$$

$$\varepsilon_t' = B \xi_t \quad (13)$$

Where B is estimated using the following (Eq. 14):

$$\widehat{B}\widehat{B}^T = \widehat{M}_0 \quad (14)$$

Where \widehat{M}_0 represents the zero-lag autocorrelation matrix, which is calculated from the following matrix (Eqs. 15 and 16):

$$\widehat{M}_k = \begin{bmatrix} r_k^{11} & r_k^{12} & \dots & r_k^{1n} \\ r_k^{21} & r_k^{22} & \dots & r_k^{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_k^{n1} & r_k^{n2} & \dots & r_k^{nn} \end{bmatrix} \quad (15)$$

$$r_k^{ij} = \frac{\sum_{t=1}^{N-k} (\varepsilon_t^{(i)} - \bar{\varepsilon}_t^{(i)})(\varepsilon_{t+k}^{(j)} - \bar{\varepsilon}_{t+k}^{(j)})}{\text{Sqrt} \left[\sum_{t=1}^{N-k} (\varepsilon_t^{(i)} - \bar{\varepsilon}_t^{(i)})^2 \cdot \sum_{t=1}^{N-k} (\varepsilon_{t+k}^{(j)} - \bar{\varepsilon}_{t+k}^{(j)})^2 \right]} \quad (16)$$

Where $\bar{\varepsilon}_t^{(i)}$ is the mean of the $N-k$ data corresponding to station i , and $\bar{\varepsilon}_{t+k}^{(j)}$ is the mean of the $N-k$ data corresponding to station j . Finally, the parameter matrix of the CARMA(p,q) model is obtained using the following relationship (Eq. 17) (Matalas, 1967):

$$\hat{A}_l = \hat{M}_l \hat{M}_0^{-1} \quad (17)$$

2.6. Model evaluation criteria

In order to assess the accuracy and efficiency of the models under consideration, the Root Mean Square Error (RMSE) and the Nash-Sutcliffe Efficiency (NSE) statistics were used as follows (Eqs. 18 and 19):

1- Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Q_i - Q_i')^2}{n}} \quad (18)$$

2-Nash-Sutcliffe Efficiency Coefficient

$$NSE = 1 - \frac{\sum_{i=1}^T (Q_i - Q_i')^2}{\sum_{i=1}^T (Q_i - \bar{Q})^2} \quad (19)$$

Where: Q_i is the actual or observed data, Q_i' represents the predicted data, \bar{Q} is the mean of the data, and n is the number of data points (Nash and Sutcliffe, 1970; Swinscow and Campbell, 2002; Salas et al., 1980).

3. Results and discussion

In this section, the results of daily rainfall simulation based on selected climatic components are presented to compare the performance and accuracy of the models used. The results of the simulations were evaluated using statistical measures, followed by the correlation coefficient during the training and testing phases. Before simulating the precipitation values at the studied station using existing models, the trend of changes in precipitation values was examined using the modified Mann-Kendall test, which accounted for internal autocorrelation. The results showed an incremental but insignificant trend in the annual precipitation values, with Z-statistic values = 0.1.

3.1. Selection the climatic predictors corresponding to daily rainfall

In this study, to select input values for the models under investigation, the correlation between climatic predictors and rainfall values at the study station was examined. Based on the higher correlation coefficient, the predictors Mean Sea Level Pressure, 1000 hPa Divergence of True Wind, 500 hPa Geopotential, and Air Temperature at 2 m were selected from 26 predictors extracted from the CanESM2 climate model. Table 2 shows the climatic predictors, with the bolded items selected for simulating rainfall values at the study station. Ultimately, using the climatic predictors and rainfall values from the Khorramabad rain gauge station, the variations in rainfall values were simulated considering the components of climate change. In this regard, RF, GPR, and CARMA models were employed in two phases: training and testing.

3.2. Results of simulating rainfall values at the Khorramabad station using the RF model

The simulation of daily rainfall values at the Khorramabad station using the Random Forest (RF) model during the training phase is presented in Fig. 3. As shown in Fig. 3, the 1:1 line does not form a 45-degree bisector, indicating that the simulated rainfall values do not align well with the observed values.

Considering the confidence interval of the simulation (95%), it can be observed that there is significant overestimation and underestimation in the RF model results at the Khorramabad station. The error rate (RMSE) calculated is 3.98 mm, and the model efficiency is 32 percent during the training phase, which is not considered satisfactory. The results of simulating daily rainfall values using the Random Forest model in the testing phase are also presented in Fig. 4. According to Fig. 4, the performance of the Random Forest model in simulating daily rainfall, considering the

components of climate change, has decreased in the testing phase compared to the training phase. The model error has increased to 4.1 mm (RMSE), and the model efficiency (NSE) has dropped to 20 percent, which is not satisfactory given the confidence interval of the simulations and indicates weak certainty. Based on the results presented in Figs 3 and 4, it can be concluded that the Random Forest model does not possess the necessary efficiency for simulating daily rainfall values at the study station considering the climatic components.

Table 2. Climatic predictors extracted from the CanESM2 climate model corresponding to the Khorramabad station.

No	Predictor	Name
1	mslp	Mean sea level pressure
2	p1_f	1000 hPa Wind speed
3	p1_u	1000 hPa Zonal wind component
4	p1_v	1000 hPa Meridional wind component
5	p1_z	1000 hPa Relative vorticity of true wind
6	p1th	1000 hPa Wind direction
7	p1zh	1000 hPa Divergence of true wind
8	p5_f	500 hPa Wind speed
9	p5_u	500 hPa Zonal wind component
10	p5_v	500 hPa Meridional wind component
11	p5_z	500 hPa Relative vorticity of true wind
12	p5th	500 hPa Wind direction
13	p5zh	500 hPa Divergence of true wind
14	p8_f	850 hPa Wind Speed
15	p8_u	850 hPa Zonal wind component
16	p8_v	850 hPa Meridional wind component
17	p8_z	850 hPa Relative vorticity of true wind
18	p8th	850 hPa Wind direction
19	p8zh	850 hPa Divergence of true wind
20	p500	500 hPa Geopotential
21	p850	850 hPa Geopotential
22	prcp	Total precipitation
23	s500	500 hPa Specific humidity
24	s850	850 hPa Specific humidity
25	shum	1000 hPa Specific humidity
26	temp	Air temperature at 2 m

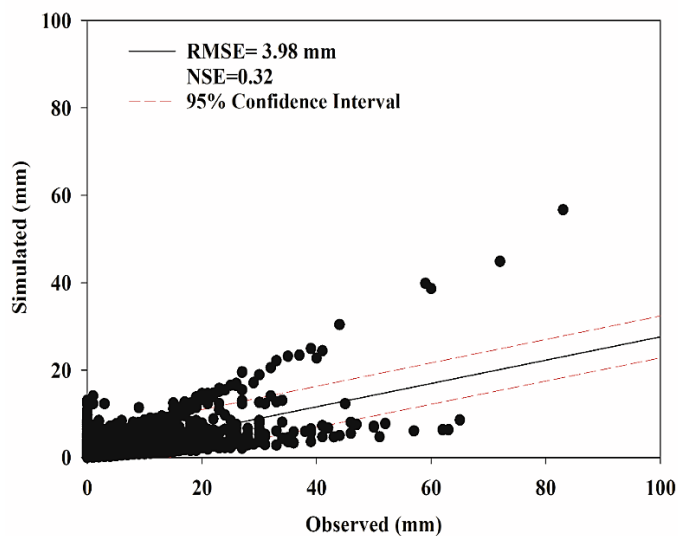


Fig. 3. Results of simulating daily rainfall values at the Khorramabad station in the training phase using the random forest model.

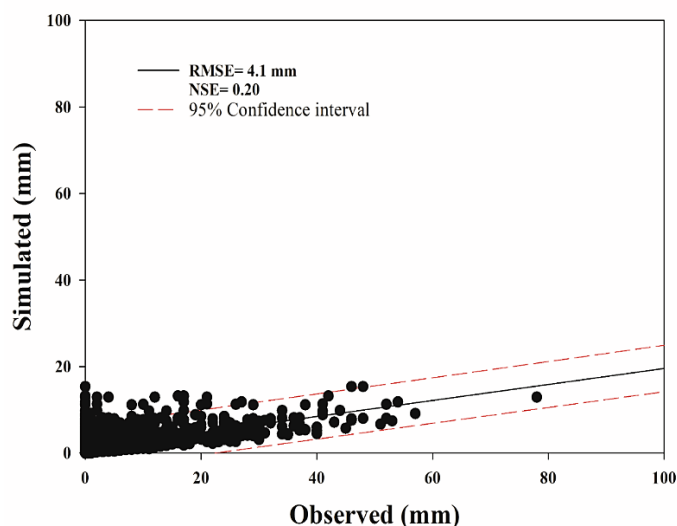


Fig. 4. Results of simulating daily rainfall values at the Khorramabad station in the testing phase using the random forest model.

3.3. Results of simulating rainfall values at the Khorramabad station using the GPR model

In the next step, the performance and efficiency of the Gaussian Process Regression (GPR) model in simulating daily rainfall values at the Khorramabad station, considering the selected climatic predictors, were evaluated in both the training and testing phases. The results of simulating daily rainfall values at the Khorramabad station using the GPR model in the training and testing phases are presented in Figs 5 and 6, respectively. As shown in Fig. 5, the performance of the GPR model in the training phase is better than that of the RF model. Specifically, the

Root Mean Square Error (RMSE) is approximately 2.55 millimeters, and the model efficiency is estimated to be 67 percent in the training phase. The 1:1 line in the training phase (Fig. 5) also shows a better fit compared to the similar instance in the RF model. Comparing Figs 5 and 3, it can be observed that the error in simulating daily rainfall at the Khorramabad station using the GPR model has decreased by about 36 percent compared to the RF model, and correspondingly, the model efficiency, as indicated by the Nash-Sutcliffe Efficiency (NSE) coefficient, has increased by approximately 109 percent. The performance of the GPR model, unlike the

RF model, is satisfactory in simulating daily rainfall.

In the testing phase, as shown in Fig. 6, the performance of the GPR model in simulating daily rainfall has decreased compared to the training phase, and this decline is evident in both the error rate and the NSE coefficient. The 1:1 line in the testing phase for the GPR model also indicates poor performance in simulating daily rainfall values at the Khorramabad station. In the testing phase, the error values, as indicated by the RMSE statistic, are equal to 4.8 millimeters, and the model

efficiency, as indicated by the NSE statistic, is around 15 percent, which shows weaker performance compared to the RF model in the testing phase. Compared to the RF model in the testing phase, the GPR model has increased the error by approximately 17 percent and decreased the model efficiency by about 25 percent. The increase in error and decrease in model performance during the testing phase compared to the training phase is reasonable; however, this level of decline indicates the Poor performance of both the RF and GPR models in simulating daily rainfall values in the study area.

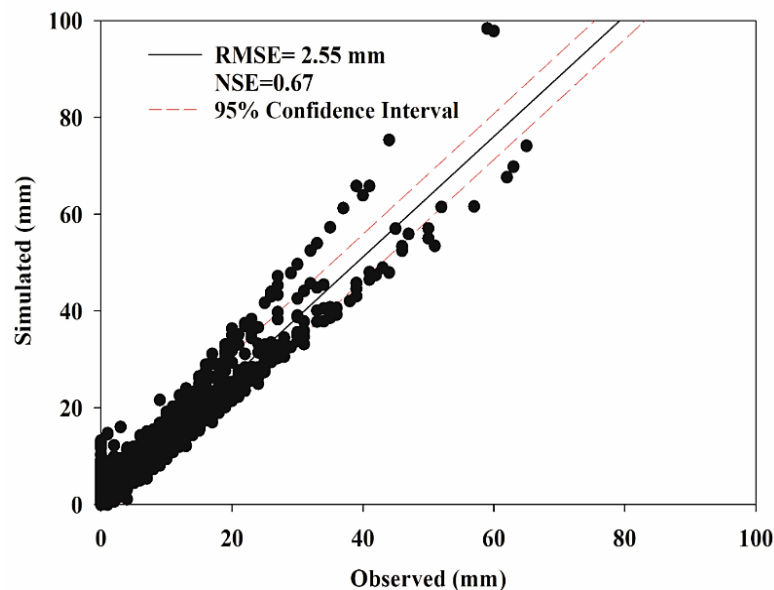


Fig. 5. Results of simulating daily rainfall values at the Khorramabad station in the training phase using the gaussian process regression model.

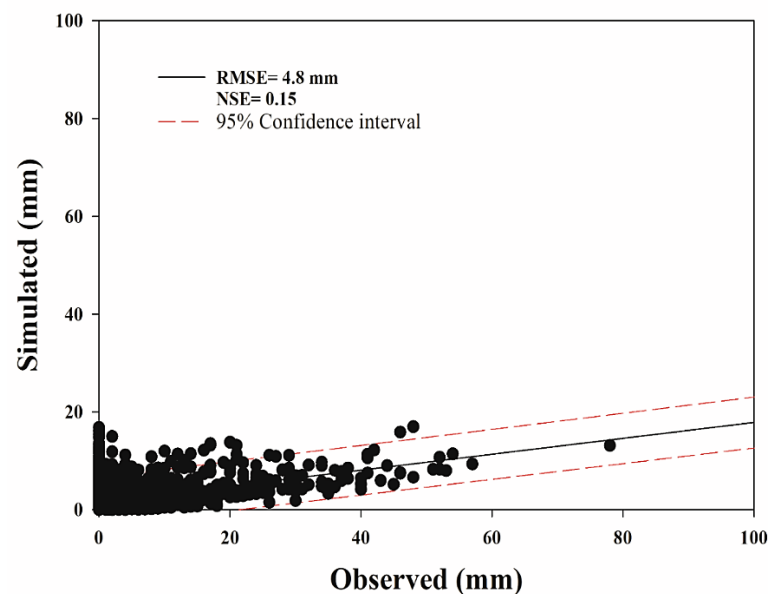


Fig. 6. Results of simulating daily rainfall values at the Khorramabad station in the testing phase using the gaussian process regression model.

3.4. Results of simulating rainfall values at the Khorramabad station using the CARMA model

The results of simulating rainfall values using the CARMA model are presented in Figs 7 and 8 for the training and testing phases, respectively. The CARMA model demonstrated a different performance in simulating daily rainfall while considering the selected climatic predictors. Notably, the structure of the CARMA model differs from that of the RF and GPR models. This model, as a contemporaneous ARMA model, is utilized in meteorological and hydrological studies and is inherently stochastic, allowing it to establish a suitable relationship with stochastic values. As shown in Fig. 7, the 1:1 line covers the first quartile completely, indicating that the 95 percent confidence interval reflects the performance of the CARMA model in the training phase. Although there are instances outside the 95 percent confidence interval, the CARMA model shows a satisfactory performance in simulating daily rainfall values compared to the other two models. The error rate is observed to be 1.2 mm (RMSE), with a model efficiency (NSE) of 94 percent in simulating daily rainfall at the studied station. In the training phase, this represents a reduction in error of

approximately 70 percent compared to the RF model and 53 percent compared to the GPR model, corresponding to reductions of 2.78 and 1.35 mm, respectively. In the testing phase, as shown in Fig. 8, there are no significant reductions in the simulation of rainfall values at the studied station using the CARMA model. This indicates that in the absence of rainfall data, the performance of the CARMA model in simulating based on climatic predictors has been satisfactory, and this model exhibits higher certainty compared to the RF and GPR models. According to Fig. 8, the error of the CARMA model in the testing phase has decreased by approximately 68 percent compared to the RF model and 73 percent compared to the GPR model, while its efficiency has improved by about 365 percent compared to the RF model and 520 percent compared to the GPR model. The 95 percent confidence interval and the 1:1 line further illustrate this superiority over the RF and GPR models in both the training and testing phases. Khalili and Nazeri Tahroudi (2016) found the CARMA model's results satisfactory in their studies on simulating precipitation values. They introduced this model as the best alternative to univariate time series models.

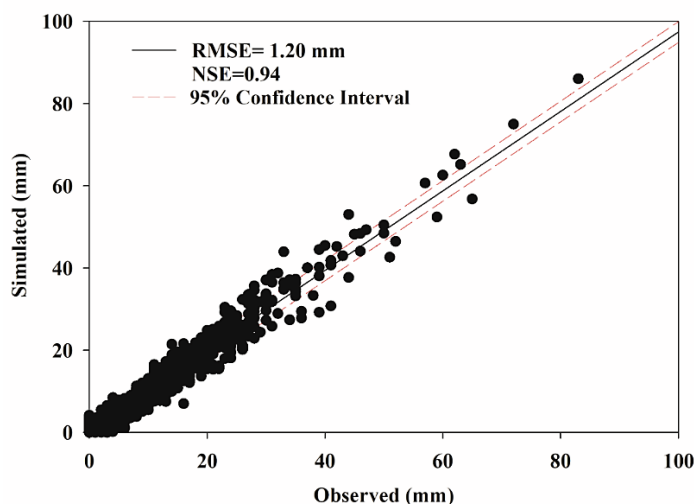


Fig. 7. Results of simulating daily rainfall values at the Khorramabad station in the training phase using the CARMA model.

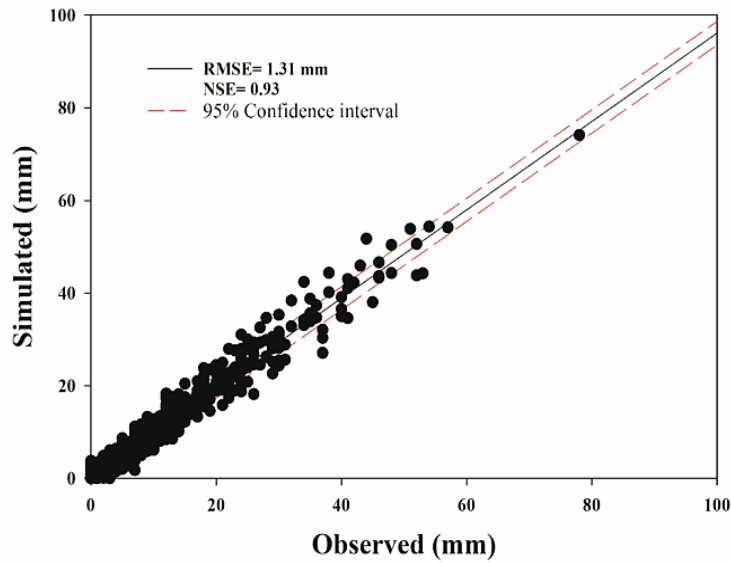


Fig. 8. Results of simulating daily rainfall values at the Khorramabad station in the testing phase using the CARMA model.

3.5. Graphical comparison of simulated daily rainfall values using the studied models

To compare the performance and efficiency of the models studied in this research, violin plots and Taylor’s diagrams were utilized, which also allow for the assessment of the certainty of the studied models, as presented in Figs 9 and 10. As shown in Fig. 9, the violin plot illustrates the observed and simulated daily rainfall values at the Khorramabad station. The results indicate that the dispersion of simulated rainfall values in the testing phase for the RF and CARMA

models are better than that of the GPR model. Additionally, Fig. 9 reveals that the GPR model has a significant overestimation compared to the other two models, while the RF model exhibits a tendency for underestimation in simulating rainfall values. From Fig. 9, it can also be observed that in terms of similarity, the violin plot corresponding to the CARMA model performs better than the other two models. Therefore, it can be concluded that the certainty of the CARMA model is considerably higher than that of the other two models, showing a strong resemblance to the observed values.

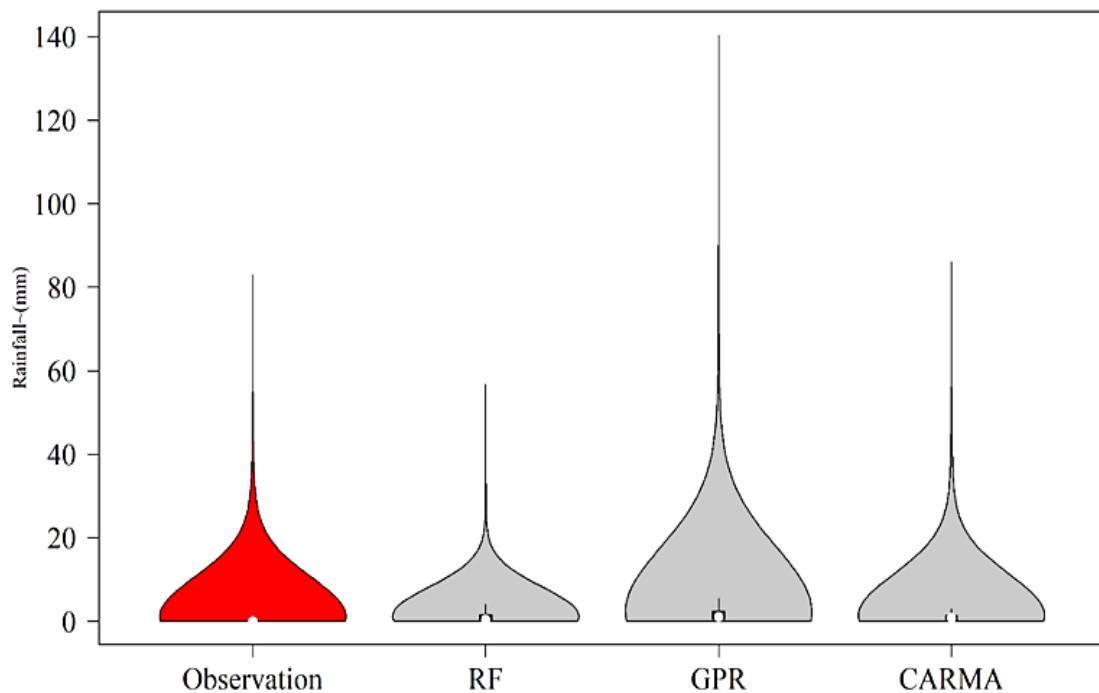


Fig. 8. Violin plot of observed and simulated values based on the studied models for analysis.

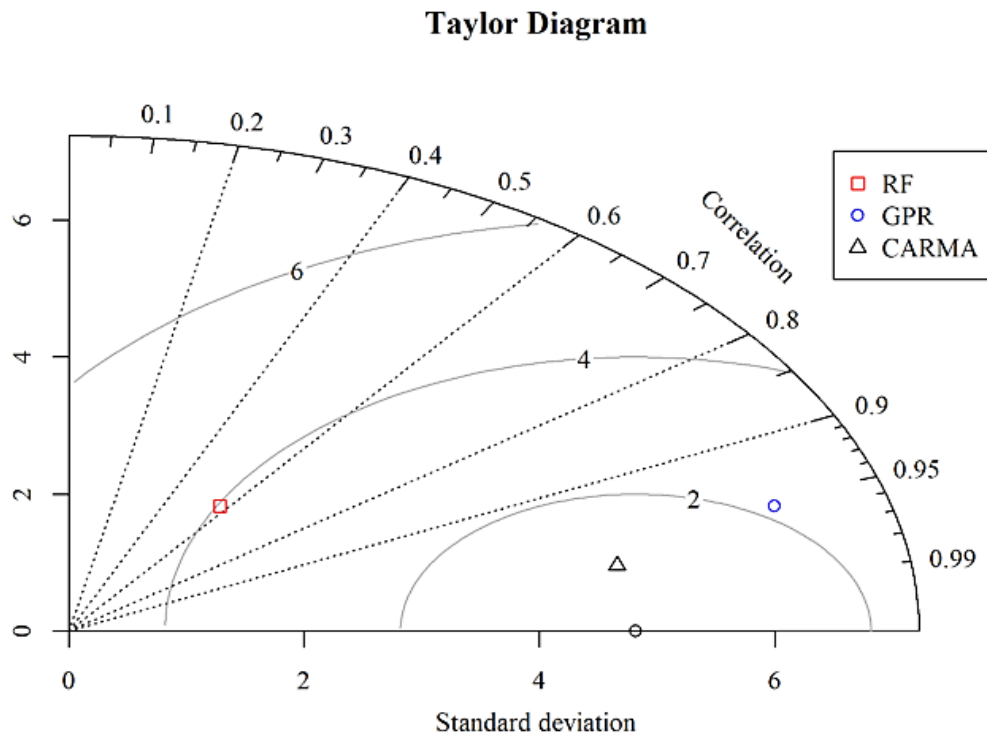


Fig. 10. Taylor diagram of observed and simulated values based on the studied models for analysis.

In Fig. 10, the Taylor diagram illustrating the observed and simulated values can also be seen. According to Figure 9, the correlation between the observed and simulated daily rainfall values at the studied station is approximately 0.6 for the RF model, about 0.95 for the GPR model, and around 0.99 for the CARMA model. The Taylor's diagram indicates that the deviation of the simulated values using the CARMA model is closer to the observed values compared to the RF and GPR models. Overall, based on the results of error statistics, the Taylor's diagram, and the violin plot, it can be concluded that the performance, certainty, and efficiency of the CARMA model are superior to those of the other studied models, thus identifying it as the best model. Khalili et al. (2016a) also reported satisfactory performance of the CARMA model in simulating meteorological and hydrological values in their studies, suggesting it as an alternative to simpler univariate ARMA models. The use of climatic predictors in simulating meteorological and hydrological variables is significant for two reasons: first, it indirectly considers the impacts of climate variability on simulations and climatic predictions; second, it employs multivariate models that incorporate influential parameters. This aspect has also been confirmed in the studies by Zarei et al. (2019). The performance of random forest and Gaussian process

regression models has been examined in various studies, and their effectiveness has been validated in the context of time series models and other different models (Merufinia et al., 2023; Vinta and Peeriga, 2024).

4. Conclusion

In this study, the aim was to simulate daily rainfall considering climate variability and climatic predictors, and to evaluate the accuracy and performance of three models: Random Forest (RF), Gaussian Process Regression (GPR), and the Autoregressive Integrated Moving Average (ARIMA) model. The results indicated that during the training phase, the performance of the GPR model was better than that of the RF model; however, in the testing phase, the RF model outperformed the GPR model based on RMSE and NSE statistics. GPR models can sometimes overfit the training data, capturing noise along with the underlying patterns. This can lead to excellent performance during training but poorer generalization to new, unseen data during testing. Also, RF models, being ensemble methods, are generally more robust to overfitting. They average the results of multiple decision trees, which helps in capturing the overall trend without being too sensitive to noise in the training data.

The 36% superiority of the GPR model in reducing error during the training phase, as indicated by the RMSE statistic compared to the RF model, demonstrated the success of the GPR model in simulating daily rainfall at the studied station. Nevertheless, in the testing phase, in the absence of observational data, the GPR model ranked third with a 17% reduction in error compared to the RF model. In contrast, the CARMA model consistently achieved the highest rank in both the training and testing phases when considering error values and model efficiency. This model demonstrated the best performance in simulating daily rainfall values, with an error of 1.2 mm and an efficiency coefficient of 94% in the training phase, and an error of 1.31 mm and an efficiency coefficient of 93% in the testing phase. GPR often outperforms RF in rainfall simulation for several reasons which have been pointed out in various studies (Mlakar et al., 2019; Shabani et al., 2020).

A: GPR models are inherently smooth and continuous, which is beneficial for simulating rainfall patterns that typically exhibit gradual changes rather than abrupt shifts. This smoothness allows GPR to capture the underlying trends more effectively.

B: GPR provides a probabilistic framework that includes uncertainty estimates for its predictions that this is particularly useful in rainfall simulation, where understanding the confidence in predictions can be crucial for decision-making and risk assessment.

C: GPR can model the spatial and temporal correlations in the data, which are essential for accurate rainfall simulation. This ability to incorporate and leverage correlations helps in producing more reliable and realistic simulations.

D: GPR tends to perform better with smaller datasets compared to RF. In many cases, high-quality rainfall data may be limited, making GPR a more suitable choice.

E: While RF is good at handling non-stationary data by partitioning the input space, GPR can be adapted to non-stationary conditions through various modifications, such as using non-stationary kernels. This adaptability enhances its performance in complex rainfall simulations.

Overall, the combination of smoothness, uncertainty quantification, and the ability to model correlations makes GPR a superior choice for rainfall simulation in many

scenarios. CARMA models are specifically designed to handle continuous-time data, making them well-suited for capturing the temporal dynamics of rainfall, which can be highly variable and continuous in nature. Also, Rainfall data often exhibit non-stationary behavior. CARMA models can effectively handle non-stationarity by incorporating both autoregressive and moving average components, providing a more accurate representation of the underlying processes. In contrast, while GPR is powerful for capturing smooth, nonlinear relationships and RF is robust for handling high-dimensional data, they may struggle with the continuous and highly variable nature of rainfall data and this is proof of the superiority of the CARMA model. Due to the stochastic and multivariate nature of the CARMA model, it can be applied to various climates around the world without restrictions.

Acknowledgments

The authors express their gratitude to the Regional Water Company of Lorestan, Iran, for supplying the data. Additionally, they declare that this article did not receive any financial support.

References

- Ahmed, K., Sachindra, D.A., Shahid, S., Demirel, M.C. & Chung, E.S., 2019. Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. *Hydrology and Earth System Sciences*, 23(11), 4803-4824.
- Ali, M., Prasad, R., Xiang, Y. & Yaseen, Z.M., 2020. Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. *Journal of Hydrology*, 584, 124647.
- Amirabadizadeh, M., Ramezani, Y., Nazeri Tahroudi, M. & Zeynali, M.J., 2019. Assessment of data-driven models in downscaling of the daily temperature in Birjand synoptic station. *AUT Journal of Civil Engineering*, 3(2), 193-200.
- Askarizadeh, S.M., Mozaffari, G. & Mazidi, A., 2018. Estimating the Fluctuations of Rainfall Extreme Indices in Mashhad for the Next Two Periods of 2011-2030 and 2046-2065 Using LARS-WG Model's Downscaling. *Journal of Geography and Regional Development*, 16(1), 25-50.
- Bageri, F., Khalili, K. & Nazeri Tahroudi, M., 2023. Evaluation of Entropy Theory Based on Random Forest in Quality Monitoring of Ground Water Network. *Water and Irrigation Management*, 13(1), 123-139.

- Bagora, P. & Narulkar, S., 2024. Enhancing Monsoon Predictions for the Upper Chambal Catchment Through Temporal and Spatial Downscaling of Predicted Future Precipitation. *Journal of The Institution of Engineers (India): Series A*, 105(3), 703-717.
- Bajirao, T.S., Kumar, P., Kumar, M., Elbeltagi, A. & Kuriqi, A., 2021. Potential of hybrid wavelet-coupled data-driven-based algorithms for daily runoff prediction in complex river basins. *Theoretical and Applied Climatology*, 145(3), 1207-1231.
- Bandyopadhyay, A., Bhadra, A., Raghuvanshi, N.S. & Singh, R., 2009. Temporal trends in estimates of reference evapotranspiration over India. *Journal of Hydrologic Engineering*, 14(5), 508-515.
- Breiman, L., 2001. Random Forests. *Machine Learning* 45(1): pp. 5–32.
- Birge, L., 2004. Model selection for Gaussian regression with random design. *Bernoulli*, 10(6), 1039-1051.
- Burn, D.H., Cunderlik, J.M. & Pietroniro, A., 2004. Hydrological trends and variability in the Liard River basin/Tendances hydrologiques et variabilité dans le bassin de la rivière Liard. *Hydrological Sciences Journal*, 49(1), 53-67.
- Burt, D., Rasmussen, C.E. & Van Der Wilk, M., 2019, May. Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning* (pp. 862-871). PMLR.
- Chengcheng, X., Kaixuan, G., Jama, A.H., Chuiyu, L., Qingyan, S., Xu, L. & Lingjia, Y., 2024. Simulation and prediction of precipitation and temperature under RCP scenarios. *Water Supply*, 24(5), 1676-1688.
- Chylek, P., Li, J., Dubey, M.K., Wang, M. & Lesins, G. J. A.C., 2011. Observed and model simulated 20th century Arctic temperature variability: Canadian earth system model CanESM2. *Atmospheric Chemistry and Physics Discussions*, 11(8), 22893-22907.
- Dou, J., Yunus, A.P., Bui, D.T., Merghadi, A., Sahana, M., Zhu, Z., ... & Pham, B.T., 2019. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Science of the Total Environment*, 662, 332-346.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
- Jain, S.K. Kumar, V. & Saharia, M., 2013. Analysis of rainfall and temperature trends in northeast India. *International Journal of Climatology*, 33, 968-978.
- Javaherian, M., Ebrahimi, H. & Aminnejad, B., 2021. Prediction of changes in climatic parameters using CanESM2 model based on RCP scenarios (case study): Lar dam basin. *Ain Shams Engineering Journal*, 12(1), 445-454.
- Kendall M.G., 1975. Rank Correlation Measures, Charles Griffin, London.
- Khalili, K. & Nazeri Tahroudi, M., 2016. Performance evaluation of ARMA and CARMA models in modeling annual precipitation of Urmia synoptic station. *Water and Soil Science*, 26(2-1), 13-28.
- Khalili, K., Tahoudi, M.N., Mirabbasi, R. & Ahmadi, F., 2016a. Investigation of spatial and temporal variability of precipitation in Iran over the last half century. *Stochastic Environmental Research and Risk Assessment*, 30(4), 1205-1221.
- Khalili, K., Tahoudi, M.N., Mirabbasi, R. & Ahmadi, F., 2016b. Investigation of spatial and temporal variability of precipitation in Iran over the last half century. *Stochastic Environmental Research and Risk Assessment*, 30, 1205-1221.
- Kim, Y. J. & Gu, C., 2004. Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2), 337-356.
- Kouhi, M., Mousavi Baygi, M., Farid hosseini, A.R., Sanaei Nejad, S.H. & Jabbari Nooghabi, H., 2012. Statistical Downscaling of Extremes of precipitation and construction of their future scenarios in the Kashfroud Basin. *Journal of Climate Research*, 1391(12), 35-53.
- Kumar, S., Merwade, V., Kam, J. & Thurner, K., 2009. Streamflow trends in Indiana: effects of long term persistence, precipitation and subsurface drains. *Journal of Hydrology*, 374(1), 171-183.
- Lee, S., Kim, J.C., Jung, H.S., Lee, M.J. & Lee, S., 2017. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk*, 8(2), 1185-1203.
- Luo, Y., Liu, S., Fu, S., Liu, J., Wang, G. & Zhou, G., 2008. Trends of precipitation in Beijiang River basin, Guangdong province, China. *Hydrological Processes*, 22(13), 2377-2386.
- Mann, H.B., 1945. Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society*, 245-259.
- Matalas, N.C., 1967. Mathematical assessment of synthetic hydrology. *Water Resources Research*, 3(4): 937-945.
- Merufinia, E., Sharafati, A., Abghari, H. & Hassanzadeh, Y. (2023). On the simulation of streamflow using hybrid tree-based machine learning models: A case study of Kurkursar basin, Iran. *Arabian Journal of Geosciences*, 16(1), 28.
- Mianabadi, A., Bateni, M. M. & Mohammadi, S., 2023. Projection of Change in the Distribution of Precipitation and Temperature Using Bias-Corrected Simulations of CMIP6 Climate Models (Case Study: Kerman Synoptic Station). *Climate Change Research*, 4(14), 65-84.
- Mlakar, M., Tusar, T. & Filipic, B., 2019. Comparing random forest and gaussian process modeling in the GP-demo algorithm. *Information Security Education Journal (ISEJ)*, 6(1), 9.
- Mohammadi, H., Azizi, G., Khoshahklagh, F. & Ranjbar, F., 2017. Analysis of daily precipitation extreme indices trend in Iran. *Physical Geography Research*, 49(1), 21-37.
- Nash, J.E. & Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology*, 10(3), 282-290.

- Nazeri Tahroudi, M. & Mirabbasi, R., 2025. Evaluating the efficiency and accuracy of the copula-based rainfall-runoff model. *Earth Science Informatics*, 18(1), 122.
- Nazeri Tahroudi, M., Ahmadi, F. & Mirabbasi, R., 2023. Performance comparison of IHACRES, random forest and copula-based models in rainfall-runoff simulation. *Applied Water Science*, 13(6), 134.
- Nazeri Tahroudi, M., Amirabadyzadeh, M. & Zeynali, M. J., 2018. Evaluation of the accuracy of artificial intelligence and regression models in simulation the daily temperature. *Journal of Meteorology and Atmospheric Science*, 1(1), 65-76.
- Sain, S.R., Baggerly, K.A. & Scott, D.W., 1994. Cross-validation of multivariate densities. *Journal of the American Statistical Association*, 89(427), 807-817.
- Salas, J.D., Delleur, J.W., Yevjevich, V. & Lane, W.L., 1980. *Applied Modeling of Hydrologic Time Series*. Water resource Publications, P. O. Box 2841. Littleton, Colorado .80161, U.S.A. 484 P.
- Salas, J. D. (1980). *Applied Modeling of Hydrologic Time Series*. Water Resources Publication.
- Shabani, S., Samadianfard, S., Sattari, M.T., Mosavi, A., Shamshirband, S., Kmet, T. & Varkonyi-Koczy, A.R., 2020. Modeling pan evaporation using Gaussian process regression K-nearest neighbors random forest and support vector machines; comparative analysis. *Atmosphere*, 11(1), 66.
- Swart, N.C., Cole, J.N., Kharin, V.V., Lazare, M., Scinocca, J.F., Gillett, N.P. ... & Winter, B., 2019. The Canadian earth system model version 5 (CanESM5. 0.3). *Geoscientific Model Development*, 12(11), 4823-4873.
- Swinscow, T.D.V. & Campbell, M.J., 2002. *Statistics at square one* (No. Ed. 10, pp. viii+-158). London: Bmj.
- Tabatabaei, S.M., Dastourani, M., Eslamian, S. & Nazeri Tahroudi, M., 2025. Comparison of kriging methods in rainfall estimation based on entropy-copula (case study: Simineh river, lake Urmia Basin, Iran). *Earth Science Informatics*, 18(1), 75.
- Vijayakumar, S. & Ramaraj, A.P., 2024. CMIP5 multi-model ensemble-based future climate projection for the Odisha state of India. *Current Science (00113891)*, 127(11).
- Vinta, S.R. & Peeriga, R., 2024. Rainfall prediction using xgb model with the australian dataset. *EAI Endorsed Transactions on Energy Web*, 11.
- Yu, H., Nghia, T., Low, B.K.H. & Jaillet, P., 2019, July. Stochastic variational inference for Bayesian sparse Gaussian process regression. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
- Zarei, K., Mohammadi, H. & Bazgeer, S., 2019. Simulation of Gorgan Synoptic Station Temperature and Precipitation with RCP Scenarios. *Physical Geography Research*, 51(4), 563-579.