Development of stepwise m5 tree model to determine the influential factors on rainfall prediction and overcome the greedy problem of its algorithm

Khalil Ghorbani (✉ ghorbani.khalil@gmail.com)

Gorgan University of Agricultural Sciences and Natural Resources

Meysam Salarjazi

Gorgan University of Agricultural Sciences and Natural Resources

Nozar Ghahreman

Tehran University: University of Tehran

Research Article

Keywords:

Posted Date: January 31st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1260445/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Large scale climatic phenomenon which has a delayed effect may be used as important variables for stepwise prediction of rainfall, but the interaction of these signals on the occurrence of rainfall leads to non-linear, complex nature of relations. A model tree is a promising tool for modeling complex systems and recognition of most significant variables. The model tree approach uses a greedy algorithm in which the increased number of variables does not necessarily improve the accuracy of the model, hence the model should be run stepwise. In this study, a stepwise M5 model tree was used in the prediction of annual rainfall in Hashem Abad station, north of Iran, using observed data and 17 climatic signals during the 1985-2019 period to determine the most significant variables. For this purpose, 131017 subsets consisting of 17 members were produced, and the M5 model tree was fitted on each of them. The best combination of variables with the highest accuracy simulated the rainfall with 36mm error (less than 5-6%) and a correlation coefficient of 94%. Among the climatic signals, the Sun Spot (SP) was placed in tree root (most significant), the Nino 4, EA and NAO were ranked as the other significant predictors respectively. The results also indicated that due to the nature of rainfall variations and the greedy algorithm of the M5 model, it is necessary to perform stepwise modeling. The delayed effects of teleconnections may be considered as a suitable feature for early prediction of next year's rainfall and capturing the inter-annual variation.

Introduction

Considering the interannual variations of rainfall and droughts occurrence, early prediction of rainfall is quite important in agricultural water management. Therefore many scientist have attempted to predict rainfall in different time scales. In a group of these studies, the governing statistical rules of time series have been employed. For example, Abdelaziz et.al (2013) used seasonal ARIMA (SARIMA) model for monthly rainfall in Ghana. SARIMA has been also used by Piari et.al (2013) Papacharalampous et.al, Dayal et.al (2019) for monthly precipitation. Wang et al. (2013) predicted seasonal precipitation with the seasonal Arima model (SARIMA) and point out that despite the good fit that the results of the model have with observational precipitation data, if an abnormal change occurs in the previous sequence of the time series, the prediction is accompanied by an error swain et al. (2018) predicted the monthly precipitation in Khorda District, Odisha, India using the ARIMA model. The results show that although the values of R^2 and NSE coefficient of determination were 0.72 and 0.86, respectively, However, should be noted that the model is not able to predict high limit values. Time series models have often been used to predict drought Karimi et.al (2019) used ARIMA model for drought monitoring in Karkheh basin in Iran. The result showed a good agreement between observed and predicted data, and confined the skill off linear stochastic models for prediction of SPI3 time series.

Mishra and Desai (2005) used ARIMA and SARIME models for drought prediction in India. According to their findings, 2 steps forward prediction did agree well with observations and by increasing the time steps, the model precision decreased.

In several other studies, time series models have been compared by data mining techniques. Zhang et.al(2019) compared the drought prediction in Sanjiang plain of China. Their result indicated that ARIMA model perform better with $R^2 = 0.9$ comparing to WNN and SVM with $R^2 = 0.837$ and 0.833 respectively.

Mishra and Desai (2006) compared linear stochastic models (ARIMA / SARIMA), Recursive Multistage Neural Network (RMSNN), and Direct Multistage Neural Network (DMSNN) to predict drought. Models were used to predict droughts based on the Standard Precipitation Index (SPI) in the Indian River Basin. The results showed the superiority of RMSNN models over stochastic and DMSNN models in one-step predictions, but in other time steps DMSNN model was superior to stochastic and RMSNN models.

Dewiredi et.al (2019) compared the simulated values of monthly rainfall obtained by ARIMA an ANN model in Gujarat state, India using 32 years data (1980-2011) for training and 5 years data for testing the model. The obtained correlation coefficient were 0.75 and 0.79 respectively.

Fathabadi et al. (2009) studied the drought condition in 5 stations in northwestern Iran Using the SPI index. The SPI values were predicted using Artificial Neural Network (ANN), time series models, and K-Nearest Neighbors (KNN). The results showed that the time series method is superior to ANN and KNN methods in modeling and predicting SPI values, and SPI is better in 9 and 12-month time scales than SPI in 3 and 6-month time scales.

According to result of Choubin et al. (2018) on rainfall prediction in Kerman province, South Iran using ARIMA, ANFIS models and CART model tree, CART showed better result in compare to other two models.

The selected above-mentioned studies have been performed knowing only the precipitation time series behavior without using any other variable for early prediction of precipitation. Never the less the rainfall occurrence is the result of atmospheric circulations such that an air mass with its specific thermal and moisture condition can determine or change the precipitation of a region. The interactions between these air masses or a teleconnection (climatic signals) are effective on dry/wet spells across the globe.

Masoudian et al. (2009) investigated the relationship between the oscillations of the North Atlantic to the rainfall in Iran. The time series of the North Atlantic Oscillation (NAO) index, for each month was entered separately with the time series of the same month on each of the cells of its digital maps in a linear model. This study showed that Iran is associated with North Atlantic fluctuations in November, March, December, August, and October. In the best condition, a maximum of 35% of the changes in precipitation of some parts of Iran is explained by the fluctuation of the north Atlantic. In October and August, on average, about 12% of changes in precipitation on the Caspian coast are explained by fluctuations in the North Atlantic, but their relationship is direct in October and inverse in August.

Deo et.al (2017) conducted a study for drought forecasting using SOI, PDO, IOD, ENSO and Nino 3, 34 signals, in eastern Australia. The drought modeling was carried out by multivariate adaptive regression spline (MARS), LSSVM and M5 model tree. Comparisons indicated the superior performance of MARS to the other models, with lower MAE. Mulualet al. (2020) used an artificial neural network to predict standardized precipitation-evapotranspiration index in the Nile water basin and concluded that by participating in large-scale climate indicators, accurate results are obtained in predicting the SPEI index.

Kisi et al. (2019) predicted the stream flow pattern in Mediterranean region by incorporating climate signal data, using MARS, M5 model tree and least square support vector machine (LSSVM). The result proved the application of LSSVM model over the others, while integrating climate signals for flow modeling. Choubin et al. (2019) investigated the effect of large-scale climatic signals on snow cover in the Khorasan watershed in Iran by MARS and the M5 Tree method and concluded that the correlation between snow cover and signals in the time delay approach is higher than the concurrency approach. Geopotential height and Nino 1 + 2 are the most important signals related to snow cover. Modeling results also show that the M5 Tree is more useful in areas with more snowfall.

The temporal variability of relation between several climate signals and the Ethiopian kiremt (June-September) season precipitation has been studied by Broan et al. (2019). The findings suggested that temporal changes in seasonal precipitation correlates to sea surface temperature teleconnections, mainly EL-Nino and La-Nina.

Atif et al. (2020) investigated the relationship between hyperlinks and rainfall limit events (EPEs) in Saudi Arabia during the wet season and concluded that the limit events of precipitation are related to the ENSO index so that in the positive (negative) phase of ENSO, the frequency of EPEs in the country increases (decreases). However, El Nino (with CGT positive) amplifies the frequency of EPEs in Saudi Arabia, while the opposite occurs for the La Nina phase (with negative CGT).

Olaya-Arenas et al. (2020) investigated the relationship between long-term precipitation and temperature trends with global atmospheric indices including ENSO, NAO, TNA, and AMO in the Locelo Mountains. They concluded that precipitation and temperature were strongly associated with changes in the tropical Atlantic than with changes in the North Atlantic or Pacific.

By looking at the results, can be seen that the prediction models, which only use precipitation data for one step forward prediction of precipitation or drought (which is common and predicting these conditions is critical in water resources management), are not accurate enough in the event of sudden changes in drought and precipitation. Some studies have shown that these models have lower accuracy in short-term periods than in long-term periods to predict rainfall and drought intensity. One of the reasons for this can be the application of a one-month lag for long-term time scales and the percentage of overlap in calculating the cumulative precipitation or drought index of two consecutive records. For example, in calculating the SPI index in the 12-month time scale for December, precipitation is considered from the beginning of January to the end of December, and for January, precipitation is studied from the beginning of February to the end of January. If the one month time lag is taken into account in their time series, the precipitation of each step is similar to that of the previous step in 11 months. Due to the small share of changes that one-month precipitation has in the total 12-month precipitation, each series itself has a high correlation with the time series of one step ahead. Overall, these models cannot be predicted before the year and have to wait for 11 months and use the data to predict that year's precipitation. However, the use of hyperlink data due to the nature of the delay in precipitation occurrence seems to be a good tool for predicting precipitation. But the question that arises is which variables or hyperbolic indices affect precipitation and whether these indices have a simple linear relationship with precipitation or whether their relationship is more complex and non-linear. Accordingly, the use of gray box modeling methods such as the decision tree model in which the modeling method is transparent can be a suitable method. So far, various researchers have used tree models in predicting precipitation, which is studied by Wei et al. (2020) in seasonal predicting of marginal precipitation in China using climatic indicators, Tong et al. (2019) for the determinants of atmospheric circulation for July-August precipitation in northern China based on tree models and Choubin et al. (2018) in predicting precipitation using CART tree model and comparing it with ANFIS and ARIMA models in Kerman province of Iran. But decision trees use different method for learning, which among them, the Greedy approach is a very popular one. This method divides the datasets into two subsets based on a greedy selection of attributes till reaching a predefined criteria or certain amount (e.g. minimum number of leaves on branch or specific class label in samples). Although these methods usually provide good accuracy but in general, are not reliable sufficiently, therefore application of these complicated trained trees might be difficult, while imposing more limitation on model performance (Verhaeghe et al. 2019).

Considering the summary of the literature review is concluded that precision of early precipitation prediction may be improved by incorporating climate signals and stepwise stochastic modeling involving the lag time and non-linear interactions. Accordingly, the purpose of this study, which has been conducted at Hashem Abad Synoptic Meteorological Station in Gorgan, is to determine the role of teleconnections affecting rainfall and present the best tree model to predict rainfall and prove its greed.

Materials And Methods

Study area and data

The study was performed using the data of Hashem Abad station, located in Gorgan province, north of Iran. The average annual rainfall and temperature of this station are 570 mm 17 °C respectively. The region has a Mediterranean climate according to Mar tone climatic classification. The data used in this research consist of observed monthly rainfall and several climatic signals (listed in Table 1) during 1984-2020. The SOI values were acquired from the National Oceanic and Atmospheric Administration (<https://www.ncdc.noaa.gov>).

Table 1
Teleconnection used in this study

Index	Full Name	Source
NAO	North Atlantic Oscillation	https://www.ncdc.noaa.gov/teleconnections/nao/data.csv
EA	East Atlantic Pattern	
WP	West Pacific	
PNA	Pacific North American	
EA/WR	East Atlantic/West Russia Pattern	
SCA	Scandinavia Pattern	
POL	Polar/ Eurasia Pattern	
Expl.Var	Explained Variance (%) of leading 10 modes	
NINO12	Nino SST Indices	https://ftp.cpc.ncep.noaa.gov/wd52dg/data/indices/sstoi.indices
NINO3		
NINO4		
NINO3.4		
AO	Arctic Oscillation	https://www.ncdc.noaa.gov/teleconnections/ao/data.csv
SP	Sun Spot	http://www.sidc.be/silso/datafiles
OLR	Outgoing Longwave Radiation	https://www.cpc.ncep.noaa.gov/data/indices/olr
ENSO	El Nino Southern Oscillation	https://www.esrl.noaa.gov/psd/enso/mei/
SOI	Southern Oscillation Index	https://ftp.cpc.ncep.noaa.gov/wd52dg/data/indices/soi

After retrieving the data and quality control, a matrix consisting of 17 independent variables of mean annual values of signals with a one-year time lag as predictors and annual rainfall as the dependent variable was formed and used as input of the model.

Decision tree

The decision tree approach is a well-known method for exploring the rules to determine a value or rank. This method can be applied to both continuous (regression) and discrete (classification) data. The M5 model tree was first proposed by Quinlan (1992) for predicting the values of continuous series. This model tree calculates the SDR (Standard Deviation Reduction) index for each attribute (independent variable) and considers the variable with the highest SDR in tree root (as the most significant one), and the rest of the variables are ranked in different branches or subclasses respectively.

The SDR is calculated for each attribute in all subclasses interactively and continues till the difference between the prediction and observed value (error) does not change significantly. Nevertheless, this method has some limitations such as branching and the need for pruning, keeping the error in an acceptable range.

$$SDR(T, X) = S(T) - S(T, X)$$

1

In equation 1, S(T) is the standard deviation of the total dependent variable, and S(T, X) is the weighted average of the standard deviation of the dependent variable, which is divided into two categories less and more than X. Standard deviation indicates the degree of data scatters from the mean.

A stepwise implementation of the M5 tree model:

In this study, the M5 tree model was implemented in the Rapid Miner software environment. Because the stepwise implementation of this model is not as default in this software, an RMP model was designed according to Figures 1 to 3. In this RMP model, first, the data matrix stored in an Excel file is called, and the data is entered into a loop to generate different subsets of multiple members (the number of members of each subset can be determined). The M5 tree model is then executed for each subset and the model evaluation criteria including their correlation coefficient and RMSE are calculated and stored in a text file.

Model performance evaluation criteria:

The correlation coefficient (r) and the mean square root of the error (RMSE) were used to show the degree of correlation between predicted and observational values, and to indicate the magnitude of the error, respectively. Also, the absolute value of percentage error of each sample was calculated, and their values were plotted against the percentage of samples to determine the prediction error in what percentage of samples is more or less than one value.

$$r = \frac{n(\sum_{i=1}^n O_i \cdot P_i) - (\sum_{i=1}^n O_i)(\sum_{i=1}^n P_i)}{\sqrt{\left[n\sum_{i=1}^n O_i^2 - \left(\sum_{i=1}^n O_i\right)^2\right] \cdot \left[n\sum_{i=1}^n P_i^2 - \left(\sum_{i=1}^n P_i\right)^2\right]}}$$

2

$$RMSE = \frac{\sqrt{\sum_{i=1}^n (O_i - P_i)^2}}{n}$$

3

Pi and Oi are predicted and observed precipitation values and n is the number of samples, respectively.

Results

In order to show the ability of M5 model tree in modeling of non-linear equations, its outputs were compared with multivariate regression model (MKK).

Results of multivariate regression in prediction of rainfall

First, the linear relation between rainfall and possible affecting variables was investigated. The correlation matrix was created to show the existing correlation between teleconnections and rainfall. The highest correlation of rainfall and climatic signals did not exceed 0.23, but a strong correlation between some of these signals with each other and with OLR was observed.

Table 2
The matrix of correlation percentage

	P	NAO	NINO12	NINO3	NINO4	NINO3.4	AO	SP	OLR	ENSO	SOI	EA	WP	PNA	EA/WR	SCA	POL
NAO	-4																
NINO12	11	-14															
NINO3	19	-5	87														
NINO4	14	13	47	81													
NINO3.4	21	4	70	95	93												
AO	-14	73	-5	-11	-5	-11											
SP	-2	34	-10	-10	-7	-9	46										
OLR	-23	-10	-52	-80	-87	-90	4	-8									
ENSO	23	10	70	89	86	94	-7	-3	-93								
SOI	-23	-17	-58	-82	-87	-91	3	1	92	-97							
EA	-14	-13	27	32	29	28	-6	-9	-10	13	-8						
WP	19	11	29	22	4	17	10	15	-23	28	-24	-3					
PNA	15	-29	34	36	22	31	-34	-26	-26	28	-22	20	17				
EA/WR	19	21	23	25	21	28	16	14	-45	44	-48	-37	17	-5			
SCA	3	-20	-24	-15	-6	-11	-47	-18	9	-11	8	-9	-39	-10	-1		
POL	-11	18	5	10	13	15	33	29	-24	19	-22	-41	-16	-6	42	5	
Expl.Var	-11	4	46	41	30	39	-1	6	-40	50	-43	12	26	14	23	-19	18

The result of stepwise execution of multivariate linear regression showed the values of 0.31 and 80 for R² and RMSE, respectively (Table 3). Also, the range of error was 0.7 to 247 and the standard deviation was found to be 80 mm/Year Nevertheless, the annual range of observed rainfall is 207 to 806 mm, but the predicted annual range was 360 to 655 mm (Fig. 4), which indicates that MVR cannot predict wet and dry years precisely. In general, the MVR failed to detect any statistically significant relationship between variables, hence, it was also necessary to evaluate the non-linear equations.

Table 3
The best selected sub sets of predict ants in linear multivariate regression.

Vars	R ²	NAO	NINO12	NINO3	NINO4	NINO3.4	AO	SP	OLR	ENSO	SOI	EA	WP	PNA	EA/WR	SCA	POL	Expl.Var
1	5										X							
2	13									X								X
3	17				X					X								X
4	20									X		X					X	X
5	23		X		X	X						X					X	
6	27			X	X	X						X					X	X
7	27			X	X	X				X		X					X	X
8	28			X	X	X		X				X	X				X	X
9	29			X	X	X		X				X	X	X			X	X
10	29			X	X	X		X	X			X	X	X			X	X
11	30			X	X	X		X	X	X		X	X	X			X	X
12	30	X		X	X	X		X	X	X		X	X	X			X	X
13	30	X		X	X	X	X	X	X	X		X	X	X			X	X
14	31	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X
15	31	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X
16	31	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
17	31	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Result of stepwise M5 model tree

The best subset with the highest accuracy was selected, after the subsets with 1 to 17 members of independent variables (Total 131017) were formed, and a tree model was created for each one. The obtained results are as follows: For the single-member subsets, the best tree model was determined using EXP.VAR index and RMSE values of 0.44 and 86mm respectively. SP index was ranked as the second (Table 4).

Table4. The results of rainfall prediction using M5 model tree for single member subsets.

used attributes	Expl.Var	SP	NINO3.4	SOIs	NINO3	ENSO	NINO4	EA	OLR	NINO12	EA/WR	WP	POL	PNA	AO	NAO	SCA
R (%)	44	38	31	28	26	26	26	24	23	21	19	19	18	15	14	00	00
RMSE (mm)	86	89	91	92	92	92	92	93	93	94	94	94	94	94	95	96	96

The M5 model tree results for 1 to 17 member subsets showed that using more than 10 variables decrease the accuracy of the model. Besides, in each multimember subset there exist a significant difference among the members with respect to the accuracy, such that in 24310 subsets with 9 members, correlation coefficient varies between 28.5 and 94% and prediction error between 362 to 91.6mm. (Table 5).

Table5. The statistical metrics of M5 tree model execution of different subsets

Subset	Number of Subset	R min (%)	R max (%)	RMSE min (mm)	RMSE max (mm)
1	17	0	44.4	86	95.6
2	136	0	74.5	65.6	95.6
3	680	15.3	84.3	53.7	94.4
4	2380	19.2	89.8	45.6	93.8
5	6188	18.7	90.4	44.3	93.9
6	12376	22.8	92.2	40.8	93
7	19448	25.2	92.7	39.6	92.5
8	24310	28.1	93.4	37.8	91.7
9	24310	28.5	94	36.2	91.6
10	19448	31.5	93.5	36.2	90.7
11	12376	32.2	93.7	36.7	90.5
12	6188	37.8	93.8	36.4	88.4
13	2380	74.7	93.7	36.3	64
14	680	75.7	93.8	36.6	62.8
15	136	84	93.7	36.8	52.2
16	17	88.7	92.9	38.1	45.5
17	1	89.4	89.4	44.2	44.2

In Table 6, the top selected 9 member combination are shown. It should be mention that NAO, NINP12, EA and SP are present in all combinations.

Table 6
The top 10 combination of 9 member subsets

used attributes	R (%)	RMSE (mm)
NAO, NINO12, NINO4, AO, SP, EA, WP, PNA, POL	94.0	36.2
NAO, NINO12, NINO4, AO, SP, EA, WP, SCA, POL	93.7	36.6
NAO, NINO12, NINO3, NINO4, AO, SP, EA, SCA, POL	93.4	37.2
NAO, NINO12, NINO4, AO, SP, OLR, ENSO, EA, POL	93.1	38.5
NAO, NINO12, NINO4, SP, OLR, SOIs, EA, WP, PNA	93.1	39.1
NAO, NINO12, NINO3, SP, OLR, SOIs, EA, WP, PNA	92.9	39.5
NAO, NINO12, AO, SP, OLR, ENSO, SOIs, EA, POL	92.9	39.2
NAO, NINO12, NINO3, NINO4, AO, SP, EA, WP, PNA	92.8	39.0
NAO, NINO12, NINO4, SP, OLR, SOIs, EA, WP, SCA	92.8	39.6
NAO, NINO12, NINO4, NINO3.4, SP, OLR, SOIs, EA, SCA	92.7	39.6

Even if 10 combinations of 9 variables with the lowest accuracy are considered, non-presence of EA, Expl.Var and SP in these combinations are significant. Both SP and EA indices have increased the model's accuracy which confirms their key role in the prediction of rainfall in the study region (Table 7).

Table 7
The results of model tree with 9 variables in 10 combinations with lowest accuracy

used attributes	R (%)	RMSE (mm)
NAO, NINO3, NINO4, NINO3.4, OLR, SOIs, PNA, EA/WR, SCA	31.1	90.8
NAO, NINO3, AO, OLR, ENSO, SOIs, WP, PNA, SCA	31.0	90.9
NAO, NINO3.4, AO, OLR, ENSO, SOIs, WP, PNA, SCA	31.0	90.9
NAO, NINO3, NINO3.4, AO, OLR, ENSO, SOIs, SCA, POL	30.9	90.9
NAO, NINO3, NINO4, NINO3.4, OLR, ENSO, SOIs, WP, PNA	30.3	91.1
NAO, NINO12, NINO3, NINO3.4, AO, OLR, ENSO, SOIs, PNA	29.7	91.3
NINO12, NINO3, NINO3.4, AO, OLR, ENSO, SOIs, PNA, SCA	29.7	91.3
NAO, NINO12, NINO3, NINO3.4, AO, OLR, ENSO, PNA, SCA	29.4	91.3
NAO, NINO12, NINO3, NINO3.4, AO, OLR, ENSO, SOIs, SCA	28.6	91.6
NAO, NINO3, NINO3.4, AO, OLR, ENSO, SOIs, PNA, SCA	28.5	91.6

These teleconnections involved in regression equations of M5 model tree have different weight, which indicates that effect magnitude of each teleconnection on rainfall vary significantly and might be even reversed (Table 8). M5 model tree not only increases the rainfall prediction comparing to multivariate regression but also improves the accuracy in severe wet and dry spells (Droughts) (Fig. 5 and 6).

The error percentage graph shows that about 90% of samples have an error less than 12%.

Determination of variables effecting rainfall variation using cross validation

To determine the significance of variables the best subset with 1 to 9 members was prepared (Table 9). The main point was the involvement of EA, SP, and NINO4 variables as the most significant in all combinations. The NINO12 variable permanently exists in combinations with more Than 5 Variables. To select the best subset, in addition to the accuracy of the model, the number of rules and equations used for prediction are also important. Such that in the current study, the model consisting of 9 variables with 20 rules had the highest accuracy, and the model accuracy decreased by increasing the number of variables.

Table 9
The best 1 to 9 variable sub sets of M5 model tree

Used_Attributes	Number Attribute	R (%)	R ² (%)	RMSE (mm)	Numberof Rules
Expl.Var	1	44.4	19.7	86.0	4
NINO3, SP	2	74.5	55.5	65.6	12
NINO4, SP, EA	3	84.3	71.1	53.7	17
NAO, NINO4, SP, EA	4	89.8	80.6	45.6	27
NAO, NINO4, SP, EA, WP	5	90.4	81.8	44.3	41
NAO, NINO12, NINO4, SP, EA, WP	6	92.2	85.0	40.8	27
NAO, NINO12, NINO4, AO, SP, EA, PNA	7	92.7	86.0	39.9	36
NAO, NINO12, NINO4, AO, SP, EA, WP, PNA	8	93.4	87.3	38.0	34
NAO, NINO12, NINO4, AO, SP, EA, WP, PNA, POL	9	94.0	88.4	36.2	20

The cross-validation method was used to investigate the effect of variables on model tree accuracy in rainfall prediction. For this purpose, each variable is removed in a substitution manner, then the error is calculated, and the variable is replaced again. Elimination of single variables in 3 to 9 variables tree models (Tables 10-16) showed the significant role of the sunspot variable in rainfall prediction that by removing this variable from the tree model, the amount of predicted error increased more than removing other variables. In 9 variables model tree, the SP variables produced the highest SDR value among the independent variables and were placed in tree root; in some sub-branches, SP variables caused another branching. These sub-branches in the M5 model tree produced more Homogenous subsets with more precise regression equations.

Table 10
Evaluation of M5 Model tree accuracy by eliminating one variables in 3 member subsets

	Variables	R %	RMSE(mm)
Best Variables	NINO4, SP, EA	84.3	53.745
Missed Variable	NINO4	59	77.529
	SP	44.9	85.697
	EA	57.6	78.399

In the best 4 variables model tree, NINO 4 index was also significant and if deleted, the rainfall prediction error would increase up to 20 mm per year.

Table 11
Evaluation of M5 Model tree accuracy by eliminating one variables in 4 member subsets

	Variables	R %	RMSE(mm)
Best Variables	NAO, NINO4, SP, EA	89.8	45.611
Missed Variable	NAO	84.3	53.745
	NINO4	74.9	65.16
	SP	51.9	81.942
	EA	81.9	57.915

In five variables model tree, non-inclusion of SP index, increases the prediction up to 2 times, viz 45 mm per year comparing to 13mm caused by deleting other variables

Table 12
Evaluation of M5 Model tree accuracy by eliminating one variables in 5 member subsets

	Variables	R %	RMSE(mm)
All Variables	NAO, NINO4, SP, EA, WP	89.8	45.611
Missed Variable	NAO	84.8	52.888
	NINO4	82.1	57.974
	SP	31.5	90.72
	EA	81.5	57.836
	WP	89.8	45.611

In the best six variables model tree deleting the NINO12 index had the lowest effect in decreasing the accuracy and led to a 3 mm prediction error.

Table 13
Evaluation of M5 Model tree accuracy by eliminating one variables in 6 member subsets

	Variables	R %	RMSE (mm)
All Variables	NAO, NINO12, NINO4, SP, EA, WP	92.2	40.847
Missed Variable	NAO	88.5	47.925
	NINO12	90.4	44.309
	NINO4	84.7	54.463
	SP	84.2	55.606
	EA	82.7	56.449
	WP	88.5	47.245

Again, in the seven variables model tree, deleting the SP reduces the accuracy. Although this much reduction was not observed in the six variables tree, and confirms the greedy nature of this model algorithm and indicates that adding/removing one variable can change the model structure significantly. Hence, the variables should be selected with caution and run the model stepwise.

Table 14
Evaluation of M5 Model tree accuracy by eliminating one variables in 7 member subsets

	Variables	R %	RMSE(mm)
All Variables	NAO, NINO12, NINO4, AO, SP, EA, PNA	92.7	39.881
Missed Variable	NAO	89.7	45.348
	NINO12	89.7	45.043
	NINO4	87	51.392
	AO	91.7	42.116
	SP	30.2	91.11
	EA	88.4	49.247
	PNA	91.3	42.526

In the best eight variables model tree, removing the SP index may increase the prediction error up to 17.2 mm/year and is quite more than other variables removal error (Table 15).

Table 15
Evaluation of M5 Model tree accuracy by eliminating one variables in 8 member subsets

	Variables	R %	RMSE(mm)
All Variables	NAO, NINO12, NINO4, AO, SP, EA, WP, PNA	93.4	38.021
Missed Variable	NAO	90.5	43.721
	NINO12	91.4	41.886
	NINO4	87.6	50.425
	AO	92	40.963
	SP	84.8	55.209
	EA	88.8	48.11
	WP	92.7	39.881
	PNA	92.7	39.621

Finally, in the case of the best nine variables model tree, removing the SP caused the highest error, in such that the prediction errors were increased from 36.2 to 52.02 mm per year.

Table 16
Evaluation of M5 Model tree accuracy by eliminating one variables in 9 member subsets

	Variables	R %	RMSE(mm)
All Variables	NAO, NINO12, NINO4, AO, SP, EA, WP, PNA, POL	94	36.204
Missed Variable	NAO	91.2	41.931
	NINO12	92.8	38.453
	NINO4	87.5	49.391
	AO	92.1	40.072
	SP	86.2	52.05
	EA	89.4	46
	WP	93.3	37.969
	PNA	92.6	38.887
	POL	93.4	38.018

The effect of sunspots on precipitation has been proven by many researchers. Tripathi, and Bhattacharya (2014) found a strong correlation between the average annual rainfall and the number of sunspots in India. Their results showed a decline in both the 10-year moving average of India's annual rainfall and the 10-year moving average of sunspots since 1996. Wang and Zhao (2012) Based on statistical tests, they examined the correlation between decadal variation in June precipitation in China and the number of sunspots and found a positive and reliable correlation between the 11-year cycle of sunspots and June precipitation, while the correlation was very weak in South of the Yangtze River region and the Hatao Basin. Jahanbakhsh and Edalatdoust (2008) investigated the role of sunspots on variation in annual precipitation in seven selected stations of Iran, using wavelet analysis and Fourier analysis, and

statistical analysis. They found that the correlation between the precipitation variation and sunspots variation is not constant in all parts of Iran. Precipitation variations in some coastal areas of northern Iran follow the cycle of solar activity while inland and land areas have the opposite reaction. The existence of alternating and consecutive droughts and wet years in accordance with the sunspots cycle indicates that the sun and its activities are not only a determining factor and are influencing the precipitation variation worldwide but also has a significant role in precipitation variation in Iran.

The results of this study with the use of a new method in modeling determined the role of sunspots on rainfall variation in one of the northern cities of Iran. Also, the simultaneous effect of several teleconnection indices on rainfall variation and their collective use in rainfall prediction models was another common result reached by other researchers.

Summary Of Results And Dissuasion

1. Each of the studied teleconnection signals did not show a significant correlation with annual rainfall either using simple regression or the M5 model tree. Therefore it is concluded that correlating the rainfall of a specific region to a single climatic signal is not valid, and a combination of signals affects the changes in annual rainfall.
2. The relationship between rainfall variation and teleconnections is not linear. The multivariate regression model MLR error exceeded 80 mm per year with a regression coefficient less than 0.5 and failed to predict the wet and dry (drought) year accurately while predicting these conditions in the management of agriculture and water resources are of greater importance.
3. Among the 17 selected teleconnections, as the predictors, the highest accuracy was obtained by a nine-member subset. This confirmed that using more than nine variables does not increase the prediction precision, which indicates the greedy nature of the M5 model tree algorithm. Also, it is necessary to execute this model stepwise.
4. The root variable of the M5 model tree was the number of sunspots, which led to more homogenous classification and higher accuracy. Non-inclusion of SP index in model tree increased the prediction error.
5. In the M5 model tree regression equations, indices had different weights, even with reverse effects. Therefore despite a weak correlation, a single variable cannot increase/decrease the Gorgan region rainfall, and a combination of indices should be considered. Although, the sign of the NAO index in all equations was positive, and EA and POL indices were negative.
6. In years with a higher number of sunspots, in the equation of predicting next year's rainfall using the M5 model tree, the weight of the SP index would also increase. Considering the negative sign of the SP index in equations, increased solar activities would decrease the rainfall in the study region.
7. In years with a lower number of sunspots, the NINO4 index has a high weight in the equations of the M5 model tree and vice versa, such that the weight may decrease from 102 to 1.2 in years with a higher rate of solar activity, which is significant. Therefore, in a year with low solar activity, the NINO4 index plays an important role in rainfall variations, and the increase of this index may significantly increase rainfall in Gorgan city.
8. The teleconnections with delayed effect on Gorgan region rainfall can serve as a prediction tool for next year's rainfall, and help the decision-makers for water resource management.

Declarations

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

There are no competing financial interests.

Author Contributions

Khalil Ghorbani: Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Writing - original draft, Writing - review & editing. Authorship.

Meysam Salarijazi: Validation, Review & Editing, Software.

Nozar Ghahraman: Review and translation.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethical Approval

We follow the principles of ethics, as it is mentioned, in conducting and publishing this research.

Consent to Participate

The authors are consented to participate in the research.

Consent for Publication

The authors are contented with the publication of the results.

Funding

The manuscript is extracted from a project supported by Gorgan University Agricultural Sciences and Natural Resources, Gorgan, Iran.

Acknowledgment

This manuscript is extracted from a project at the Gorgan University Agricultural Sciences and Natural Resources, Gorgan, Iran. The authors are grateful to the University for providing the conditions for conducting this research.

References

1. Atif RM, Almazroui M, Saeed S, Abid MA, Islam MN, Ismail M (2020) Extreme precipitation events over Saudi Arabia during the wet season and their associated teleconnections. *Atmos Res* 231:104655
2. Broman D, Rajagopalan B, Hopson T, Gebremichael M (2020) Spatial and temporal variability of East African Kiremt season precipitation and large-scale teleconnections. *Int J Climatol* 40(2):1241–1254
3. Choubin B, Roshan H, Sajedi-Hosseini F, Rahmati O, Melesse AM, Singh VP (2019) Effects of large-scale climate signals on snow cover in Khersan watershed, Iran. *Extreme Hydrology and Climate Variability*. Elsevier, pp 1–10
4. Choubin B, Zehtabian G, Azareh A, Rafiei-Sardooi E, Sajedi-Hosseini F, Kişi Ö (2018) Precipitation forecasting using classification and regression trees (CART) model: a comparative study of different approaches. *Environ Earth Sci* 77(8):314
5. Dayal D, Swain S, Gautam AK, Palmate SS, Pandey A, Mishra SK (2019) May). Development of ARIMA Model for Monthly Rainfall Forecasting over an Indian River Basin. *World Environmental and Water Resources Congress 2019: Watershed Management, Irrigation and Drainage, and Water Resources Planning and Management*. American Society of Civil Engineers, Reston, VA, pp 264–271
6. Deo RC, Kisi O, Singh VP (2017) Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. *Atmos Res* 184:149–175
7. Dwivedi DK, Kelaiya JH, Sharma GR (2019) Forecasting monthly rainfall using autoregressive integrated moving average model (ARIMA) and artificial neural network (ANN) model: A case study of Junagadh, Gujarat, India. *Journal of Applied and Natural Science* 11(1):35–41
8. Fathabadi A, Gholami H, Salajeghe A, Azanivand H, Khosravi H (2009) Drought forecasting using neural network and stochastic models. *American Eurasian Network for Scientific Information* 3(2):137–146
9. Jahanbakhsh S, Edalatdoust M (2008) The effect of solar activity on the variation of annual precipitation in Iran. *Geographical Research Journal*, 2008; 23 (1):3-24. (In Persian)
10. Karimi M, Melesse AM, Khosravi K, Mamuye M, Zhang J (2019) Analysis and prediction of meteorological drought using SPI index and ARIMA model in the Karkheh River Basin, Iran. *Extreme Hydrology and Climate Variability*. Elsevier, pp 343–353
11. Kisi O, Choubin B, Deo RC, Yaseen ZM (2019) Incorporating synoptic-scale climate signals for streamflow modelling over the Mediterranean region using machine learning models. *Hydrol Sci J* 64(10):1240–1252
12. Masoudian SAAF (2009) on relationship between precipitation of Iran and North Atlantic Oscillation. *Geographical Research Journal* 23(4):3–18
13. Mishra AK, Desai VR (2006) Drought forecasting using feed-forward recursive neural network. *Ecol Model* 198(1–2):127–138
14. Mishra AK, Desai VR (2005) Drought forecasting using stochastic models. *Stoch Environ Res Assess* 19:326–339
15. Mulua GM, Liou YA (2020) Application of Artificial Neural Networks in Forecasting a Standardized Precipitation Evapotranspiration Index for the Upper Blue Nile Basin. *Water* 12(3):643
16. Olaya-Arenas P, Meléndez-Ackerman EJ, Pérez ME (2020) Long-term temperature and precipitation trends in the Luquillo Mountains, and their relationships to global atmospheric indices used in climate change predictions. *Caribb J Sci* 50(1):107–131
17. Quinlan JR (1992) November. Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343-348)
18. Swain S, Nandi S, Patel P (2018) Development of an ARIMA model for monthly rainfall forecasting over Khordha district, Odisha, India. *Recent Findings in Intelligent Computing Techniques*. Springer, Singapore, pp 325–331
19. Tong X, Yan Z, Xia J, Lou X (2019) Decisive atmospheric circulation indices for July–August precipitation in North China based on tree models. *J Hydrometeorol* 20(8):1707–1720

20. Tripathi DK, Bhattacharya AB (2014) Sunspot Activity Over the Indian Rainfall Pattern. International Journal of Electronics & Communication Technology 5(2):43–44
21. Verhaeghe H, Nijssen S, Pesant G, Quimper CG, Schaus P (2019) Learning optimal decision trees using constraint programming. In The 25th International Conference on Principles and Practice of Constraint Programming (CP2019)
22. Wang JS, Zhao L (2012) Statistical tests for a correlation between decadal variation in June precipitation in China and sunspot number. Journal of Geophysical Research: Atmospheres 117:D23
23. Wei W, Yan Z, Jones PD (2020) A decision-tree approach to seasonal prediction of extreme precipitation in eastern China. Int J Climatol 40(1):255–272
24. Zhang Y, Yang H, Cui H, Chen Q (2019) Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China. Natural Resources Research, pp 1–18

Figures

Figure 1

Reading the data matrix from an excel file and entering them into the subsets loop

Figure 2

Implementation of the M5 tree model for each of the subsets and evaluating the model results

Figure 3

Saving the results of the evaluation of model implementation on every subset

Figure 4
Scatter diagram of observed and predicted rainfall by linear multivariate regression

Figure 5
Scatter diagram of observed and simulated rainfall

Figure 6

The predicted values for the last three study years 2016-1019 (Right) and error percentage vs. samples percentage (Left)