Water Resources Management (2023) 37:5871–5891
https://doi.org/10.1007/s11269-023-03563-4

Evaluation of Feature Selection Methods in Estimation of Precipitation Based on Deep Learning Artificial Neural Networks

Mohammad Taghi Sattari[1,3] · Anca Avram[2] · Halit Apaydin[3] · Oliviu Matei[2]

Received: 27 December 2022 / Accepted: 20 July 2023 / Published online: 18 November 2023

Abstract
Precipitation is the most important element of the water cycle and an indispensable element of water resources management. This paper's aim is to model the monthly precipitation in 8 precipitation observation stations in the province of Hamadan, Iran. The effects and role of different feature weights pre-processing methods (Weight by deviation, Weight by PCA, Weight by correlation and Weight by Support Vector Machine) on artificial intelligence modeling were investigated. Deep learning method based on a multi-layer feed-forward artificial neural network that is trained with Stochastic Gradient Descent using back-propagation (DL-SGD) and Convolutional Neural Networks (CNN) modelling were applied. The precipitation of each station is modeled using the precipitation values of the other stations. The best result, among all scenarios, at the Vasaj station according to the DL-SGD method (CC = 0.9845, NS = 0.9543 and RMSE = 10.4169 mm) and at the Varayineh station according to the CNN method (CC = 0.9679, NS = 0.9362 and RMSE = 16.0988 mm) were estimated.

Keywords Precipitation · Artificial intelligence · Feature selection · Deep learning · Stochastic gradient descent · Feature weights · H_2O cluster

✉ Mohammad Taghi Sattari
mtsattar@tabrizu.ac.ir

✉ Oliviu Matei
oliviu.matei@ieec.utcluj.ro

Anca Avram
anca.avram@ieee.org

Halit Apaydin
apaydin@ankara.edu.tr

1 Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran

2 Department of Electrical Engineering, Electronics and Computer Science, Technical University of Cluj-Napoca, North University Center of Baia Mare, 400114 Cluj-Napoca, Romania

3 Department of Agricultural Engineering, Faculty of Agriculture, Ankara University, 06110 Ankara, Turkey

1 Introduction

Water resources fed by precipitation are globally becoming increasingly important, in terms of both environment and socio-economic aspects. Global warming, climate change and droughts, on the one hand, and unsuccessful water resources management policies, on the other hand, adversely affect the environment and people's lives (Sattari et al. 2012; Inan 2021). Precipitation has a random and stochastic nature. Precipitation is in a complex and non-linear relationship with other meteorological variables (Zhu et al. 2022). As a matter of fact, precipitation levels in a basin and neighbouring precipitation stations in the same climate are also in connection with each other. There are different statistical models for the estimation of precipitation levels. However, in recent years, it is seen that data-based models that can make predictions by using only observed data from past periods, without caring about the physics of the event, have expanded in all branches of science. Measured or observed historical data are the basis of models such as data mining, data-driven models, black box, artificial intelligence, machine learning and deep learning.

Various studies use data-driven methods in modeling precipitation (Xie et al. 2021). Nourani et al. (2017) estimated the maximum precipitation in Iran using decision tree and association rules models. In this study, along with meteorological variables, the Black, Mediterranean and Red Seas, sea surface temperature values were used as model inputs. The estimated measures attest to the suggested hybrid data mining method's dependability in predicting extreme precipitation events while taking higher threshold values into account. Chen et al. (2020) used the Deep Learning-based multilayer perceptron (MLP) method to predict precipitation. In the study, high quality precipitation products from the ground radar network were used as target tags to train the MLP model. Apaydin and Sattari (2020) used a hybrid approach of deep learning and geographic information system based on spatio-temporal variables to model the amount of precipitation on the Turkish coastline. Spatial variables such as latitude, longitude, altitude, distance to the sea and the change in monthly precipitation were considered as temporal variables. Yan et al. (2021) successfully predicted precipitation using 5-year meteorological data from 26 stations in the Beijing-Tianjin-Hebei region of China, with a new Attentive Interpretable Tabular Learning Neural network (TabNet) approach using machine learning to monitor and method precipitation with the help of satellite system. Kassem and Gökçeku (2021) estimated monthly precipitation in Nigeria using artificial intelligence and mathematical models. They used a multilayer feedforward neural network, graded feedforward neural network and radial basis neural network models together with quadratic and Poisson regression models.

With the help of cutting-edge soft computing methods including multivariate adaptive regression splines (MARS), classification and regression trees (CART), and gene expression programming (GEP), Chaplot (2021) sought to concentrate on the rainfall prediction in India's Udaipur district, which receives 631 mm of annual rainfall. Diriba and Debusho (2021) used a multivariate conditional modeling approach to analyze the behavior of common precipitation events in South Africa to explore the co-dependence of extreme precipitation events. Vathsala and Koolagudi (2021) used a combination of data mining and neuro-fuzzy inference system to predict precipitation in India. Precipitation levels were successfully estimated with the if–then rules derived according to the proposed method in the study. Ahi et al. (2023) aimed to predict evaporation in the Karaidemir Reservoir in Turkey using ANN and 30 years of daily meteorological data. Afshari Nia et al. (2023) combined the deep learning model with ANN to predict monthly precipitation in the Kashan plain of Iran. Three different models were developed to compare the results of the study. That the literature review showcases how artificial intelligence and machine learning techniques consistently make effective and accurate predictions worldwide (Lin et al. 2021; Ng et al. 2021; Tabatabaei et al. 2021; Roslan et al. 2021).

The aim of this research is modelling the monthly precipitation of the neighbouring station in the Hamadan region by deep learning method based on a multi-layer feed-forward artificial neural network that is trained with Stochastic Gradient Descent using back-propagation (DL-SGD) and Convolutional Neural Networks (CNN). The methods used in the study to forecast precipitation at the neighbouring stations are well-known, but it is unclear which stations should be included in the model to estimate the data of each station. This study considered how the weights should be taken into account in order to use the least number of neighboring station data. The novelty of the research consists in applying the Simple Additive Weighted (SAW) technique, then designing and applying three types of scenarios when choosing the correlated sources based on the computed aggregated weights (AW): (1) ALL scenario – no selection of correlated sources, (2) W1 scenario – all sources having AW > 1, (3) W2 scenario – all sources having AW > 2. In addition, our study seeks to improve the precipitation forecast by using data from fewer stations by using four different feature weights pre-processing (Weight by deviation, Weight by PCA, Weight by correlation and Weight by SVM) method and to investigate the effects of stations on one another.

2 Material and Methods

2.1 Case Study

Hamadan province is located in the Western part of Iran and is characterized by its mountainous topography. The region has moderate temperatures in summer and very cold in winter (Baaghideh et al. 2017). The input data came from 8 stations: Varayineh, Kheyirabad, Sarabi, Khosroabad, Namileh, Pihan, Vasaj (Fig. 1). For each of these locations the information available was the value for the rainfall. The available data was collected for 300 months. The monthly precipitation data of 8 precipitation observation stations between 1995–2019 water year were used. The geographical location of these stations and the statistical properties of precipitation amounts are given in Table 1.

As can be seen in Table 1, Varayineh is the station with the highest rainfall with a mean annual precipitation of 557.64 mm (monthly mean average of 46.47 mm) and Namileh has the least mean annual precipitation with 322.44 mm (26.87 mm monthly mean precipitation). At the same time, according to Table 1, it is seen that the study area is rugged and mountainous with an altitude of at least 1525 m (Khosroabad) and at most 1925 m (Sarabi).

When the monthly precipitation levels are examined (Fig. 2 and Table 2), it is seen that the averages fall below 50 mm, and the extreme precipitations levels are below 200 mm with the exception of Varayineh and Babapirali stations. Unprecedented precipitation occurred in the region in April 2019 (295th month) in last 70 years (Sinamet 2018), and the highest values observed in all stations were measured in this event. The driest months in the region are July–August with the wettest occurring in April.

According to Fig. 3, where the precipitation time series is given, the water years of 1999 and 2008 are the driest; The water years of 1995, 2016 and 2019 were the wettest periods. Varayineh was the station with the most rainfall, while Namileh and Pihan had the least rainfall.

Fig. 1 Study locations

2.2 Methods

The purpose of the research is to determine the best approach for predicting the rainfall of a specific location, when also having the input of other stations in the proximity. Various combinations and scenarios have been determined using different model inputs. In

Table 1 Statistics of yearly and monthly precipitation data between October 1994 and September 2019 and station locations

		Babapirali	Kheyirabad	Khosroabad	Namileh	Pihan	Sarabi	Varayineh	Vasaj
Monthly	Minimum (mm)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Maximum (mm)	278.0	306.0	277.0	276.0	246.5	302.0	320.0	276.0
	Mean (mm)	38.15	32.49	32.08	26.87	27.13	37.21	46.47	34.16
	Stdev (mm)	47.79	40.46	40.83	34.12	34.71	44.44	55.14	41.21
Yearly	Minimum (mm)	251.5	223.5	163.0	126.5	173.0	228.0	155.3	292.5
	Maximum (mm)	739.0	705.0	711.0	598.0	626.0	786.5	705.0	891.0
	Mean (mm)	457.8	389.8	385.0	321.3	324.5	446.6	409.9	557.6
	Stdev (mm)	136.6	116.8	129.8	98.8	98.2	130.1	123.5	141.3
Longitude (^{0})		34.63	34.25	35.51	34.27	34.13	34.53	34.08	34.32
Latitude (^{0})		48.35	49.02	49.31	48.82	48.87	48.47	48.40	48.23
Elevation (m)		1917	1770	1525	1773	1870	1925	1795	1566

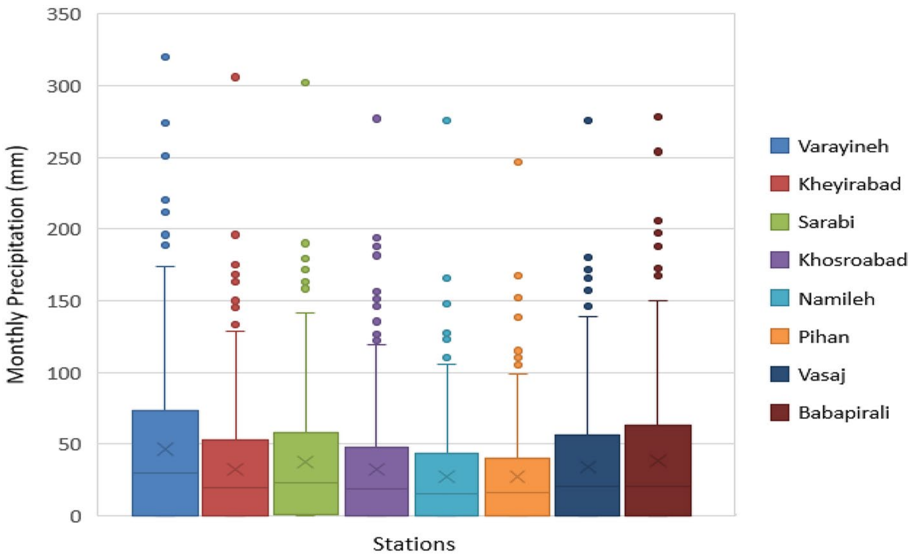

Fig. 2 Distribution of monthly precipitation and extreme values

addition, various feature weights were taken into account in the selection of scenarios. To compute the relevance of the attributes (correlated sources) for each location, additional methods were used, described as follows.

2.2.1 Feature Weights, Selection and Scenarios

Weight by Deviation Computes the weight of attributes with respect to the predicted attribute based on the (normalized) standard deviation of the attributes. It is only applicable to numerical predictions.

Table 2 Average monthly precipitation of stations according to 25 water years

Months	Babapirali	Kheyirabad	Khosroabad	Namileh	Pihan	Sarabi	Varayineh	Vasaj
Oct	9.0	9.9	8.0	10.7	7.8	11.0	13.5	11.2
Nov	76.6	64.0	62.4	48.1	48.2	71.8	79.0	67.0
Dec	57.4	55.9	54.7	37.3	43.9	61.3	76.7	54.1
Jan	49.8	36.9	40.6	29.5	33.1	42.2	58.3	43.3
Feb	66.5	47.1	55.7	37.2	38.0	66.2	80.1	57.1
Mar	46.0	43.9	43.3	33.6	35.7	48.1	65.8	45.9
Apr	96.3	80.6	72.4	74.9	75.0	92.5	115.2	82.9
May	47.2	41.4	41.5	39.3	35.6	42.7	60.5	40.6
Jun	5.6	6.8	4.2	6.3	4.7	6.3	6.1	5.6
Jul	1.0	0.8	1.0	1.4	1.1	0.9	0.5	0.5
Aug	0.9	1.2	0.3	0.6	0.4	0.6	0.5	0.4
Sep	1.3	1.2	0.9	2.4	1.0	3.1	1.2	1.3
Mean	**38.15**	**32.49**	**32.08**	**26.78**	**27.13**	**37.21**	**46.47**	**34.16**

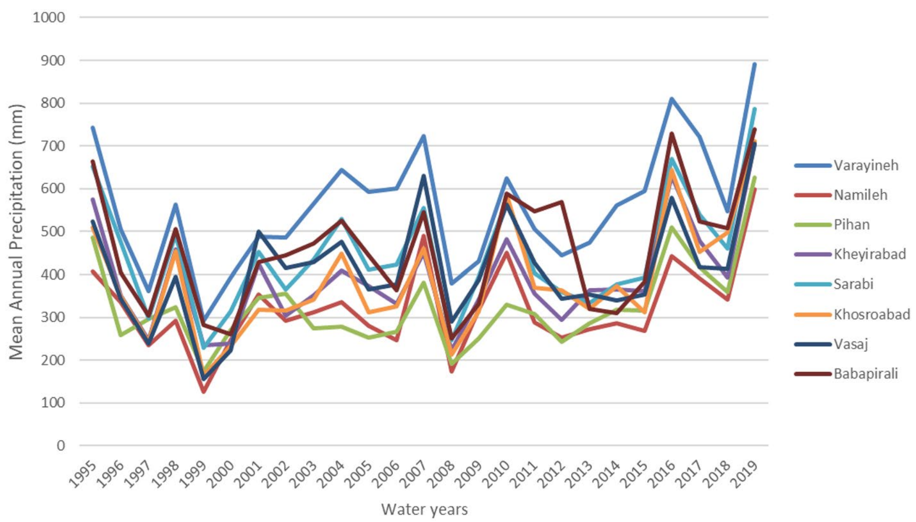

Fig. 3 Yearly time series of precipitation

Weight by PCA (Principal Component Analysis) Uses an orthogonal transformation to transform an initial (correlated) set of data into an equal or smaller set of so called principal components (a new set of uncorrelated attributes).

Weight by Correlation Computes the correlation between each attribute in the initial data set and the attribute predicted. The value returned is the absolute or squared value of the resulted correlation for each attribute.

Weight by SVM Calculates the relevance of the attributes by computing for each attribute of the initial data set, the weight with respect to the predicted attribute. The coefficients of a hyperplane calculated by an SVM are set as attribute weights.

The attributes with the highest weights are considered to have the highest relevance for all the above methods.

Each of the eight available locations was considered as predicted target and the values of the proposed weights were computed for all the other locations with respect to the predicted one. Once the results were obtained, the next step was to normalize the weights value, since these were in different ranges. According to Schowe (2011) there are three possibilities to combine the resulted weights: (1) by counting the number of times a feature was selected among top k features – the k features with the highest count are considered; (2) by defining a minimum number of iterations that a feature has to be selected; (3) by accumulating weights. In the present research, the third option was considered and the sum of the featured weights was computed for each location. It is also known as SAW (Prasetiyo and Baroroh 2016; Kaliszewski and Podkopaev 2016; Setyawan et al. 2017).

In the scenario creation phase, 3 alternatives were determined. In the first alternative (Called Scenario All), all stations surrounding a station participated in the modeling. In the second scenario (Scenario W2), stations providing the condition of the locations with a scored sum of weights $> = 2$ were included in the modeling. In the third scenario (Scenario

W1), stations providing the condition of the locations with a scored sum of weights $>=1$ were included in the modeling.

2.2.2 Deep Learning Based on a Multi-Layer Feed-Forward Artificial Neural Network

Deep Learning models was performed using RapidMiner. This algorithm is based on a multi-layer feed-forward artificial neural network, trained with stochastic gradient descent using back-propagation. The RapidMiner operator starts a 1-node local H_2O cluster and runs the algorithm on it. In the implemented scenarios, it attempts to minimize the network error by using the gradient descent method, more specific, by moving down the gradient of the error curve (Alsmadi et al. 2009).

Parameter Optimization The network can have multiple hidden layers of neurons, with the activation functions being *Tanh*, *Rectifier* or *Maxout* (Neamt et al. 2017). Before running the final experiment, we needed to know what the best combination of parameter inputs for the algorithm is. For this purpose, the following combinations of parameters were considered:

- Activation: *Tanh, Rectifier, Maxout, ExpRectifier*;
- Epochs: 10, 20, 30, 40, 50;
- Learning rate: 0.1 to 1 with a step of 0.1.

The best combination of parameters was chosen by looking at the results that produced the lowest Root Mean Squared Error (RMSE) per scenario (more than 2000 combinations per each feature weight scenario). Overall, the best results have been obtained for the ExpRectified activation function, 10 as number of epochs and learning rate different by case.

After the optimal values for the parameters were determined, the process described in Avram et al. (2020) has been applied to the input data in order to obtain the results for the predicted location (Fig. 4). The 300 months of input data were split into 70% data for learning the model and 30% data for testing it.

2.2.3 Convolutional Neural Networks (CNN) in Python

CNN were developed and inspired by biological processes in that the connection model between neurons resembles the organization of the animal visual cortex, and was first presented at the Neural Information Processing Workshop in 1987. CNN are deep artificial neural networks mainly used for classifying images, clustering similarities, and object identification. CNN is very successful in revealing the features of the data. A basic CNN network architecture usually has a ReLU layer as the activation function, followed by hidden layers such as fully connected layers, normalization layers, and pooling layers. Inputs and outputs are masked by the activation function.

Python programming language and its libraries were used for CNN modelling (https://github.com/Hapaydin/Hamadan/). The alternative Learning Rate (LR), Activation function, Decay (DE) and epochs of the models are determined using the grid search algorithm. In the initial stage LR, 0.1 to 1 with a step of 0.1 was used but produced worse results. Instead, $1e^{-1}$, $1e^{-2}$, $1e^{-3}$, $1e^{-4}$ and 1^{-5} were chosen. ReLU, Tanh and Sigmoid chosen as activation function. Decay from $1e^{-1}$ to $1e^{-5}$, and 5, 10, 50, 100, 500 and 1000 epochs have been tried.

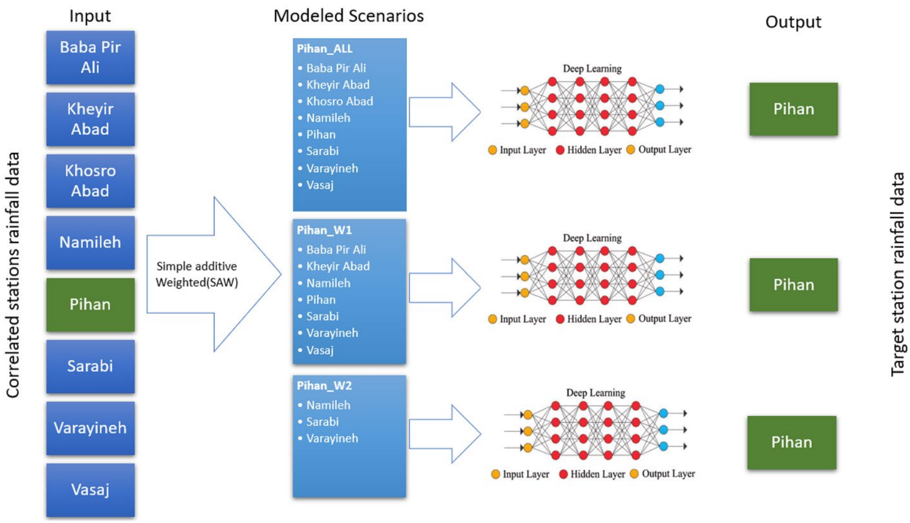

Fig. 4 Schematic example of the methods exemplified for the Pihan station

The assumption that data collected for the study is unbiased and that there is no error during the collection of measurements was taken into account. In addition, in the preparation of the experiments and parameter optimization phase, since the number of combinations might be infinite and the resources limited, we based the selection of the best option by looking at the RMSE and the overall results of about 2000 combinations per scenario.

2.3 Evaluation Metrics

More well-known metrics such as Pearson correlation coefficient (r), Nash–Sutcliffe coefficient (NS), Willmott index of agreement (WI), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) were computed in the evaluation of the models (Hyndman and Koehler 2006).

3 Results

3.1 Feature Selection

In this study, firstly, a station was selected as the target station and the weights of the other stations were calculated according to the feature weights algorithms. Then this process was applied for each station. For the purpose of comparison between the calculated weights, Table 3 presents the summary of the obtained aggregated feature weights results, ordered by descending value, for each predicted location.

As can be seen from Table 3, the weights vary between 0 and 3.6195. Three alternatives were considered for the formation of the scenarios. To predict the target station; i) All stations as input (w = > 0), ii) Stations with resulted sum of weights > = 2 and, iii) Stations with resulted sum of weights > = 1 were considered as model input and scenarios were selected as in Table 4.

Table 3 Results of aggregated weights (AW) for each location with respect to the predicted location

Target	Babapirali	AW	Kheyirabad	AW	Khosroabad	AW	Namileh	AW	Pihan	AW	Sarabi	AW	Varayineh	AW	Vasaj	AW
Surrounding stations	Varayineh	3.2379	Sarabi	3.0254	Varayineh	3.2324	Varayineh	3.1212	Varayineh	3.5014	Varayineh	3.6195	Sarabi	3.5888	Sarabi	3.02397
	Sarabi	3.0181	Varayineh	2.8169	Sarabi	3.0212	Sarabi	2.3234	Sarabi	2.0208	Vasaj	2.6357	Vasaj	2.4224	Varayineh	2.85573
	Vasaj	2.5671	Vasaj	2.0183	Babapirali	2.8847	Vasaj	2.1597	Namileh	2.0000	Kheyirabad	2.4191	Babapirali	2.3432	Kheyirabad	1.89875
	Khosroabad	2.1819	Babapirali	1.4065	Vasaj	2.4195	Kheyirabad	2.1187	Kheyirabad	1.9971	Babapirali	2.3586	Kheyirabad	1.8208	Babapirali	1.80852
	Kheyirabad	1.9919	Namileh	1.0723	Kheyirabad	1.9073	Pihan	2.0000	Vasaj	1.7407	Khosroabad	1.6463	Khosroabad	0.9819	Khosroabad	1.36458
	Namileh	0.2263	Khosroabad	0.7283	Pihan	0.2140	Babapirali	1.3444	Babapirali	1.3665	Namileh	0.3020	Pihan	0.8528	Namileh	0.62744
	Pihan	0.0582	Pihan	0.5194	Namileh	0.0000	Khosroabad	0.8532	Khosroabad	0.6388	Pihan	0.0582	Namileh	0.5830	Pihan	0.05972

Table 4 Scenarios modeled based on feature selection. W2 is used as suffix for scenarios where weight > = 2; W1 for scenarios where weight > 1; All stations as input (w > 0)

Predicted Station	Scenario	Source station1	Source2	Source3	Source4	Source5	Source6	Source7
Babapirali	All	Varayineh	Kheyirabad	Sarabi	Khosroabad	Namileh	Pihan	Vasaj
Kheyirabad		Sarabi	Khosroabad	Namileh	Pihan	Vasaj	Babapirali	Varayineh
Khosroabad		Namileh	Pihan	Vasaj	Babapirali	Varayineh	Kheyirabad	Sarabi
Namileh		Pihan	Vasaj	Babapirali	Varayineh	Kheyirabad	Sarabi	Khosroabad
Pihan		Vasaj	Babapirali	Varayineh	Kheyirabad	Sarabi	Khosroabad	Namileh
Sarabi		Khosroabad	Namileh	Pihan	Vasaj	Babapirali	Varayineh	Kheyirabad
Varayineh		Kheyirabad	Sarabi	Khosroabad	Namileh	Pihan	Vasaj	Babapirali
Vasaj		Babapirali	Varayineh	Kheyirabad	Sarabi	Khosroabad	Namileh	Pihan
Babapirali	W2	Varayineh	Sarabi	Vasaj	Khosroabad			
Kheyirabad		Sarabi	Varayineh	Vasaj				
Khosroabad		Varayineh	Sarabi	Babapirali	Vasaj			
Namileh		Varayineh	Sarabi	Vasaj	Kheyirabad	Pihan		
Pihan		Varayineh	Sarabi	Namileh				
Sarabi		Varayineh	Vasaj	Kheyirabad	Babapirali			
Varayineh		Sarabi	Vasaj	Babapirali				
Vasaj		Sarabi	Varayineh					
Babapirali	W1	Varayineh	Sarabi	Vasaj	Khosroabad	Kheyirabad		
Kheyirabad		Sarabi	Varayineh	Vasaj	Babapirali	Namileh		
Khosroabad		Varayineh	Sarabi	Babapirali	Vasaj	Kheyirabad		
Namileh		Varayineh	Sarabi	Vasaj	Kheyirabad	Pihan	Babapirali	
Pihan		Varayineh	Sarabi	Namileh	Kheyirabad	Vasaj	Babapirali	
Sarabi		Varayineh	Vasaj	Kheyirabad	Babapirali	Khosroabad		
Varayineh		Sarabi	Vasaj	Babapirali	Kheyirabad			
Vasaj		Sarabi	Varayineh	Kheyirabad	Babapirali	Khosroabad		

In this case, either all stations or stations with a value greater than 1 or 2 according to the degree of weights were taken into account in order to estimate the precipitation of the target station. Naturally, the accuracy rate will increase when all stations are considered as model inputs, but it is aimed to avoid the size and complexity of the model by modeling with the least number of stations. In this context, the performance of each scenario in each station and method was evaluated separately.

3.2 Results of DL-SGD Method

The results obtained according to the deep learning based on a multi-layer feed-forward artificial neural network (DL-SGD) method are given in Table 5. The DL-SGD method predicted monthly precipitation at all stations with great success. In general, the results of all 3 scenarios follow each other with very little difference. At this stage, the results of each target station were discussed one by one.

1. **Babapirali**: In this station, the best result was obtained as CC = 0.9626 and RMSE = 15.2109 mm in the test phase according to the W1 scenario (Vrayyineh, Sarabi, Vasaj, Khosroabad and Kheyirabad). As can be seen here, Namileh and Pihan stations did not have an effect on the forecast of Babapirali station precipitation. As can be seen from the map in Fig. 1, Namileh and Pihan stations are located at a far distance from Babapirali station, and in this case, it is logical that these two stations have little or no physical effect on Babapirali precipitation.
2. **Kheyirabad**: In this station, although the CC value of the All scenario is slightly higher, the RMSE value is 1 mm lower, so it was chosen as the best alternative (CC = 0.9779

Table 5 Results of DL-SGD method

Target	Scenario	LR	Epoch	CCTr	CCTt	NSTr	NSTt	WITr	WITt	RMSETr	RMSETt	MAETr	MAETt
Babapirali	All	0.1	10	0.9424	0.9578	0.8879	0.8875	0.9700	0.9660	14.6597	18.8577	9.1742	15.0496
	W1	0.3	10	0.9373	**0.9626**	0.8784	**0.9252**	0.9670	0.9800	15.2847	**15.2109**	9.1218	9.2098
	W2	0.9	10	0.9369	0.9560	0.8631	0.9012	0.9570	0.9700	16.2214	17.4827	10.6627	11.5329
Kheyirabad	All	0.1	10	0.9759	0.9840	0.9507	0.8822	0.9870	0.9590	7.5963	17.6271	4.9919	11.2510
	W1	0.3	10	0.9752	**0.9779**	0.9497	**0.8940**	0.9870	0.9640	7.7477	**16.6576**	5.0348	11.6126
	W2	0.9	10	0.9693	0.9740	0.9394	0.8895	0.9840	0.9630	8.5036	17.0460	5.4089	10.5784
Khosroabad	All	0.1	10	0.9530	0.9253	0.9026	0.8373	0.9760	0.9470	11.0607	20.3763	7.8108	13.1388
	W1	0.3	10	0.9526	0.9231	0.9038	0.8349	0.9760	0.9470	11.0148	20.5744	7.4357	11.9064
	W2	0.9	10	0.9529	**0.9285**	0.8943	**0.8598**	0.9680	0.9620	11.5497	**18.9193**	8.3236	12.0012
Namileh	All	0.1	10	0.9569	**0.9477**	0.9144	0.8712	0.9760	0.9620	8.7277	**15.2309**	5.7906	11.3007
	W1	0.3	10	0.9554	0.9330	0.9122	0.8537	0.9760	0.9570	8.8749	16.1602	6.1643	10.7025
	W2	0.9	10	0.9561	0.9418	0.9139	0.8691	0.9770	0.9640	8.7887	15.3053	5.4109	10.2268
Pihan	All	0.1	10	0.9485	**0.9707**	0.8823	0.8865	0.9720	0.9650	10.3476	**14.7623**	7.2485	10.4748
	W1	0.3	10	0.9493	0.9689	0.8948	**0.8954**	0.9740	0.9690	9.7799	14.1320	6.9500	10.1955
	W2	0.9	10	0.9480	0.9630	0.8827	0.8296	0.9720	0.9400	10.3354	18.1108	7.1219	11.7248
Sarabi	All	0.1	10	0.9780	**0.9840**	0.9492	0.9356	0.9860	0.9800	8.8357	13.5916	6.1481	9.4954
	W1	0.3	10	0.9746	0.9805	0.9497	0.9255	0.9870	0.9770	8.9042	14.6107	6.5843	10.5946
	W2	0.9	10	0.9767	0.9821	0.9492	**0.9527**	0.9860	0.9860	8.9520	**11.6246**	6.8286	9.0351
Varayineh	All	0.1	10	0.9576	**0.9670**	0.9156	0.8947	0.9770	0.9770	14.6934	20.6893	9.2862	14.7963
	W1	0.3	10	0.9613	0.9501	0.9230	0.8602	0.9800	0.9580	14.0472	24.3281	9.1097	15.9766
	W2	0.9	10	0.9594	0.9504	0.9198	**0.8998**	0.9780	0.9740	14.3424	**20.2104**	9.4402	12.5135
Vasaj	All	0.1	10	0.9642	**0.9845**	0.9281	**0.9543**	0.9800	0.9870	10.0645	**10.4169**	6.9171	8.6760
	W1	0.3	10	0.9640	0.9722	0.9289	0.9013	0.9810	0.9710	10.0100	15.5058	6.8324	12.8675
	W2	0.9	10	0.9571	0.9721	0.9159	0.8583	0.9770	0.9540	10.8902	18.8646	7.2076	16.6388

Bold-red values indicate the most successful scenario

CC Correlation Coefficient, *NS* Nash–Sutcliffe efficiency index, *WI* Willmont index, *RMSE* Root mean square error, *MAE* Mean absolute error, *Tr* Train, *Tt* Test and RMSE = 16.6576 mm) in the testing phase compared to the W1 scenario (Sarabi, Varayineh, Vasaj, Babapirali and Namileh). As can be seen here, Khosroabad and Pihan stations had no effect on the forecasting of precipitation at Kheyirabad station. As can be seen from the map in Fig. 1, Khosroabad station is located at a far distance from Kheyirabad station and also Pihan station is 100 m higher than Kheyirabad station. In this case, it is thought that these two stations have no effect on Kheyirabad precipitations.

3. **Khosroabad**: At this station, the best result (CC = 0.9285 and RMSE = 18.9193 mm) was determined in the test phase according to the W2 scenario (Vrayyineh, Sarabi, Babapirali and Vasaj). As can be seen here, Kheyirabad, Namileh and Pihan stations did not have an effect on the forecast of Khosroabad station precipitation. As can be seen from the map in Fig. 1, Khosroabad station is located at a far distance from Kheyirabad, Namileh and Pihan stations. In this case, it is thought that these three stations will not have any effect on Khosroabad precipitation.
4. **Namileh**: In this station, the best result was determined as CC = 0.9477 and RMSE = 15.2309 mm in the test phase according to the All scenario (all stations).
5. **Pihan**: In this station, the best result CC = 0.9707 and RMSE = 14.7623 mm was determined in the test phase according to the All scenario (all stations).
6. **Sarabi**: At this station, the best result (CC = 0.9840 and RMSE = 13.5916 mm) was determined during the test phase according to the All scenario (all stations). As can be seen from the map in Fig. 1, Sarabi station is located in the approximate center of the study area. From this point of view, it is expected that the All scenario and all stations are effective for precipitation estimation.
7. **Varayineh**: At this station, the best result (CC = 0.9670 and RMSE = 20.6893 mm) was determined during the test phase according to the All scenario (all stations).
8. **Vasaj**: At this station, the best result (CC = 0.9845 and RMSE = 10.4169 mm) was determined during the test phase according to the All scenario (all stations). As can be seen from the map in Fig. 1, Vasaj station is also located in the relatively central position of the study area. From this point of view, it is expected that the All scenario and all stations will be effective for precipitation prediction.

Considering the results of all 8 stations, the DL-SGD method successfully predicted monthly precipitation levels at Vasaj station with the highest accuracy rate and the least error according to the evaluation criteria.

In order to make a better visual evaluation and comparison, the time series and scatter plot graphics of all 8 stations according to the DL-SGD method are given in Fig. 5. As can be seen from Fig. 5, the precipitation levels of all stations in general have been estimated at an acceptable level according to all 3 scenarios. At the same time, smaller levels of precipitation are better predicted than medium and high levels. It is seen that the elevation of the stations and their positions relative to the other stations are affected by the estimation values. In general, it was concluded that the DL-SGD method was effective at estimating monthly precipitation.

3.3 Results of CNN Method

The results obtained according to the CNN method are given in Table 6. Table 6 shows that when precipitation is calculated at all stations and in all 3 scenarios according to the CNN method, the CC value is high, but the amount of error is high at other stations with the

Fig. 5 Time series and scatter plots of the DL-SGD method

Fig. 5 (continued)

Table 6 The best results of different scenarios for CNN method

Target	Scenario	LR	DE	Epoch	CCTr	CCTt	NSTr	NSTt	WITr	WITt	RMSETr	RMSETt	MAETr	MAETt
Babapirali	All	0.01	0.01	100	0.9517	0.9562	0.8979	0.9127	0.9708	0.9775	13.9767	16.4297	7.9185	8.9460
	W1	0.01	0.01	50	0.9399	**0.9615**	0.8593	**0.9204**	0.9568	0.9777	16.4023	**15.6860**	9.2957	8.7877
	W2	0.01	0.01	10	0.9364	0.9592	0.8750	0.9078	0.9668	0.9772	15.4631	16.8850	9.2490	10.1409
Kheyirabad	All	0.01	0.01	10	0.9475	**0.9665**	0.8343	**0.9246**	0.9611	0.9812	10.1398	**10.0270**	6.3663	6.3787
	W1	0.01	0.01	5	0.9447	0.9610	0.7188	0.8887	0.9375	0.9733	13.4406	12.2394	8.8509	7.4543
	W2	0.01	0.01	10	0.9376	0.9541	0.8715	0.8953	0.9664	0.9698	8.8831	11.8614	5.6685	7.7207
Khosroabad	All	0.01	0.01	10	0.8880	0.8728	0.7127	0.7285	0.9289	0.9306	13.5318	19.8512	8.6748	10.3868
	W1	0.01	0.01	50	0.9121	0.8685	0.8287	0.7322	0.9514	0.9174	10.6081	19.1898	6.5343	10.6825
	W2	0.01	0.01	10	0.9079	**0.8720**	0.7594	**0.7486**	0.9408	0.9330	12.7021	**18.5460**	8.1241	9.9017
Namileh	All	0.01	0.01	50	0.9115	**0.9421**	0.8237	**0.8853**	0.9490	0.9683	9.5740	**11.3406**	5.6449	6.2781
	W1	0.01	0.01	50	0.9100	0.9310	0.8253	0.8646	0.9500	0.9622	9.0406	11.0929	5.2952	6.7155
	W2	0.01	0.01	50	0.9104	0.9338	0.8052	0.8587	0.9411	0.9587	10.7754	11.7803	6.7198	7.0426
Pihan	All	0.01	0.01	50	0.9253	**0.9430**	0.8527	**0.8697**	0.9583	0.9625	8.3209	**11.1221**	5.1602	6.2187
	W1	0.01	0.01	50	0.9157	0.9411	0.7966	0.8221	0.9362	0.9448	9.8582	13.0846	5.8776	7.4861
	W2	0.01	0.01	10	0.8700	0.9355	0.6787	0.8665	0.9186	0.9657	12.6207	11.2636	7.8050	7.1483
Sarabi	All	0.01	0.01	50	0.9590	0.9629	0.9133	0.9258	0.9762	0.9802	8.3422	10.3960	5.4713	6.6094
	W1	0.01	0.01	50	0.9572	0.9709	0.8912	0.9198	0.9688	0.9770	9.3738	10.8172	6.1050	7.0370
	W2	0.01	0.01	5	0.9488	**0.9753**	0.8996	**0.9497**	0.9733	0.9867	8.9962	**8.5302**	5.9678	5.3839
Varayineh	All	0.01	0.01	500	0.9812	**0.9679**	0.9536	**0.9362**	0.9871	0.9835	10.8802	**16.0988**	6.2167	10.3021
	W1	0.01	0.01	100	0.9635	0.9601	0.9091	0.9059	0.9732	0.9739	17.9626	25.2831	10.4618	14.6068
	W2	0.01	0.01	5	0.9597	0.9569	0.9132	0.9039	0.9752	0.9740	14.8816	19.7583	9.0036	10.5227
Vasaj	All	0.01	0.01	50	0.9281	0.9730	0.8577	0.9127	0.9630	0.9796	10.1795	10.2347	6.5883	6.5624
	W1	0.01	0.01	50	0.9257	**0.9700**	0.8479	**0.9403**	0.9567	0.9845	10.5400	**8.4304**	6.7986	5.4850
	W2	0.01	0.001	5	0.9093	0.9534	0.7757	0.8718	0.9289	0.9611	12.9347	12.4714	8.3032	7.2833

Bold-red values indicate the most successful scenario

CC Correlation Coefficient, *NS* Nash–Sutcliffe efficiency index, *WI* Willmont index, *RMSE* Root mean square error, *MAE* Mean absolute error, *Tr* Train, *Tt* Test

exception of the Babapirali and Varyaneh stations. All 8 stations were evaluated separately, taking into account the evaluation criteria given in Table 6.

1. **Babapirali**: At this station, the best results (CC = 0.9615 and RMSE = 15.6860 mm) were obtained in the W1 scenario (Varayineh, Sarabi, Vasaj, Khosroabad and Kheyirabad) during the test phase. As can be seen, the Namileh and Pihan stations had no effect on the forecasting of precipitation at Babapirali station. As explained in the same DL-SGD method, the map in Fig. 1 shows that the Namileh and Pihan stations are located at a remote distance from the Babapirali station, and it is logical that these two stations have no physical effect on the Babapirali precipitation levels. Thus, we can see that DL-SGD and CNN methods provide similar results at Babapirali station.
2. **Kheyirabad**: At this station, the best result (CC = 0.9665 and RMSE = 10.0270 mm) was obtained from the All scenario (all stations). Considering the CC and RMSE values obtained here, it cannot be said clearly which of the CNN and DL-SGD methods is more effective.
3. **Khosroabad**: According to the W2 scenario (Varayineh, Sarabi, Babapirali and Vasaj) at this station, the best result in the test phase was determined as CC = 0.8720 and RMSE = 18.5460 mm. The Kheyirabad, Namileh and Pihan stations did not have an effect on the prediction of precipitation at the Khosroabad station. While the RMSE values of the DL-SGD and CNN methods are close to each other, the CC value of the DL-SGD method is higher.
4. **Namileh**: According to the All scenario (all stations) at this station, the best result was determined as CC = 0.9421 and RMSE = 11.3406 mm during the test phase. While the CC values of the DL-SGD and CNN methods are close to each other, the RMSE value of the CNN method is lower.
5. **Pihan**: At this station, the best results (CC = 0.9430 and RMSE = 11.1221 mm) were obtained in the All scenario (all stations). Considering the CC and RMSE values obtained here, it cannot be said clearly which of the CNN and DL-SGD methods is more successful.

Fig. 6 Time series and scatter plots based on feauture selection methods for all stations on CNN method

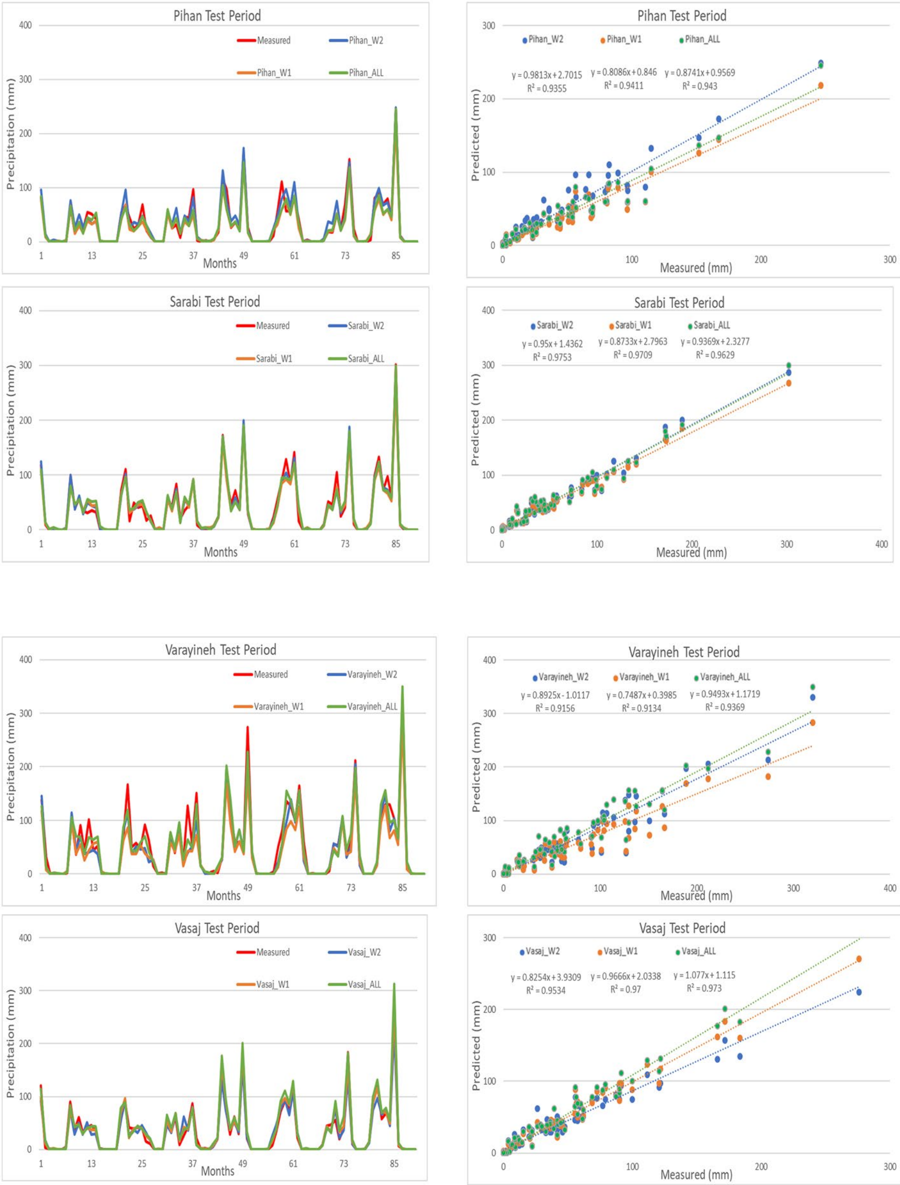

Fig. 6 (continued)

6. **Sarabi**: In this station, the best result was obtained as CC = 0.9753 and RMSE = 8.5302 mm in the test phase according to the W2 scenario (Varayineh, Vasaj, Kheyirabad and Babapirali). While the CC values of the DL-SGD and CNN methods are close to each other, the RMSE value of the CNN method is lower.
7. **Varayineh**: In this station, the best result in the All scenario (all stations) alternative was determined as CC = 0.9679 and RMSE = 16.0988 mm. While the CC values of the DL-SGD and CNN methods are close to each other, the RMSE value of the CNN method is lower.
8. **Vasaj**: At this station, the best result (CC = 0.9700 and RMSE = 8.4304 mm) was obtained in the W1 scenario (Sarabi, Varayineh, Kheyirabad, Babapirali and Khosroabad). While the CC values of the DL-SGD and CNN methods are close to each other, the RMSE value of the CNN method is lower.

When the results of these 8 stations are examined, it is seen that the estimation results of the CNN method are quite accurate in other stations except for Khosroabad station. Considering the RMSE values, CNN gave more accurate results than the DL-SGD method. Considering the CC and NS values, a definite superiority cannot be determined. In order to make a more effective evaluation and comparison visually, the time series and scatter plot graphics of all 8 stations according to the CNN method are given in Fig. 6.

As can be seen in Fig. 6, the precipitations of all stations are generally estimated at an acceptable level based on the CC and RMSE values for each of the 3 scenarios. However, when we consider the margins of error, only the Babapirali and Varayineh stations have made really accurate estimations. When we look at the scatter plots here, it is seen that the points are sparse and scattered to the far points of the diagonal line, and, consequently, it is inevitable that the error value increases.

4 Discussion and Conclusions

In this study, the monthly precipitation levels of 8 precipitation observation stations were estimated according to the precipitation levels of neighbouring stations and based on the principle that precipitation events between neighbouring stations may be similar. This was done by applying the Stochastic Gradient Descent using back-propagation (DL-SGD) and Convolutional Neural Networks (CNN). The novelty proposed in the paper is designing scenarios by reducing the number of sources based on the computed AW (aggregated weights).

Four feature selection algorithms were used in evaluating the available data sources (Weight by deviation, Weight by PCA, Weight by correlation and Weight by SVM) and SAW technique was used to combine the results. Three distinct scenarios, involving a different number of sources, based on the value of the aggregated weights, were proposed, and applied using the DL-SGD and CNN methods. The following conclusions were reached after analysing the outcome:

- In the selection of the scenarios, it was observed that the weight values were compatible with the station locations and the distance between the stations.

- The CNN model predicted precipitation at other stations with low error values, with the exception of one station. It can be stated that the CNN method provided better results overall, since all the scenarios produced results closer to the observed value than when applying the DL-SGD method.
- Since this study was conducted in a semi-arid and mountainous region, the results may not yield similar results in all climates and regions. This may be considered as a disadvantage of the study.
- All three proposed scenarios produced similar results in estimating the monthly precipitations, hence the SAW technique can be used to reduce the number of correlated sources.
- Another enhancement that could be done is expanding the list of machine learning algorithms applied and extending the research to different areas while addressing other practical prediction problems.

One of the main limitations of the study is that it refers only to a certain selected areal (province of Hamadan, Iran) and because of its specificity, the extracted conclusions cannot be generalized. However, the proposed approach of considering only a subset of the correlated sources, by applying the SAW technique is a procedure that could be further enhanced by exploring the limit of this value that would still allow for accurate predictions.

Acknowledgements This work was supported by the project "Collaborative Framework for Smart Agriculture" – COSA that received funding from Romania's National Recovery and Resilience Plan PNRR-III-C9-2022-I8, under grant agreement 760070.

Authors' Contributions Mohammad Taghi Sattari performed conceptualization, methodology, validation, writing—original draft preparation and supervision phases. Anca Avram performed methodology, software, validation, formal analysis, writing—original draft preparation and visualization stages. Halit Apaydin performed software, validation, data curation and visualization stages. O Oliviu Matei performed conceptualization, writing—review and editing, supervision phases. All authors read and approved the final manuscript.

Data Availability Data are available on request due to privacy or other restrictions.

Code Availability Rapidminer (educational license) was used. Python code are available on https://github.com/Hapaydin/Hamadan.

Declarations

Ethics Approval No need/Not applicable.

Consent to Participate No need/Not applicable.

Consent for Publication No need/Not applicable.

Conflict of Interest/Competing Interests The authors declare no conflict of interest.

References

Afshari Nia M, Panahi F, Ehteram M (2023) Convolutional neural network- ANN- E (Tanh): A new deep learning model for predicting rainfall. Water Resour Manag 37:1785–1810. https://doi.org/10.1007/s11269-023-03454-8

Ahi Y, Coşkun Dilcan Ç, Köksal DD (2023) Reservoir evaporation forecasting based on climate change scenarios using ANN. Water Resour Manag 37:2607–2624. https://doi.org/10.1007/s11269-022-03365-0

Alsmadi MKS, Omar K, Noah SA (2009) Back propagation algorithm: The best algorithm among the multi-layer perceptron algorithm. Int J Comput Sci Netw Secur 378–383

Apaydin H, Sattari MT (2020) Deep-learning GIS hybrid approach in precipitation modeling based on spatio-temporal variables in the coastal zone of Turkey. Climate Res 9:81. https://doi.org/10.3354/cr01612

Avram A, Matei O, Pintea C, Anton C (2020) Innovative platform for designing hybrid collaborative & context-aware data mining scenarios. Mathematics 5:8. https://doi.org/10.3390/math8050684

Baaghideh M, Fallah Ghalhari G, Hajimohammadi H, Rezaei H (2017) Investigating the role of irregularities in the formation of regions and sub-regions of Hamadan Province. Quart Geogr Data 26(103):109–122. https://doi.org/10.22131/sepehr.2017.28897

Chaplot B (2021) Prediction of rainfall time series using soft computing techniques. Environ Monit Assess 193:721. https://doi.org/10.1007/s10661-021-09388-1

Chen H, Chandrasekar V, Cifelli R, Xie P (2020) A machine learning system for precipitation estimation using satellite and ground radar network observations. IEEE Trans Geosci Remote Sens 2:58. https://doi.org/10.1109/TGRS.2019.2942280

Diriba TA, Debusho LA (2021) Statistical modelling of extreme rainfall indices using multivariate extreme value distributions. Environ Model Assess 8:26. https://doi.org/10.1007/s10666-021-09766-6

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22:679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Inan HI (2021) Spatial data model for rural planning and land management in Turkey. J Agric Sci 27(3):254–266. https://doi.org/10.15832/ankutbd.983096

Kaliszewski I, Podkopaev D (2016) Simple additive weighting—A metamodel for multiple criteria decision analysis methods. Expert Syst Appl 7:54. https://doi.org/10.1016/j.eswa.2016.01.042

Kassem Y, Gökçekuş H (2021) Do quadratic and Poisson regression models help to predict monthly rainfall? Desalin Water Treat 215. https://doi.org/10.5004/dwt.2021.26397

Lin K, Zhou J, Liang R, Hu X, Lan T, Liu M, Gao X, Yan D (2021) Identifying rainfall threshold of flash flood using entropy decision approach and hydrological model method. Nat Hazards 9:108. https://doi.org/10.1007/s11069-021-04739-0

Neamt L, Matei O, Chiver O (2017) Finite element method combined with neural networks for power system grounding investigation. Int J Adv Comput Sci Appl 8(2):187–192

Ng CWW, Liu ZQ, Kwan JSH, Yang B (2021) Spatiotemporal modelling of rainfall-induced landslides using machine learning. Landslides 7:18. https://doi.org/10.1007/s10346-021-01662-0

Nourani V, Sattari MT, Molajou A (2017) Threshold-based hybrid data mining method for long-term maximum precipitation forecasting. Water Resour Manag 7:31. https://doi.org/10.1007/s11269-017-1649-y

Prasetiyo B, Baroroh N (2016) Fuzzy simple additive weighting method in the decision making of human resource recruitment. Lontar Komputer: Jurnal Ilmiah Teknologi Informasi 12. https://doi.org/10.24843/LKJITI.2016.v07.i03.p05

Roslan N, Md Reba MN, Sharoni SMH, Hossain MS (2021) The 3D neural network for improving radar-rainfall estimation in monsoon climate. Atmosphere 5:12. https://doi.org/10.3390/atmos12050634

Sattari MT, Apaydin H, Ozturk F, Baykal N (2012) Application of a data mining approach to derive operating rules for the Eleviyan irrigation reservoir. Lake Reservoir Manag 28(2):142–152. https://doi.org/10.1080/07438141.2012.678927

Schowe B (2011) Feature selection for high-dimensional data with RapidMiner. Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011), Aachen

Setyawan A, Akhlis I, Arini FY (2017) Comparative analysis of simple additive weighting method and weighted product method to new employee recruitment decision support system (DSS) at PT. Warta Media Nusantara. Sci J Inform 5:4. https://doi.org/10.15294/sji.v4i1.8458

Sinamet (2018) Analysis of unprecedented rainfall in April 2018 in Hamedan province. Ministry of Roads and Urban Development, National Meteorological Organization, General Directorate of Meteorology. Access date: 06 Jun 2023. http://www.sinamet.ir/data/prsinamet/pr/baresh%20bi%20sabegheh.pdf

Tabatabaei SM, Hamraz BS, Nazeri Tahroudi M (2021) Comparison of the performances of GEP, ANFIS, and SVM artificial intelligence models in rainfall simulation. Időjárás 125. https://doi.org/10.28974/idojaras.2021.2.2

Vathsala H, Koolagudi SG (2021) Neuro-fuzzy model for quantified rainfall prediction using data mining and soft computing approaches. IETE J Res 4. https://doi.org/10.1080/03772063.2021.1912648

Xie X, Xie B, Cheng J, Chu Q, Dooling T (2021) A simple Monte Carlo method for estimating the chance of a cyclone impact. Nat Hazards 107(3):2573–2582. https://doi.org/10.1007/s11069-021-04505-2

Yan J, Xu T, Yu Y, Xu H (2021) Rainfall forecast model based on the TabNet model. Water 4:13. https://doi.org/10.3390/w13091272

Zhu X, Xu Z, Liu Z, Liu M, Yin Z, Yin L, Zheng W (2022) Impact of dam construction on precipitation: a regional perspective. Mar Freshw Res 74:877–890. https://doi.org/10.1071/MF22135