



Application of machine learning ensemble models for rainfall prediction

Hasan Ahmadi¹ · Babak Aminnejad¹ · Hojat Sabatsany¹

Received: 6 May 2022 / Accepted: 6 October 2022 / Published online: 10 November 2022

© The Author(s) under exclusive licence to Institute of Geophysics, Polish Academy of Sciences & Polish Academy of Sciences 2022

Abstract

Practical information can be drawn from rainfall for making long-term water resources management plans, taking flood mitigation measures, and even establishing proper irrigation systems. Given that a large amount of data with high resolution is required for physical modeling, this study proposes a new standalone sequential minimal optimization (SMO) regression model and develops its ensembles using Dagging (DA), random committee (RC), and additive regression (AR) models (i.e., DA-SMO, RC-SMO, and AR-SMO) for rainfall prediction. First, 30-year monthly data derived from the year 1988 to 2018 including evaporation, maximum and minimum temperatures, maximum and minimum relative humidity rates, sunshine hours, and wind speed as input and rainfall as the output were acquired from a synoptic station in Kermanshah, Iran. Next, based on the Pearson correlation coefficient (r -value) between input and output variables, different input scenarios were formed. Then, the dataset was separated into three subsets: development (1988–2008), calibration (2009–2010), and validation (2011–2018). Finally, the performance of the developed algorithms was validated using different visual (scatterplot and boxplot) and quantitative (percentage of BIAS, root mean square error, Nash–Sutcliffe efficiency, and mean absolute error) metrics. The results revealed that minimum relative humidity had the greatest effect on rainfall prediction. The most effective input scenario featured all the input variables except for wind speed. Our findings indicated that the DA-SMO ensemble algorithm outperformed other algorithms.

Keywords Dagging · Ensemble models · Machine learning · Performance evaluation · Rainfall prediction · SMO

Introduction

Accurate rainfall prediction leads to not only better prevention of disasters in the event of flood, but also proper management of water, agricultural and aquatic ecosystems, and drought. While low rainfall results in drought, heavy rainfall increases the probability of landslides, floods, and other natural disasters. It is expected that the recent rise in the concentration of atmospheric carbon dioxide and the consequent increase in the global temperature will have a substantial impact on the rainfall pattern on local, regional, and global scales (Wang et al. 2013; Adefisan 2018; Zhu et al.

2022a,b). Since rainfall patterns are highly chaotic, the level of uncertainty in its prediction increases. Therefore, early warning systems should enjoy greater accuracy so that risks to life and property can be reduced, if not averted. Rainfall occurrences mainly depend on several meteorological variables including evaporation, maximum relative humidity, minimum relative humidity, maximum temperature, minimum temperature, sunshine hours, and wind speed (Zhao et al. 2021a,b).

Near real-time rainfall prediction is generally done through physical/numerical models, which are developed based on dynamical equations. These models facilitate acquiring extensive information about time and space (Toth et al. 2000; Liu et al. 2022a, b). Nevertheless, requiring a significant amount of data with high resolution for physical/numerical modeling and their time-consuming implementation are the major drawbacks of the mentioned models. To overcome these obstacles, attempts have been made to develop statistical formulae in conjunction with statistical models for a more efficient rainfall prediction based

Edited by Dr. Ankit Garg (ASSOCIATE EDITOR) / Dr. Michael Nones (CO-EDITOR-IN-CHIEF).

✉ Hasan Ahmadi
hgahmadi@riau.ac.ir

¹ Department of Civil Engineering, Islamic Azad University Roudehen Branch, Tehran, Iran

on historical data. The autoregressive time-series prediction methods like autoregressive integrated moving average (ARIMA) and multiple linear regression (MLR), which have been applied to many different fields of study (Delleur and Kavvas 1978; Yevjevich 1987; Zhang et al. 2019a, b, 2022; Osouli et al. 2022). However, these models are not robust or flexible enough to predict complicated and chaotic phenomena like rainfall.

So far, different types of machine learning (ML) models have drawn the attention of many researchers around the world. These models enjoy the following advantages: nonlinear structure, ability to model nonlinear phenomena, requiring small input datasets, user-friendliness, and ability to find a relationship between input and output variables to predict the target variable (Elbaz et al. 2019, 2020; Khosravi et al. 2021, 2022a,b; Liu et al. 2020). Artificial neural networks (ANN) are derived from the pioneering ML models whose applicability to rainfall forecasting is not much prevalent (e.g., Luk et al. 2000; Aksoy and Dahamsheh 2009; Samantaray et al. 2020, among others), although they have been found relevant to fields of geoscience (Ghumman et al. 2011; Oyebode and Stretch, 2019; Kadam et al. 2019). Generally, ANN model has low generalization power and low convergence speed (Aksoy and Dahamsheh 2009; Wang et al. 2021; Quan et al. 2022). To overcome these shortcomings, ANN have been integrated with fuzzy logic to develop neuro-fuzzy interface systems (ANFISs). Although ANFIS benefits from both ANN and fuzzy logic and has higher performance and generalization power, determining the weights of membership functions is still a challenging task in these cases. Other alternatives such as support vector machine (SVM) (Nhu et al. 2019), gene expression programming (GEP) (Sheikh Khozani et al. 2017), and extreme learning machine (ELM) (Atiquzzaman and Kandasamy 2018; Zhao et al. 2021a, b, c; Wang et al. 2022) have also been applied in recent years.

Recently, to enhance the prediction accuracy of the models through data pre- or post-processing approaches, ML ensemble algorithms have been developed by integrating two or more types of models (e.g., Zhang et al. 2019a, b; Xie et al. 2021a, b) to increase the modeling performance, thus benefiting from multiple advantages of two model types at the same time. An ensemble predictive model was developed by Sivapragasam et al. (2001) for one-day-ahead forecasting using SVM and singular spectrum analysis models. They employed singular spectrum analysis in order to disintegrate rainfall data for training SVM through a supervised approach. The authors argued that the ensemble algorithm of the singular spectrum analysis-SVM model could outperform nonlinear prediction methods. Similarly, Chau and Wu (2010) revealed the higher predictive power of the singular spectrum analysis-SVM than the traditional ANN model. Wu et al. (2010) reported that particle swarm optimization

(PSO) as a metaheuristic algorithm could enhance the performance of the standalone SVM model. While most of the neuron-based algorithms suffer from inadequate accuracy in determining the weights of membership functions, metaheuristic algorithms are able to determine accurate weights automatically. Kisi and Shiri (2011) examined the efficiency of the ensemble-based wavelet-GEP algorithm in predicting daily rainfall and compared it with a hybrid wavelet-neuro-fuzzy method in terms of predictive power. The performance of the wavelet-GEP was reported better than that of the neuro-fuzzy method. Yaseen et al. (2019) used a hybrid ANFIS model with three metaheuristic algorithms, namely PSO, genetic algorithm (GA), and differential evolution (DE), for monthly rainfall prediction. Finally, the performance of the hybrid ANFIS model was compared with that of the standalone ANFIS model. The results revealed that all the hybrid models outperformed the standalone model.

All in all, these studies have illustrated the necessity of using hybrid models for precise forecasting of long-term rainfall. However, despite their complicated structures, these models are of higher accuracy than standalone models in most cases. Accordingly, there is an ongoing effort to increase their modeling accuracy and reduce their complexity. Ridwan et al. (2021) compared Bayesian linear regression (BLR), boosted decision tree regression (BDTR), decision forest regression (DFR), and neural network regression (NNR) in terms of rainfall prediction accuracy in Malaysia. They found that monthly dataset facilitated higher prediction accuracy than daily and weekly datasets and revealed that BDTR model had favorable performance in most cases.

In recent years, advanced ML models in combination with sequential minimal optimization (SMO) have successfully been applied for regression and classification purposes in different fields of science. For example, Hashmi et al. (2015) compared model tree with SMO in terms of predicting the minimum surface roughness value and indicated the high predictive power of both algorithms. Gao et al. (2019) successfully employed SMO algorithm for designing energy-efficient residential buildings.

In this paper, standalone SMO algorithms as well as three new hybrid models developed through effective integration of additive regression (AR), Dagging (DA), and random committee (RC) algorithms (i.e., AR-SMO, DA-SMO, and RC-SMO) are presented for rainfall prediction. As stated in the literature review, although there are many applicable models for rainfall prediction, most of them are classified as traditional with high uncertainty in performance. In the present study, authors tried to boost the accuracy of new ML models through hybridization and apply them to rainfall prediction. In addition, they intend to explore what types of input variable among hydrometeorological data are effective in rainfall prediction and what input scenario has the

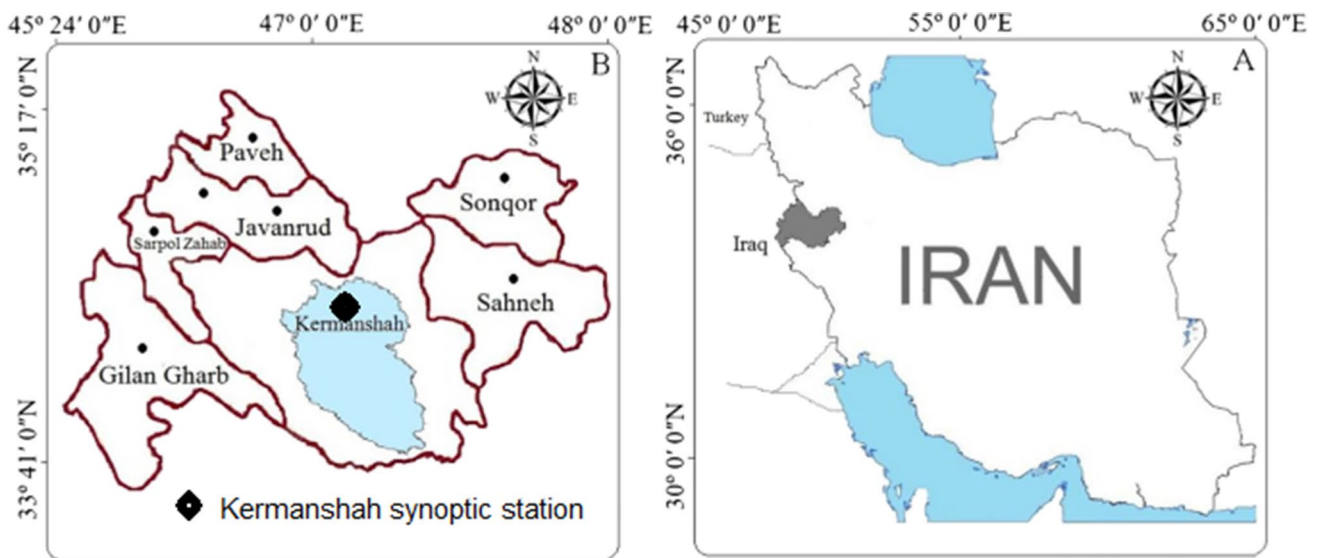


Fig. 1 The location of the Kermanshah synoptic station

highest flexibility. To the best of our knowledge, this is the first time that these types of hybrid algorithms have been applied to hydrology, especially for rainfall prediction. Other new aspects of the current study are: (a) investigating the effectiveness of each input variable (about 7 input variables) in the result, and (b) finding an appropriate input scenario.

Case study

Kermanshah synoptic station with an elevation of 1318 m at latitude of $34^{\circ} 21' N$ and a longitude of $47^{\circ} 9' E$ in Kermanshah province, west of Iran (Fig. 1), was considered as our case study. The mean annual rainfall of 450 mm was recorded by this station located in a mountainous area, and the corresponding long-term hydrometeorological information was achieved. The wind that blows from the west and carries humidity from the Mediterranean and Atlantic Oceans is the main source of rainfalls and snow-falls during spring and winter, respectively, while the wind blowing in the summer is hot.

Methodology

The methodological framework of the current study is presented in Fig. 2.

Dataset collection

Thirty-year monthly data from January 1988 to December 2018 including evaporation (Eva), maximum relative humidity (MARH), minimum relative humidity (MIRH),

maximum temperature (MAT), minimum temperature (MIT), sunshine hours (SSH), wind speed (WS), and rainfall were recorded from Kermanshah synoptic station by Kermanshah Regional Water Authority. Rainfall was determined as the target variable, while the remaining variables were incorporated into the model for estimating the rainfall. The whole data were divided into three groups: 70% of the data from January 1988 to December 2008 used for model building; 15% from January 2009 to December 2010 for model calibration; and the remaining 15% from January 2011 to December 2018 for model validation. Statistical analysis of the development, calibration, and validation datasets is given in Table 1.

Input scenarios

R-value between the input variables and rainfall (as output) was calculated to construct different input scenarios for variables and investigate their effectiveness in rainfall modeling. The r -value pie-chart between input and output variables is shown in Fig. 3. The r -value shows that all input variables, except WS, contribute to high rainfall prediction power.

Seven different input scenarios were built based on the r -value. First, the variable with the highest r -value (i.e., MIRH) was considered as a single input scenario. Then, the potential variable with the next highest correlation value was added to the first input for building the input combination number 2. Each variable was added in a stepwise approach to the previous scenario and this approach continued until the variable with the lowest r -value (i.e., WS) was added.

Fig. 2 A basic flow chart of the current study

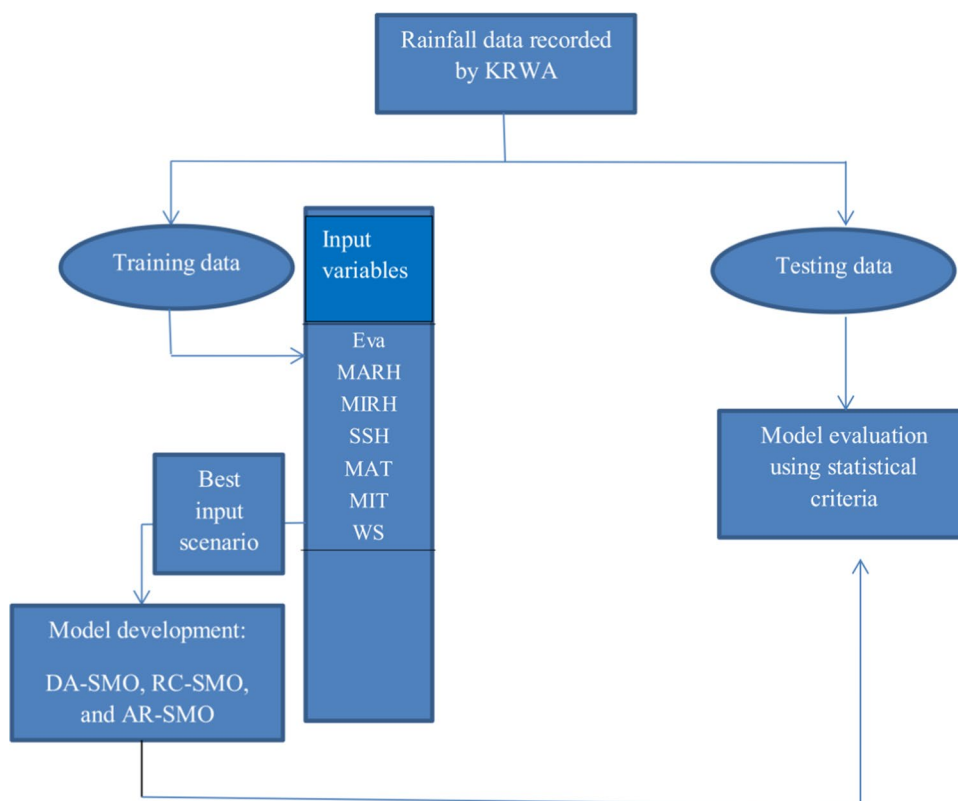


Table 1 Descriptive statistics for training, calibration, and validation

Variables	Development dataset				Calibration dataset				Validation dataset			
	Min	Max	Mean	St. D	Min	Max	Mean	St. D	Min	Max	Mean	St. D
Eva	13.04	0.10	5.28	3.75	8.84	0.56	3.43	2.51	12.69	0.47	4.85	3.56
MARH	95.48	20.65	61.26	22.89	87.00	27.45	61.19	22.20	91.84	24.33	61.85	21.85
MIRH	71.03	7.03	28.96	17.24	53.74	7.61	27.02	14.95	70.87	6.55	27.44	15.21
SSH	12.79	2.84	8.34	2.73	11.84	4.29	8.09	2.26	12.60	2.80	8.20	2.43
MAT	38.26	0.22	21.56	10.98	37.57	6.05	22.53	9.99	38.55	4.02	22.59	10.51
MIT	24.82	-10.84	8.12	8.25	21.25	-3.65	9.06	7.53	23.23	-3.79	8.90	7.86
WS	5.37	0.27	2.81	0.98	3.95	0.69	2.15	0.77	3.79	0.58	2.04	0.68
Rainfall	171.50	0.00	34.65	40.70	85.00	0.00	26.69	25.15	143.00	0.00	26.86	27.07

St. D standard deviation

To achieve the most effective input scenario, all the input scenarios were fed as input into the rainfall model. Then, the results were compared and the scenario with the highest *r*-value was determined as the optimum input scenario (Table 2).

Optimum weights

Obtaining the optimum weight of each parameter in the model is a crucial phase in the modeling process, which can significantly enhance the predictive power of a model. This step was taken in the Waikato Environment for Knowledge

Analysis (WEKA 3.9) software using the trial-and-error method. First, default values were considered, and then, the models were developed. Next, based on the achieved results, higher and lower values were applied until the optimal value was determined for the model. The lowest RMSE value was considered as a criterion in this step to determine the optimum values for the models.

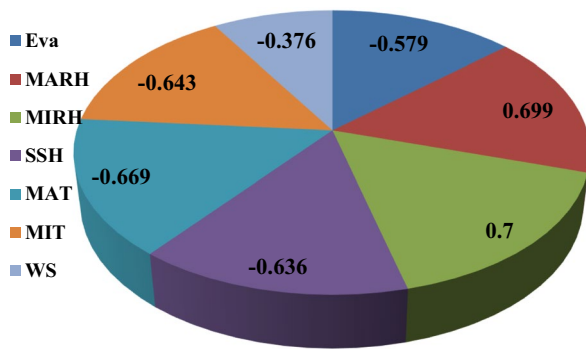


Fig. 3 Importance of input variables in terms of Pearson correlation coefficient

Table 2 Input variable scenarios

No	Different input scenarios
1	MIRH
2	MIRH, MARH
3	MIRH, MARH, MAT
4	MIRH, MARH, MAT, MIT
5	MIRH, MARH, MAT, MIT, SSH
6	MIRH, MARH, MAT, MIT, SSH, Eva
7	MIRH, MARH, MAT, MIT, SSH, Eva, WS

Description of the models

Sequential minimal optimization (SMO) regression algorithm

SMO algorithm is one of the well-known function algorithms that was first invented by Platt (1988) at Microsoft Research Center. This iterative algorithm is developed for solving optimization problems, e.g., in quadratic programming, which is a popular weakness of the SVM algorithm. Generally, the need for quadratic programming arises during the training of the SVM algorithm, and SMO algorithm is an alternative to QP methods, which are sophisticated and need expensive third-party QP solvers. SMO, on the other hand, can divide such problems into smaller parts, called sub-problems. The main advantage of these sub-problems is that they can be analytically solved. Of note, the SMO algorithm implements SVM for regression problems. Hence, the parameters can be learned using various algorithms. The algorithm is selected by setting the reg-optimizer as the most popular algorithm. Consider a binary classification with a dataset $(x_1, y_1), \dots, (x_n, y_n)$, where x_i is an input vector and $y_i \in \{-1, +1\}$ is a binary label corresponding to it. A soft-margin SVM is trained by solving Q_p , which is explained in a dual form as follows:

$$\text{Max} \sum_{i=1}^n \alpha_i - 0.5 \sum_{i=1}^n \sum_{j=1}^n y_i y_j K(x_i, x_j) \alpha_i \alpha_j, \quad (1)$$

subject to:

$$0 \leq \alpha_i \leq C, \text{ for } i = 1, 2, \dots, n, \quad (2)$$

$$\sum_{i=1}^n y_i \alpha_i = 0, \quad (3)$$

where C is an SVM hyperparameter and $K(x_i, x_j)$ is the kernel function, both supplied by the user; and the variables α_i are Lagrange multipliers.

Random committee (RC)

An ensemble-based model developed by combining more than two artificial intelligence techniques is called Committee Machines (CMs). CMs enjoy flexibility in modeling and know-how in resolving the deficiencies currently witnessed in the respective standalone model (Ghiasi-Freez et al. 2012; Chen et al. 2022; Yin et al. 2022a, b). The RC model is a kind of CM learning technique applied for addressing classification and regression issues and is considered to be an effective ensemble model (Niranjan et al. 2017). The RC model is able to develop hybridized randomizable base regressors or classifiers, each functioning based on the same data. However, a specific random seed is utilized and the final model response is obtained by averaging the estimations made by each standalone model (Witten and Frank 2005).

Disjoint aggregating (Dagging)

Developed by Ting and Witten (1997), the disjoint aggregating (Dagging) approach works on a proportionate stratified sampling scheme, dividing a dataset into a series of stratified folds in such a way that one individual learner manages a specific fold (Chen et al. 2020). Dagging is considered a robust tool for improving the accuracy of the weak single models in which a series of weak regressors are coupled using the voting rule for calculating the output of the model (Pham et al. 2020). The Dagging model can be established as follows: (i) The training dataset is split into a suite of "L" subsets; (ii) everyone should have "K" samples, and each sample belongs to only one subset; (iii) a regression model is developed for each subset, and in total, "L" regression models are obtained; and (iv) based on the comparison of the results from each regression model, the model with enough votes is selected for the Dagging algorithm (Chen and Li 2020).

Additive regression (AR)

AR model was introduced by Stone (1985). From a mathematical point of view, "Y" and explanatory " x_i " variables should be linked through a suite of nonparametric regression functions " f_i ", which are in turn defined to be a regression function of the independent variables " x_i ". Consequently, each independent variable " x_i " contributes to the final model in the form of ($f_i(x_i)$) compared to the linear form ($\lambda_i x_i$) used in the MLR model (Xu et al. 2017). The AR is a general (potentially nonlinear) regression model that includes linear regression as a special case. Suppose that variable $Y_i (i = 1, 2, \dots, n)$ is a function of unrestricted functions $f_j (j = 1, 2, \dots, p)$, which are determined by the input variables $X_{i1}, X_{i2}, \dots, X_{ip}$, respectively. The AR model is evaluated based on the following equation (Xu and Lin 2017; Tian et al. 2021a, b):

$$Y_i = \sum_{j=1}^p f_j(X_{ij}) + \mu_i, \quad \mu_i \sim iid(0, \sigma^2), \quad (4)$$

where $f_j(X_{ij})$ is a nonparametric function fit to the data. The random error term (μ_i) has zero mean and variance of σ^2 . More AR details can be found in the study of Cui et al. (2010).

Evaluation and comparison of the models

RMSE, percentage of bias (PBIAS), Nash–Sutcliffe efficiency (NSE), and mean absolute error (MAE) are computed as follows (Legates et al. 1999; Moriasi et al. 2007a, b):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Q^{Pre} - Q^{Obs})^2}, \quad (5)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |Q^{Pre} - Q^{Obs}|, \quad (6)$$

$$NSE = 1 - \frac{\sum_{i=1}^N (Q^{Pre} - Q^{Obs})^2}{\sum_{i=1}^N (Q^{Obs} - \overline{Mean Q^{Obs}})^2}, \quad (7)$$

$$PBIAS = \left[\frac{\sum_{i=1}^N (Q^{Pre} - Q^{Obs})}{\sum_{i=1}^N Q^{Pre}} \right] * 100, \quad (8)$$

where Q^{Obs} is the observed rainfall, Q^{Pre} is the predicted rainfall, $\overline{Mean Q^{Obs}}$ is the mean of measured rainfall values, and N is the number of data samples (the number of test datasets). The lower the RMSE and MAE values, the higher the model performance. NSE ranges between $-\infty$ and 1, and the model with $NSE = 1$ is ideal. NSE is classified as unsatisfactory,

acceptable, satisfactory, good, and very good performance for $NSE \leq 0.4$, $0.40 < NSE \leq 0.50$, $0.50 < NSE \leq 0.65$, $0.65 < NSE \leq 0.75$, and $0.75 < NSE \leq 1.00$, respectively (Moriasi et al. 2007a, b). PBIAS can be used for model performance classification similar to NSE, and it shows the overall model underestimation or overestimation. Negative PBIAS indicates the model overestimation, while positive PBIAS represents underestimation (Legates and McCabe 1999).

Results

Most effective input scenario

Based on the results provided in Fig. 4, the input scenario (6) composed of the variables MIRH, MARH, MAT, MIT, SSH, and Eva had the highest r -values at both training and validation phases and showcased the most effective variables in rainfall prediction. Hence, the developed models were tested using the input scenario (6), the performance of which is discussed in the following section.

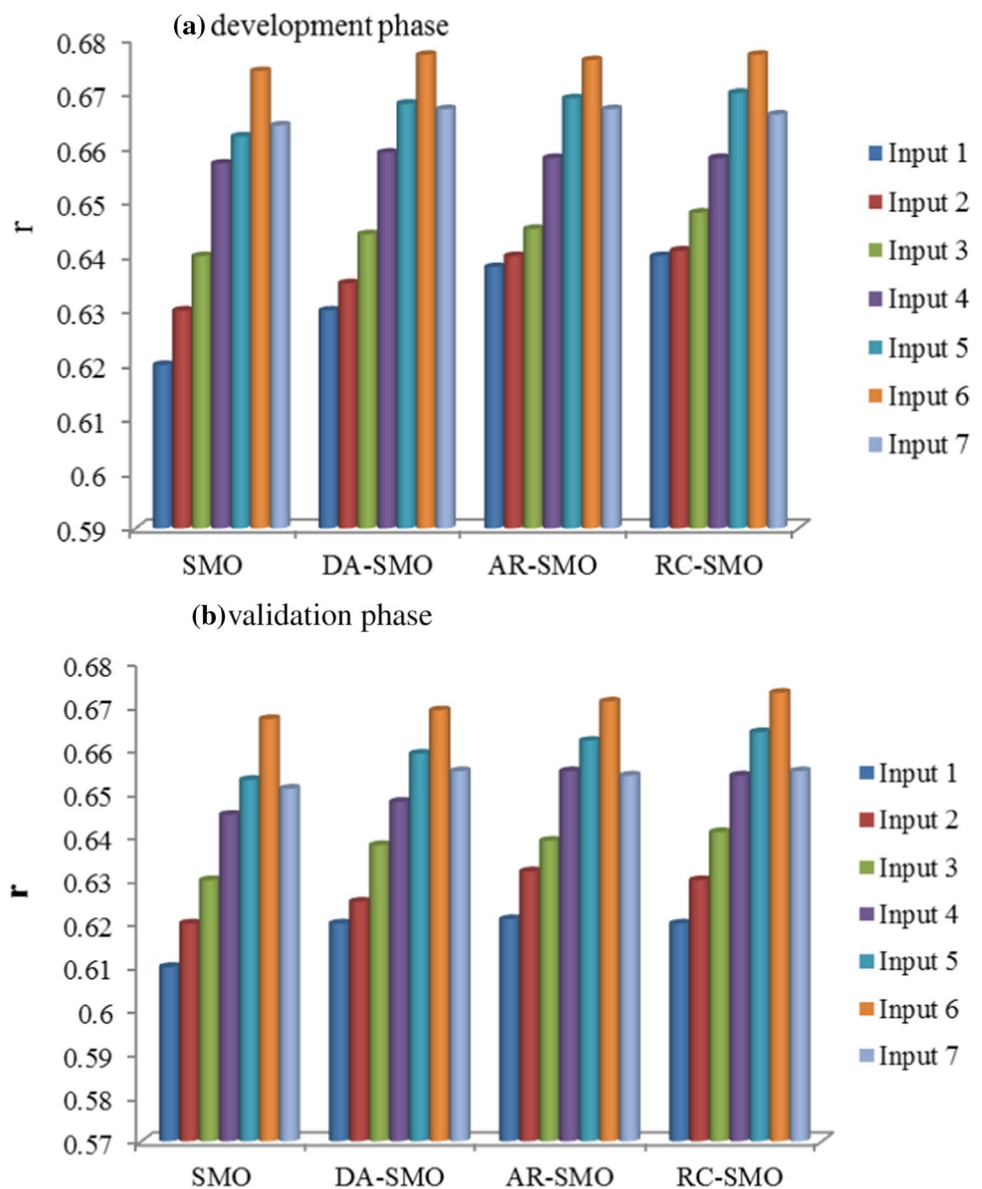
Model evaluation

After calibrating and validating the developed models, the performance of the models was evaluated using the test data. The observed versus estimated rainfall values by different models are visualized as time-varying and scatter plots in Fig. 5. It was observed that the standalone models predicted rainfall with an R^2 of 0.568 and were less accurate than other models. In contrast, the estimated rainfall values obtained by all the developed ensemble models (DA-, RC-, and AR-based models) were much closer to the measured values with R^2 of 0.739, 0.735, and 0.738. It was demonstrated that the DA algorithm was much more robust and had higher predictive power than RC and AR models.

It is observed that the DA-SMO and AR-SMO are able to predict the median rainfall (Q_{50}) much closer to the measured rainfall values (Fig. 6). The first quartile (Q_{25}) of all the developed models, except for the standalone SMO model, is close to the measured Q_{25} of the rainfall values, while the estimated values of the third quartile (Q_{75}) for all the developed models are higher than the measured Q_{75} of rainfall values. The SMO is inaccurate in estimating the minimum and maximum rates of rainfall. However, the ensemble DA-SMO is the most accurate model in estimating minimum and maximum rainfall rates, indicating the predictive power of the ensemble models.

In terms of the error metrics (Table 3) (i.e., RMSE and MAE), the SMO model had the lowest predictive power (RMSE = 22.61 mm, MAE = 13.98 mm). All the ensemble

Fig. 4 Determination of the most effective input scenario: (a) development phase, and (b) validation phase



models decreased the error occurring in the standalone models. It was observed that DA-, AR-, and RC-based SMO models decreased RMSE (MAE) of the standalone SMO by 21.1% (12.6%), 20.7% (11.8%), and 21.7% (13.0%), respectively. In terms of the NSE metric and based on the findings of Moriasi et al. (2007a, b), SMO with the NSE of 0.57 has a satisfactory performance ($0.5 < \text{NSE} \leq 0.65$), while three ensemble-based models have a good performance ($0.65 < \text{NSE} \leq 0.75$). The PBIAS indicator shows that all the models are classified as having good performance ($10\% < \text{PBIAS} \leq 15\%$) based on the findings of Moriasi et al. (2007a, b). Moreover, the PBIAS values for all the developed models are negative, indicating the overestimation of rainfall values.

Discussion

Choosing the best input variables is a difficult task in estimating rainfall. A number of studies rely on nonlinear methods such as gamma test in order to determine appropriate predictors in rainfall estimation (Ahmadi et al. 2015). However, the results of the present study indicate that employing a linear methodology by incorporating the correlation coefficient can also be a promising means to choose the input variables. It should be noted that while only a single combination for input is selected automatically in a nonlinear method, we constructed and examined different combinations for input variables in this article. This matter helped identify the input combination with the highest effectiveness and, at the same time, determine the impact of each input variable on the final results. The findings of the present

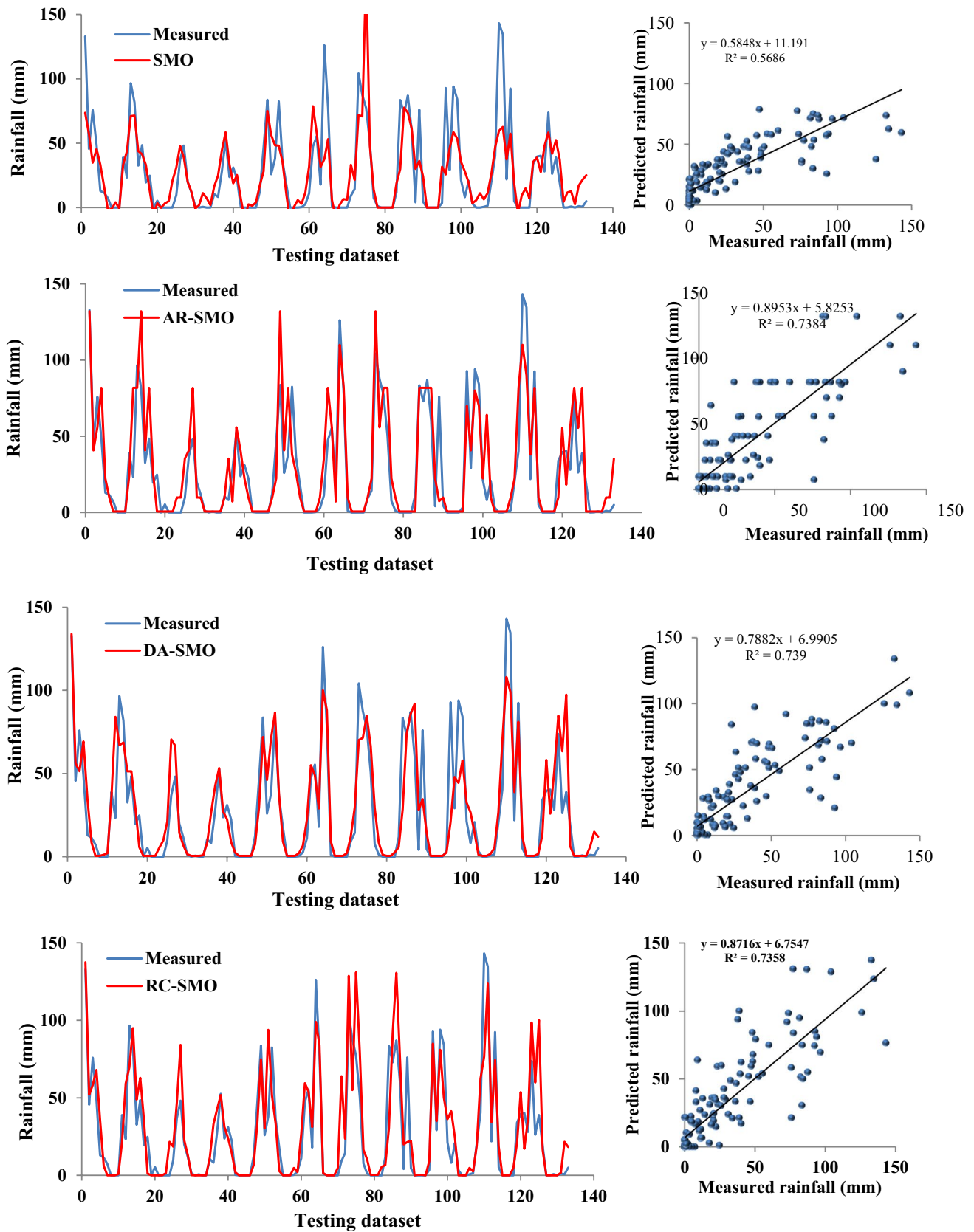
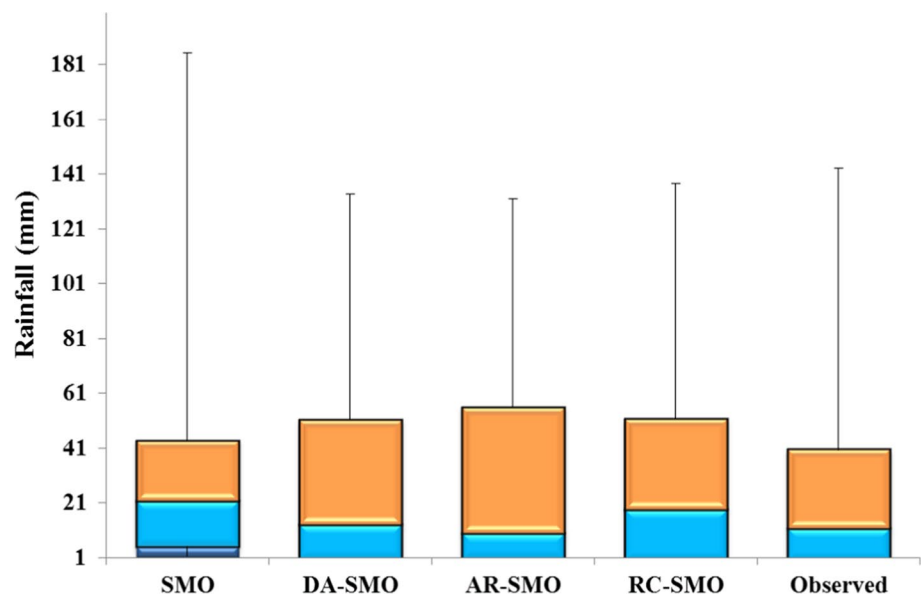


Fig. 5 Line-graphs and scatter plots for predicted vs. measured rainfall values during the validation phase

Fig. 6 Box plot of the measured and predicted rainfall values**Table 3** Performance evaluation of the models based on quantitative metrics

	RMSE (mm)	MAE (mm)	NSE	PBIAS%
SMO	22.61	0.67	0.570	-13.00
AR-SMO	18.92	0.81	0.703	-11.94
DA-SMO	17.73	0.81	0.730	-11.86
RC-SMO	18.80	0.79	0.705	-11.73

paper are supported by Khosravi et al. (2020a), who found the linear Pearson correlation coefficient effective in finding the best input variables for prediction models.

Due to the diversity in structure and complexity of the algorithms, DM models in various forms produce different results when employed for rainfall estimation. According to the findings of the present paper, the developed hybrid models, which are of nonlinear pattern in nature, had higher accuracy in modeling rainfall events than ML models in the standalone mode. In addition, the hybrid algorithms were more flexible than the latter. These findings support some research studies conducted previously, in which hybrid models were employed in order to simulate nonlinear hydrological processes, proving the superiority of hybrid models as they decreased both bias and variance (Hong et al. 2018; Jiang et al. 2021; Zuo et al. 2020; Ebrahimi et al. 2022). Another notable merit of hybrid models is that they can tackle the problem of over-fitting in regression modeling (Chen et al. 2020). For instance, among such studies, Bui et al. (2020) concluded that hybrid models functioning based on bagging (BA) (e.g., BA-random forest (RF), BA-M5P, BA-random tree (RT), and BA-reduced error pruning tree (REPT)) would enhance the capability of individual models

for predicting the indices of water quality. In addition, the study conducted by Khosravi et al. (2020b) found the BA-based models superior to the individual M5P, RF, and REPT when adopted for bedload transport rate modeling. In the field of landslide susceptibility mapping, Nguyen et al. (2019) observed that hybrid models performing on the basis of BA and DA would achieve better results than the alternating decision trees. Furthermore, the findings of Chen et al. (2020) indicated that hybrid J48 Decision Trees established based on BA and DA outperformed the standalone J48 model for the mapping of groundwater spring potential. All in all, BA- and DA-based hybrid models have proven viably satisfactory and reliable for prediction purposes. The results of the present paper further prove the robustness of some other hybrid models that are introduced as promising algorithms for modeling with prediction purposes.

As mentioned earlier, the main reasons behind the lower prediction performance of ML models with regard to natural phenomena like flood, drought, etc., lie in high non-linearity, stochastic process, and complexity of the occurrence of precipitation (e.g., Sánchez-Monedero et al. 2014; Hashim et al. 2016). As a result, natural phenomena prediction, particularly rainfall forecasting, continues to be subject to uncertainty. Furthermore, relevant studies have considered the strong correlation between rainfall and cloud information. For this purpose, total perceptible water, equivalent potential temperature, humidity, wind speed, wind direction, convective available potential energy, and convective inhibition have been introduced as a reliable combination for input. It can be concluded that some missing data of the mentioned parameters are the major reason for uncertainty with regard to the results of the present research.

It is suggested that future studies apply the proposed hybrid models to other hydrological fields such as water quality. Moreover, AR, DA, and RC algorithms can be utilized as ensemble learners in order to develop hybrid models using other ML techniques, e.g., decision trees and rule-based, lazy-based, and neuron-based algorithms, in different fields of geoscience.

Conclusion

Rainfall is one of the main components of hydrologic cycle that has a significant impact on the infiltration process, flood occurrences, soil erosion rate, water resources management, and irrigation system. Therefore, rainfall prediction is one of the hot topics today in the fields of hydrology and water resources management. Given that rainfall is highly stochastic and chaotic in behavior, it is not an easy task to predict it. The current study proposed standalone SMO models and three new ensemble-based algorithms of DA-SMO, AR-SMO, and RC-SMO for rainfall prediction in Kermanshah synoptic station, Iran. Moreover, different input scenarios were investigated to explore the effectiveness of different input combinations in the result. The main achievements drawn from the findings of this study can be summarized as follows:

1. Minimum relative humidity had the highest effect on rainfall prediction, followed by maximum temperature, relative humidity, minimum temperature, sunshine hours, evaporation, and wind speed.
2. Wind speed attenuated the predictive power of a model; in the present study, an input scenario featuring all input variables, except wind speed, was identified as the most effective scenario in rainfall estimation.
3. DA, AR, and RC algorithms enhanced the performance of the standalone SMO algorithm by about 12.8, 20.7%, and 21.7%, respectively, based on the RMSE metric.
4. While the standalone SMO algorithm had a satisfactory performance, ensemble models were of good performance in terms of NSE metric.
5. DA-SMO ensemble model outperformed other models, followed by RC-SMO, AR-SMO and standalone SMO models.
6. All of the developed models in the current study tended to overestimate the rainfall amount.
7. Hybrid algorithms were more accurate than the standalone model in capturing extreme values. The different machine learning algorithms explored in the current study are applicable to mountainous areas in Iran and may potentially yield accurate rainfall predictions for other mountainous regions in the world.

Author contributions HSS was involved in conceptualization, methodology, software, writing—original draft. HA contributed to supervision, review and editing. BA was involved in supervision, review and editing.

Funding Authors did not receive any funding for this paper.

Availability of data and materials Available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that there is no conflict of interest associated with this research or manuscript.

References

- Adefisan E (2018) Climate change impact on rainfall and temperature distributions over West Africa from three IPCC scenarios. *J Earth Sci Clim Change* 9:476
- Ahmadi A, Han D, Kakaei Lafdani E, Moridi A (2015) Input selection for long-lead precipitation prediction using large-scale climate variables: a case study. *J Hydroinf* 17(1):114–129
- Aksoy H, Dahamsheh A (2009) Artificial neural network models for forecasting monthly precipitation in Jordan. *Stoch Env Res Risk Assess* 23(7):917–931
- Atiquzzaman M, Kandasamy J (2018) Robustness of Extreme Learning Machine in the prediction of hydrological flow series. *Comput Geosci* 120:105–114
- Bui DT, Khosravi K, Tiefenbacher J, Nguyen H, Kazakis N (2020) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci Total Environ* 721:137612
- Chau KW, Wu CL (2010) A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *J Hydroinf* 12(4):458–473
- Chen W, Zhao X, Tsangaratos P, Shahabi H, Ilia I, Xue W, Ahmad BB (2020) Evaluating the usage of tree-based ensemble methods in groundwater spring potential mapping. *J Hydrol* 583:124602
- Chen Z, Liu Z, Yin L, Zheng W (2022) Statistical analysis of regional air temperature characteristics before and after dam construction. *Urban Clim*. <https://doi.org/10.1016/j.uclim.2022.101085>
- Cui X, Penh H, Wen S, Zhi L (2010) Component selection in the additive regression model. *Scand J Stat* 40(3):491–510
- Delleur JW, Kavvas ML (1978) Stochastic models for monthly rainfall forecasting and synthetic generation. *J Appl Meteorol* 17(10):1528–1536
- Ebrahimi M, Rostami H, Osouli A, Rosanna Saindon RG (2022) Use of Geoelectrical Techniques to Detect Hydrocarbon Plume in Leaking Pipelines, ASCE Lifelines Conference 2021–2022, Los Angeles
- Elbaz K, Shen S, Sun W, Yin Z, Zhou A (2020) Incorporating improved particle swarm optimization into ANFIS. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2020.2974058>
- Elbazz K, Shen S, Zhou A, Yuan D, Xu Y (2019) Optimization of EPB Shield Performance with Adaptive Neuro-Fuzzy Inference System and Genetic Algorithm. *Appl Sci* 9(4):780. <https://doi.org/10.3390/app9040780>
- Gao Q-Q, Bai Y-Q, Zhan Y-R (2019) Quadratic kernel-free least square twin support vector machine for binary classification problems. *J Oper Res Soc China* 7:539–559
- Ghiasi-Freeze J, Kadkhodaie-Ilkhchi A, Ziaii M (2012) Improving the accuracy of flow units prediction through two committee machine

- models: An example from the South Pars Gas Field, Persian Gulf Basin, Iran. *Comput Geosci* 46:10–23
- Ghumman AR, Ghazaw YM, Sohail AR, Watanabe K (2011) Runoff forecasting by artificial neural network and conventional model. *Alexandria Eng J* 50(4):345–350
- Hashim R, Roy C, Motamedi S, Shamshirband S, Petković D, Gocic M, Lee SC (2016) Selection of meteorological parameters affecting rainfall estimation using neuro-fuzzy computing methodology. *Atmos Res* 171:21–30
- Hashmi S, Halawani MO, AmirAhmad MB (2015) Model trees and sequential minimal optimization based support vector machine models for estimating minimum surface roughness value. *Appl Math Model* 39(3):1119–1136
- Hong H, Liu J, Bui DT, Pradhan B, Acharya TD, Pham BT, Zhu AX, Chen W, Ahmad BB (2018) Landslide susceptibility mapping using J48 decision tree with adaboost, bagging and rotation forest ensembles in the Guangchang area (China). *CATENA* 163:399–413
- Jiang S, Zuo Y, Yang M, Feng R (2021) Reconstruction of the Cenozoic tectono-thermal history of the dongpu depression, bohai bay basin, China: constraints from apatite fission track and vitrinite reflectance data. *J Petrol Sci Eng* 205:108809. <https://doi.org/10.1016/j.petrol.2021.108809>
- Kadam AK, Wagh VM, Muley AA, Umrikar BN, Sankhua RN (2019) Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. *Model Earth Syst Environ* 5:951–962
- Khosravi K, Barzegar R, Miraki S, Adamowski J, Daggupati P, Alizadeh MR, Pham B, Alami M (2020b) Stochastic modeling of groundwater fluoride contamination: introducing lazy learners. In press, *Groundwater*
- Khosravi K, Golkarian A, Boojij M, Barzegar R, Sun W, Yaseen ZM, Mosavi A (2021) Improving daily stochastic streamflow prediction: comparison of novel hybrid data-mining algorithms. *Hydrol Sci J* 66:1457–1474. <https://doi.org/10.1080/02626667.2021.1928673>
- Khosravi K, Golkarian A, Barzegar R, Aalami MT, Heddami S, Omidvar E, Keestra S, Opez-Vicente M (2022a) Multi-step-ahead soil temperature forecasting at multiple-depth based on meteorological data: integrating resampling algorithms and machine learning models. Under press, *Pedosphere*
- Khosravi K, Golkarian A, Melesse A, Deo R (2022b) Suspended sediment load modeling using advanced hybrid rotation forest based elastic network approach. *J Hydrol* 610:127963. <https://doi.org/10.1016/j.jhydrol.2022.127963>
- Khosravi K, Cooper, J. R., Daggupati, P., Pham, B. T., & Bui, D. T. (2020b). Bedload transport rate prediction: application of novel hybrid data mining techniques. *Journal of Hydrology*, 124774.
- Kisi O, Shiri J (2011) Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. *Water Resour Manage* 25(13):3135–3152
- Legates DR, McCabe GJ (1999) Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resour Res* 35:233–241
- Liu Y, Zhang K, Li Z, Liu Z, Wang J, Huang P (2020) A hybrid runoff generation modelling framework based on spatial combination of three runoff generation schemes for semi-humid and semi-arid watersheds. *J Hydrol (amsterdam)* 590:125440. <https://doi.org/10.1016/j.jhydrol.2020.125440>
- Liu B, Spiekermann R, Zhao C, Püttmann W, Sun Y, Jasper A, Uhl D (2022a) Evidence for the repeated occurrence of wildfires in an upper Pliocene lignite deposit from Yunnan, SW China. *Int J Coal Geol* 250:103924. <https://doi.org/10.1016/j.coal.2021.103924>
- Liu S, Liu Y, Wang C, Dang X (2022b) The Distribution characteristics and human health risks of high-fluorine groundwater in coastal plain: a case study in Southern Laizhou Bay. *Frontiers in Environmental Science, China*. <https://doi.org/10.3389/fenvs.2022b.901637>
- Luk KC, Ball JE, Sharma A (2000) A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting. *J Hydrol* 227(1–4):56–65
- Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007a) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans ASABE* 50(3):885–900
- Moriasi DN, Arnold JG, Van Liew MW, Binger RL, Harmel RD, Veith TL (2007b) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans ASABE* 50:885–900. <https://doi.org/10.13031/2013.23153>
- Nguyen H, Mehrabi M, Kalantar B, Moayedi H, Abdullahi MM (2019) Potential of hybrid evolutionary approaches for assessment of geohazard landslide susceptibility mapping. *Geomat Natural Hazard Risk* 10(1):1667–1693
- Nhu V-H, Khosravi K, Cooper JR, Karimi M, Kisi O, Pham BT, Lyu Z (2020) Monthly suspended sediment load prediction using artificial intelligence: testing of a new random subspace method. *Hydrol Sci J* 65(12):2116–2127
- Niranjan A, Haripriya DK, Pooja R, Sarah S, Deepa Shenoy P, Venugopal KR (2018) EKRV: Ensemble of kNN and Random Committee Using Voting for Efficient Classification of Phishing. In: *Advances in Intelligent Systems and Computing*, vol 713, pp 403–414
- Oyebode O, Stretch D (2019) Neural network modelling of hydrological systems: a review of implementation techniques. In: *Natural resource modelling*. Wiley, pp 1–14. <https://doi.org/10.1002/nrm.12189>
- Osouli A, Ebrahimi M, Alzamora D, Shoup HZ, Pagenkopf J (2022) Multi-criteria assessment of bridge sites for conducting PSTD/ISTD: case histories. *Transp Res Rec J Transp Res Board*. <https://doi.org/10.1177/03611981221108153>
- Pham BT, Le LM, Le T-H, Thi Bui K-T, Minh V, Prakhsh I (2020) Development of advanced artificial intelligence models for daily rainfall prediction. *Atmos Res* 237:104845
- Quan Q, Liang W, Yan D, Lei J (2022) Influences of joint action of natural and social factors on atmospheric process of hydrological cycle in Inner Mongolia. *China Urban Clim* 41:101043. <https://doi.org/10.1016/j.uclim.2021.101043>
- Ridwan WM, Sapitang M, Aziz A, Kushiar KF, Ahmed AN, El-Shafie A (2021) Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. *Ain Shams Eng J* 12(2):1651–1663
- Samantaray S, Tripathy O, Sahoo A, & Ghose DK, (2020). Rainfall forecasting through ANN and SVM in Bolangir Watershed, India. In: *smart intelligent computing and applications*. Springer, Singapore (pp. 767–774)
- Sánchez-Monedero J, Salcedo-Sanz S, Gutiérrez PA, Casanova-Mateo C, Hervás-Martínez C (2014) Simultaneous modelling of rainfall occurrence and amount using a hierarchical nominal–ordinal support vector classifier. *Eng Appl Artif Intell* 34:199–207. <https://doi.org/10.1016/j.engappai.2014.05.016>
- Sheikh Khozani Z, Bonakdari H, Ebtehaj I (2017) An analysis of shear stress distribution in circular channels with sediment deposition based on Gene Expression Programming. *Int J Sediment Res* 32(4):575–584
- Sivapragasam C, Liong SY, Pasha MFK (2001) Rainfall and runoff forecasting with SSA–SVM approach. *J Hydroinf* 3(3):141–152
- Tian H, Qin Y, Niu Z, Wang L, Ge S (2021a) Summer maize mapping by compositing time series sentinel-1a imagery based on crop growth cycles. *J Indian Soc Remote Sens* 49(11):2863–2874. <https://doi.org/10.1007/s12524-021-01428-0>
- Tian H, Wang Y, Chen T, Zhang L, Qin Y (2021b) Early-season mapping of winter crops using sentinel-2 optical imagery. *Remote*

- Sens (basel, Switzerland) 13(19):3822. <https://doi.org/10.3390/rs13193822>
- Ting, KM, Witten IH, (1997) stacking Bagged and Dagged Models. In: Fourteenth international Conference on Machine Learning, San Francisco, CA, 367-375
- Toth E, Brath A, Montanari A (2000) Comparison of short-term rainfall prediction models for real-time flood forecasting. *J Hydrol* 239(1–4):132–147
- Wang D, Hagen SC, Alizad K (2013) Climate change impact and uncertainty analysis of extreme rainfall events in the Apalachicola River basin, Florida. *J Hydrol* 480:125–135
- Wang S, Zhang K, Chao L, Li D, Tian X, Bao H, Xia Y (2021) Exploring the utility of radar and satellite-sensed precipitation and their dynamic bias correction for integrated prediction of flood and landslide hazards. *J Hydrol (amsterdam)* 603:126964. <https://doi.org/10.1016/j.jhydrol.2021.126964>
- Wang Y, Cheng H, Hu Q, Liu L, Jia L, Gao S, Wang Y (2022) Pore structure heterogeneity of Wufeng-Longmaxi shale, Sichuan Basin, China: Evidence from gas physisorption and multifractal geometries. *J Pet Sci Eng* 208:109313. <https://doi.org/10.1016/j.petrol.2021.109313>
- Witten IH, Frank E (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Second edn, p 558
- Wu J, Liu M, Jin L (2010) A hybrid support vector regression approach for rainfall forecasting using particle swarm optimization and projection pursuit technology. *Int J Comput Intell Appl* 9(02):87–104
- Xie W, Li X, Jian W, Yang Y, Liu H, Robledo LF, Nie W (2021a) A novel hybrid method for landslide susceptibility mapping-based geodetector and machine learning cluster: a case of Xiaojin County China. *ISPRS Int J Geo-Inf* 10(2):93. <https://doi.org/10.3390/ijgi10020093>
- Xie W, Nie W, Saffari P, Robledo LF, Descote, P.,... Jian, W. (2021b) Landslide hazard assessment based on Bayesian optimization–support vector machine in Nanping City. *China Nat Hazard (dordrecht)* 109(1):931–948. <https://doi.org/10.1007/s11069-021-04862-y>
- Xu B, Lin B (2017) Does the high–tech industry consistently reduce CO₂ emissions? Results from nonparametric additive regression model. *Environ Impact Assess Rev* 63:44–58
- Yaseen ZM, Ebtehaj I, Kim S, Sanikhani H, Asadi H, Ghareb MI, Shahid S (2019) Novel hybrid data-intelligence model for forecasting monthly rainfall with uncertainty analysis. *Water* 11(3):502
- Yevjevich V (1987) Stochastic models in hydrology. *Stoch Hydrol Hydraul* 1(1):17–36
- Yin L, Wang L, Keim BD, Konsoer K, Zheng W (2022a) Wavelet analysis of dam injection and discharge in three gorges dam and reservoir with precipitation and river discharge. *Water* 14(4):567. <https://doi.org/10.3390/w14040567>
- Yin L, Wang L, Zheng W, Ge L, Tian J, Liu, Y.,... Liu, S. (2022b) Evaluation of empirical atmospheric models using swarm-c satellite data. *Atmosphere* 13(2):294. <https://doi.org/10.3390/atmos13020294>
- Zhang K, Ali A, Antonarakis A, Moghaddam M, Saatchi S, Tabatabaenejad A, Moorcroft P (2019a) The sensitivity of North American terrestrial carbon fluxes to spatial and temporal variation in soil moisture: an analysis using radar-derived estimates of root-zone soil moisture. *J Geophys Res Biogeosci* 124(11):3208–3231. <https://doi.org/10.1029/2018JG004589>
- Zhang K, Wang S, Bao H, Zhao X (2019b) Characteristics and influencing factors of rainfall-induced landslide and debris flow hazards in Shaanxi Province, China. *Nat Hazard* 19(1):93–105. <https://doi.org/10.5194/nhess-19-93-2019>
- Zhang K, Shalehy MH, Ezaz GT, Chakraborty A, Mohib KM, Liu L (2022) An integrated flood risk assessment approach based on coupled hydrological-hydraulic modeling and bottom-up hazard vulnerability analysis. *Environ Model Softw* 148:105279. <https://doi.org/10.1016/j.envsoft.2021.105279>
- Zhao F, Song L, Peng Z, Yang J, Luan G, Chu C, Xie Z (2021a) Night-time light remote sensing mapping: construction and analysis of ethnic minority development index. *Remote Sens (basel, Switzerland)* 13(11):2129. <https://doi.org/10.3390/rs13112129>
- Zhao F, Zhang S, Du Q, Ding J, Luan G, Xie Z (2021b) Assessment of the sustainable development of rural minority settlements based on multidimensional data and geographical detector method: a case study in Dehong China. *Socio-Econ Plan Sci*. 78:101066
- Zhao X, Xia H, Pan L, Song H, Niu W, Wang R, Qin Y (2021c) Drought monitoring over yellow river basin from 2003–2019 using reconstructed modis land surface temperature in google earth engine. *Remote Sens (basel, Switzerland)* 13(18):3748. <https://doi.org/10.3390/rs13183748>
- Zhu B, Zhong Q, Chen Y, Liao S, Li Z, Shi K, Sotelo MA (2022a) A novel reconstruction method for temperature distribution measurement based on ultrasonic tomography. *IEEE Trans Ultrason Ferroelectr Freq Control*. <https://doi.org/10.1109/TUFFC.2022.3177469>
- Zhu Z, Zhu Z, Wu Y, Han J (2022b) A Prediction method of coal burst based on analytic hierarchy process and fuzzy comprehensive evaluation. *Front Earth Sci (lausanne)*. <https://doi.org/10.3389/feart.2021.834958>
- Zuo Y, Jiang S, Wu S, Xu W, Zhang J, Feng R, Santosh M (2020) Terrestrial heat flow and lithospheric thermal structure in the Chagan depression of the Yingen-Ejinaqi Basin, north central China. *Basin Res* 32(6):1328–1346. <https://doi.org/10.1111/bre.12430>

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.