Natural Hazards (2026) 122:418
https://doi.org/10.1007/s11069-026-08147-0

ORIGINAL PAPER

A stacked generalization-based multi-model ensemble framework for precipitation downscaling and drought characterization in a semi-arid region

Amirhossein Mirdarsoltany[1] · Matineh Imani Borhan[2] · Leila Rahimi[3] · Carlo De Michele[4]

Received: 10 September 2025 / Accepted: 2 April 2026

Abstract
Drought, as a recurrent climatic phenomenon driven by prolonged precipitation deficits, has intensified in recent decades, causing widespread impacts on agriculture, water resources, and socio-environmental sustainability. This trend is particularly evident in arid and semi-arid regions such as Iran, highlighting the critical need for continuous monitoring and assessment. This study proposes a novel downscaling framework using Stacked Generalization, which differs from traditional ensemble methods by using base model predictions as inputs to a secondary model. This two-stage approach captures complex dependencies, leading to improved downscaling performance. The refined outputs were then used to analyze drought characteristics based on the best-performing Global Climate Models (GCMs). Then, the next objective is to better determine future drought characteristics by utilizing the more accurate results obtained from the proposed downscaling approach. It further aims to improve the identification of future drought characteristics using the more accurate results obtained from the proposed downscaling approach. The results indicated that the Stacked method consistently outperformed the individual base ones (MLP, SVR, and RF), achieving the highest Nash–Sutcliffe Efficiency (NSE) across all stations and climate models, and exhibiting the lowest Mean Squared Error (MSE) compared to the other methods. Additionally, the Standardized Precipitation Index (SPI) was calculated using a parametric method at 3-, 6-, and 12-month timescales. The findings indicated that, while short-term drought characteristics remain stable, long-term droughts, as represented by SPI-12, are projected to become longer and more severe, particularly in certain regions.

Keywords Machine learning · Stacked generalization · Downscaling · SPI · Global climate models

Extended author information available on the last page of the article

Published online: 30 April 2026

1 Introduction

An extreme event like drought, is one of the most widespread challenges to the resilience of socio-environmental systems, imposing serious adverse impacts on social, economic, and environmental sustainability (Tirivarombo et al. 2018; Laimighofer and Laaha 2022; Sunusi and Auliana 2025). Drought is a natural climatic recurring event caused by over prolonged periods, which is typically characterized by precipitation levels falling below the long-term average, notably in the case of meteorological drought (Dai 2011; Huang et al. 2016). In recent decades, droughts driven by precipitation deficits and extreme dryness have led to cascading impacts such as crop failures, reduced agricultural productivity, decreased water availability, food shortages, loss of hydropower production, wildfires, and other significant damages worldwide (Laimighofer and Laaha 2022; Bhardwaj et al. 2025; Sunusi and Auliana 2025). In this regard, assessment and monitoring of its effects and progress are essential.

Over the past few decades, droughts have become more frequent and intense in arid and semi-arid countries (Shayeghi et al. 2024), such as Iran in the northern hemisphere (Kazemzadeh et al. 2022). Among the different types of droughts, including hydrological, agricultural, socio-economic, and meteorological (Yerdelen et al. 2021; He et al. 2023; Behfar et al. 2024), the meteorological drought may occur earlier than the other types (Kazemzadeh and Malekian 2016), which is caused by precipitation deficiency over a spatial and temporal scales (Fiorillo and Guadagno 2010; Nabaei et al. 2019; Banfi et al. 2024). The Standardized Precipitation Index (SPI), developed by McKee et al. 1993 to define and monitor droughts, special for determining meteorological drought (Svoboda et al. 2012). The SPI is a useful tool for drought contingency planning due to its simplicity and ability to identify the beginning and end of drought events. SPI is commonly used to assess meteorological drought over various timescales and enables comparison across regions with different climates (Naresh Kumar et al. 2009; Sobral et al. 2019; Lorenzo et al. 2024; Guria et al. 2025). It reflects soil moisture deficit at short timescales and relates to groundwater or reservoir levels at longer ones. SPI in 3- or 6-month moving average can be considered as a meteorological drought indicator and at 12 months or more time windows for hydrological drought analysis and applications (Lorenzo et al. 2024; Zeleňáková et al. 2025). This index indicates as a useful tool in drought monitoring in Iran (Kazemzadeh and Malekian 2016; Shirvani and Landman 2022; Mirdarsoltany et al. 2025), a country with variety of climate conditions.

Monitoring of drought characteristics (e.g., frequency, intensity, etc.) due to their variabilities under climate change is becoming increasingly important. Addressing these challenges requires both historical data to understand drought patterns in the past, and future climate projections to evaluate potential risks. Global Climate Models (GCMs) are the most advanced tools to generate future spatiotemporal climate variable projections through large-scale resolution (Vasiliades et al. 2009; Yeganeh-Bakhtiary et al. 2022). Although general GCMs are widely used to evaluate the impacts of climate change (Tabari et al. 2021), they operate at coarse temporal and spatial resolutions, typically ranging from 100 to 300 km, and often exhibit consistent biases when compared to observational data (Takayabu et al. 2016; Ahmed et al. 2019). Downscaling and bias correction play a vital role on addressing the spatial scale mismatch between the coarse resolution outputs of GCMs and fine resolution of hydro climatic variables in local scale, which are required for hydrology studies (Wilby and Wigley 1997; Hertig and Jacobeit 2008; Sachindra et al. 2018). Indeed, down-

scaling is an avenue to derive high-resolution data from coarse-resolution GCMs. Downscaling techniques are classified into dynamical and statistical methods (Baghanam et al. 2020). Statistical downscaling (SD) was used here due to its straightforward nature and reduced computational expense in comparison with dynamic downscaling (DD) (Zhang et al. 2022). SD methods can be broadly categorized into three types: regression-based techniques, weather classification-based approaches, and weather generators (Wilby et al. 2004). Regression-based statistical downscaling techniques have gained popularity out of the above three categories owing to their simplicity in application. The regression techniques are widely used in statistical downscaling (Sachindra et al. 2018), including Multi Regression (MLR) (Chen et al. 2010; Nishant et al. 2023), Artificial Neural Networks (ANNs) (Ahmed et al. 2015; Laddimath and Patil 2019; Nourani et al. 2019; George and Athira 2024), Support Vector Machine (SVM) (Anandhi et al. 2008; Yazdian et al. 2023; Kalu et al. 2024), Genetic Programming (GP) (Coulibaly 2004; Sachindra et al. 2018; Kumar et al. 2021), Relevance Vector Machine (RVM) (Okkan and Inan 2015), Random Forest (RF) (Kalu et al. 2024), and Gene Expression Programming (GEP) (Sachindra et al. 2016; Pour et al. 2023). Owing to the learning abilities from data, techniques such as ANN, SVM, RVM, RF, and GP are often called machine learning techniques.

Several previous studies have focused on the performance of different downscaling approaches developed by machine learning techniques. Coulibaly (2004) showed that GP-based downscaling models performed better in simulating both daily minimum and maximum temperature in comparison to MLR-based downscaling models. Ghosh and Mujumdar (2008) proposed a statistical downscaling method using Relevance Vector Machine (RVM) with sparse Bayesian learning to model monsoon streamflow, the results showed that the RVM model outperformed SVM and projected a decreasing future streamflow trend. Duhan and Pandey (2015) utilized three downscaling techniques, namely, multiple linear regression, artificial neural network, and least square support vector machine (LS-SVM), for the projections of monthly maximum and minimum air temperature in the Tons River basin, which is a sub-basin of the Ganges River in Central India. Their results show that all the models are able to simulate temperature; however, LS-SVM models perform slightly better than ANN and MLR. Sachindra et al. (2018) evaluated four ML techniques, including GP, ANN, SVM, and RVM, for downscaling reanalysis data to monthly precipitation in Victoria, Australia. Their findings indicated that SVM and RVM-based algorithms, so that Polynomial kernels, led to the best performance for both SVM and RVM across all climate regimes. Nishant et al. (2023) compared ML downscaling using multi-layer perceptron (MLP) with dynamical downscaling for refining precipitation data in Australia. The MLP approach, integrating coarse-scale meteorological inputs with fine-scale surface data, outperforms DD in capturing precipitation climatology and extremes, especially at higher resolutions. Results highlight that ML can offer superior accuracy and lower computational cost than DD, even when using coarser input data. Kalu et al. (2024) used four different machine learning algorithms, including SVM, partial least squares, Gaussian process regression, and random forest, to downscale the original Gravity Recovery and Climate Experiment (GRACE) data. Their result demonstrated that the Gaussian process regression algorithm gave the best result. However, despite their effectiveness, these models often suffer from limitations such as overfitting, limited generalizability, or reduced interpretability, which encouraged the development of a more robust and integrated approach outlined in this study.

Since the identification of the most suitable GCM depends on the applied downscaling technique, and different techniques can yield varying results, selecting an appropriate downscaling method is critical. This decision directly affects the estimation of key drought characteristics, such as duration, frequency, and intensity, and consequently, the reliability of drought projection and risk assessment, influencing early warning systems, adaptation planning, and efforts to mitigate future drought-related impacts and stress. To enhance the performance of traditional machine learning methods in downscaling, the present study introduces a *novel downscaling* framework based on Stacked Generalization (Wolpert 1992). While previous studies have commonly employed ensemble methods, combining multiple machine learning methods to enhance downscaling accuracy, this approach is fundamentally different. Rather than merging ML models or averaging their outputs, the SG method leverages the individual predictions generated by various base models as inputs to a secondary learning stage. These predictions are compiled into a new dataset, which serves as the input for the SG method. This two-stage framework allows the SG method to learn complex relationships and dependencies among the outputs of the base models, thereby achieving superior downscaling results in our study without explicitly combining the models themselves. This structure represents a novel alternative to traditional ensemble techniques in the downscaling domain. Subsequently, drought characteristics were derived using the outputs of the best-performing GCM model identified through this enhanced downscaling method. Thus, this research's primary objectives and novel aspects are as follows.

1) Employing a novel model called the Stacked Generalization algorithm with the based models SVM, MLP, and RF to downscale multiple GCMs precipitation in the Latyan basin (Iran).
2) Comparing the outputs of based machine learning methods (SVM, MLP, and RF) with the new method (SG) to demonstrate Stacked Generalization capability in precipitation downscaling.
3) Better determination of drought characteristics in the future through more accurate results obtained from downscaling.

The rest of the paper is organized as follows: after describing the study area (Sect. 2), we introduce the proposed framework (Sect. 3), present the results (Sect. 4), and finally discuss the conclusions and future directions (Sect. 5).

2 Study area and data

Iran, is a vast country in the Middle East spanning 1.684 million square kilometers. It has faced significant climatic challenges, particularly those leading to severe water shortages (Saemian et al. 2022). Indicators of reduced surface water include drying rivers, lakes, and wetlands, along with deforestation, soil erosion, and frequent sand and dust storms under climate change. The rising demand for water to support a growing population, coupled with frequent climate change-driven meteorological droughts, is intensifying the depletion of water resources across the country. In this study, the impacts of climate change on drought characteristics within Latyan basin are evaluated (Fig. 1).

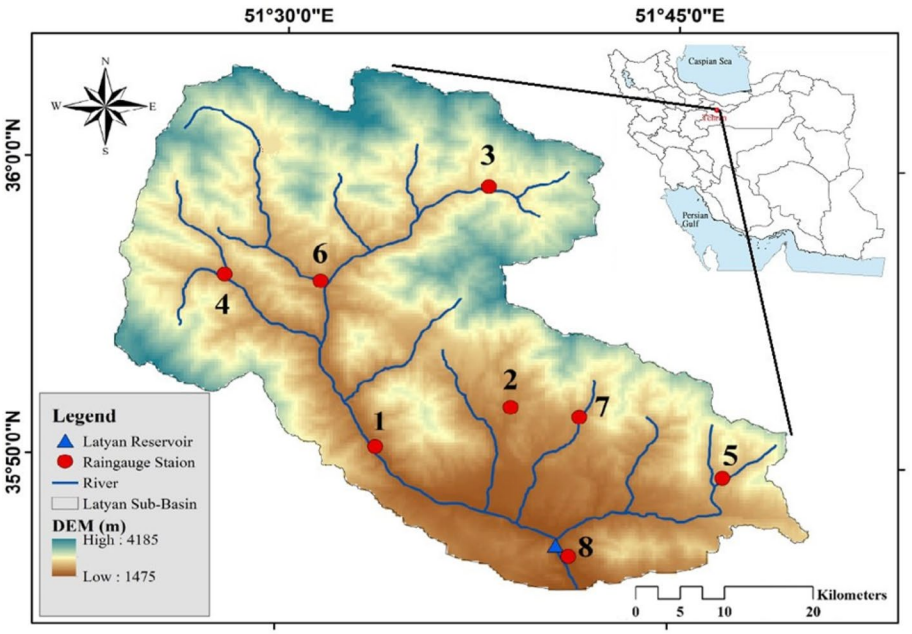

Fig. 1 Map of Study area: The Latyan basin (Iran) with the location of raingauge stations

The Latyan Basin, covering an area of 728 square kilometers, is situated adjacent to the Karaj Basin. It lies between latitudes 35° 45′ N and 36° 6′ N and longitudes 51° 22′ E and 51° 55′ E. The basin experiences an average annual rainfall of 320 mm and an average temperature of 11.4 °C. The elevation within the basin ranges from a minimum of 1,472 m above sea level to a maximum of 4,297 m. The Latyan Reservoir, located at the basin's outlet, plays a critical role in meeting the municipal water supplies of Tehran, the capital city of Iran, which has a population of approximately 11 million. Given the reservoir's importance for water supply, studying drought characteristics in this region is of particular significance.

In this study, both observational and climate model datasets were utilized for drought monitoring in the historical and future in the Latyan Basin. The observed dataset consists of monthly precipitation records covering the period from 1970 to 2023. These observed precipitation values serve as the target variable for the modeling process and provide a reliable foundation for capturing historical variability and drought conditions in the study area. The monthly precipitation data were acquired from the Water Resources Management Company (https://www.wrm.ir) for selected stations within the basin. Information of selected stations is given in Table 1.

Coupled Model Intercomparison Project Phase 6 (CMIP6) provides state-of-the-art climate models, offering improved simulations of historical and future climate conditions compared to previous phases. In this study, four GCMs (1970–2023 for the historical period and 2030–2100 for the future period) from CMIP6 are utilized. Table 2 summarizes the characteristics of the selected models. To evaluate the accuracy of the CMIP6 models, the historical period 1970 to 2023 was used as a baseline period. The selection of these specific models was guided by a main criterion. Preference was given to models that have shown

Table 1 Rain gauge stations information for Latyan basin

Station number	Station name	Latitude (N)	Longitude (E)	Elevation (m)	Average monthly precipitation (mm)
1	Rodak	35° 50′	51° 33′	1705	49.04
2	Kandsofla	35° 51′	51° 36′	1945	42.89
3	Garmabdar	35° 59′	51° 37′	2435	56.23
4	Ahar	35° 56′	51° 27′	2087	54.92
5	Lavasanbozorg	35° 49′	51° 47′	2195	45.01
6	Fasham	35° 55′	51° 31′	1966	60.31
7	Afjeh	35° 51′	51° 41′	2048	55.23
8	Latyan	35° 46′	51° 41′	1580	34.47

Table 2 List of GCMs used for downscaling

Name	Country	Resolution	Developer
ACCESS-CM2	Australia	1.25° × 1.87°	Commonwealth Scientific and Industrial Research Organization (CSIRO)
CanESM5	Canada	2.8° × 2.8°	Canadian Centre for Climate Modeling andAnalysis
HadGEM3-GC31-LL	United Kingdom	1.88° × 1.25°	Meteorological Office Hadley Centre
MIROC6	Japan	1.40° × 1.40°	National Institute for Environmental Studies, The University of Tokyo

reliable performance in simulating temperature and precipitation in regions with similar geographic or climatic characteristics, as demonstrated in past performance assessments (Guo et al. 2021; Zareian et al. 2024).

3 Methodology

To monitor drought characteristics in the historical (from 1970 to 2023) and future (from 2030 to 2100) under climate change in the case study, first are collected the historical data and projected data from GCMs. Next, machine learning methods (ANN, SVR, and RF) are developed by preprocessing the dataset, splitting it into training, validation, and testing sets, and training the models accordingly. Then, Stacked Generalization (Stacking) approach was

implemented using a set of base learners including MLP, SVR, and RF to enhance the accuracy of precipitation downscaling in historical period. The performance of these models is compared using evaluation metrics (e.g. NSE, MSE, and KGE) to select the most effective downscaling approach and GCM combination. Using the selected model, precipitation is predicted for future periods (2030–2100) under different climate scenarios (SSP1-2.6 and SSP5-8.5). Finally, the SPI at 3-, 6-, and 12-month timescales was computed to assess drought characteristics during both historical and future periods. The flowchart of the proposed method is presented in Fig. 2.

3.1 Preprocessing data

The observed dataset consists of monthly precipitation records from 1970 to 2023. Long-term observations are essential for drought analysis as they capture climate variability while

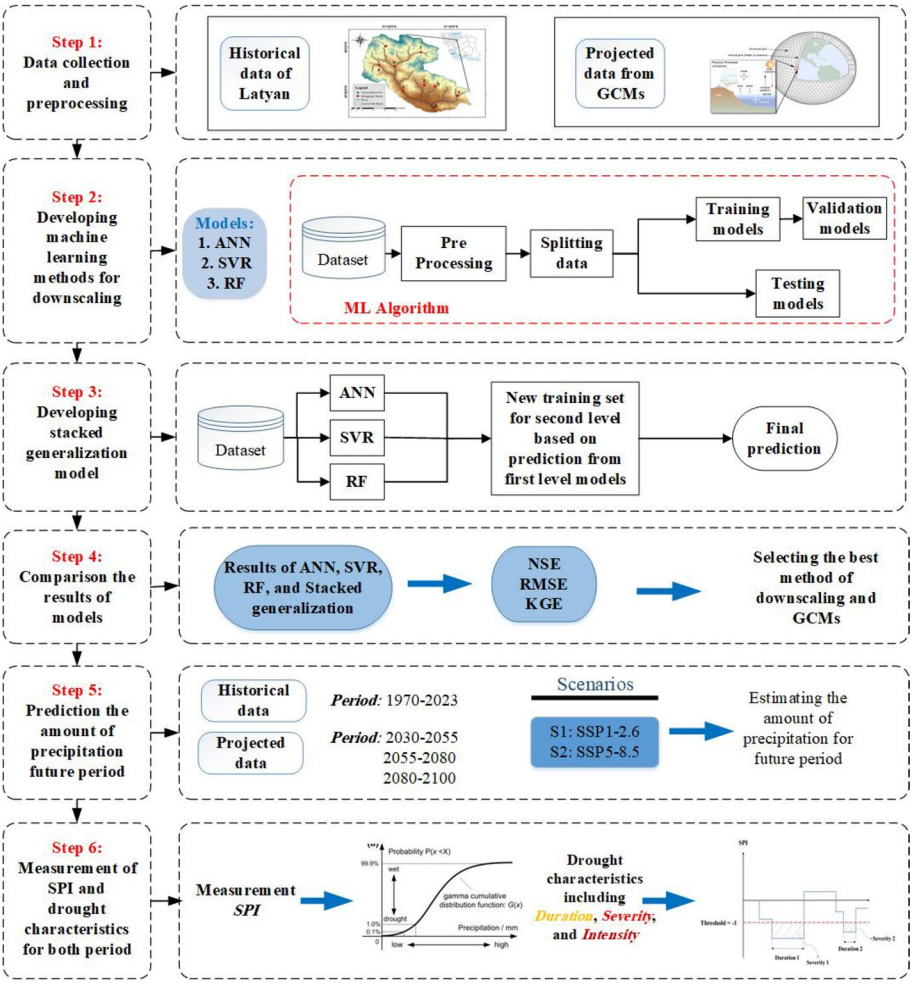

Fig. 2 The flowchart of the suggested approach

minimizing the need for data reconstruction. Prior to analysis, the dataset was screened for missing values and inconsistencies. In the time series, missing precipitation values were reconstructed using an Artificial Neural Network (ANN). A three-layer ANN with one hidden layer was implemented following Dariane and Borhan 2025, preserving the statistical characteristics of the observed precipitation series.

3.2 Downscaling

Downscaling methods are well-known approaches to address the mismatch of spatial resolution between the coarse outputs of GCMs and the finer-scale hydro climatic variables at the catchment level. In this study, multiple statistical downscaling approaches were employed, including MLP, SVR, RF. In addition, stacked generalizations were introduced to improve the accuracy of downscaling.

Initially, the input variables considered for the models included GCM-derived included surface air temperature, near-surface air temperature, geopotential height at multiple pressure levels (1000, 925, 850, 700, 600, 500, 400, and 300 hPa), relative and specific humidity at the surface and pressure levels, near surface wind speed, sea level pressure, surface air pressure. As a first screening step, the Pearson correlation coefficient between each candidate predictor and the observed precipitation series was calculated. Predictors exhibiting statistically significant correlations at the 95% confidence level were retained, resulting in a reduced subset of five potential predictors, including precipitation (pr), surface air pressure (ps), surface temperature (ts), sea level pressure (psl), and relative humidity (hur). While Pearson correlation was used as an initial screening tool to reduce dimensionality, it does not preclude the presence of nonlinear relationships, which are subsequently captured by the machine learning models.

However, based on a trial-and-error process and sensitivity analysis, only three predictors, including precipitation (pr), surface air pressure (ps), and surface temperature (ts), were selected as the most relevant inputs, as they yielded better performance and reduced redundancy in the models. The target variable for downscaling is monthly precipitation observed at local gauge stations. These observations were used to train and validate the models to ensure accurate projection of localized precipitation patterns from coarse GCM outputs. A brief description of each method is given in the following sections.

3.2.1 Multilayer perceptron (MLP)

Artificial Neural Networks (ANNs) are modeled after the human brain's biological nervous system (Agatonovic-Kustrin and Beresford 2000). ANNs have the ability to learn and generalize the relationships between input and output data, which allows them to tackle complex problems, whether linear or nonlinear (such as precipitation downscaling). This capability gives ANNs an edge over traditional statistical regression methods like Multi Linear Regression (MLR) and Generalized Linear Models (GLM).

In hydrology and climatology studies, the MLP is one of the most commonly used ANN architectures (Ahmed et al. 2015). An MLP is typically a feed-forward neural network consisting of an input layer, one or more hidden layers, and an output layer (Fig. 3). Mathematically, an MLP neural network can be expressed by the equation in Eq. (1) (Kim and Valdés 2003), where, y_k and x_j represent the outputs and inputs to the network, respectively.

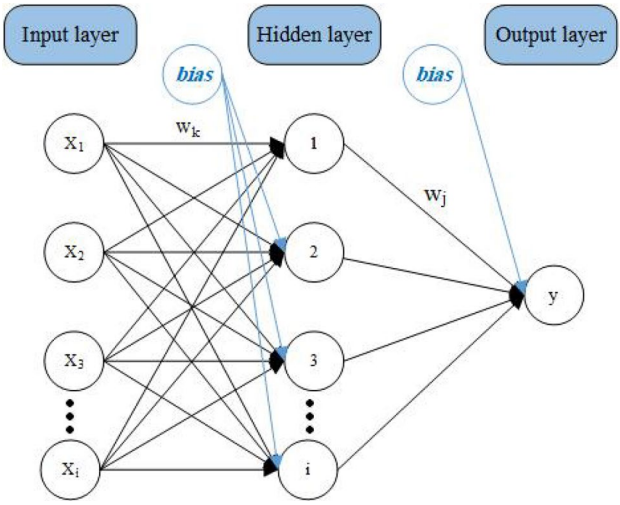

Fig. 3 ANN Structure

$$y_k = f_2\left[\sum_{i=1}^{m} w_{ki} f_1\left(\sum_{j=1}^{n} w_{ij} x_j + b_i\right) + b_k\right] \tag{1}$$

where, the subscripts i, j, and k correspond to the hidden, input, and output layers, respectively. The variables n and m denote the number of neurons in the input and hidden layers. The weights w_{ij} represent the connections between neurons in the input and hidden layers, while w_{ki} are the weights connecting neurons in the hidden and output layers. The functions f_1 and f_2 are the activation functions for the hidden and output layers, respectively. Additionally, b_i and b_k are the biases associated with the neurons in the hidden and output layers, respectively (Mekanik et al. 2013).

In this study, the sigmoid activation function was used in the single hidden layer to handle the nonlinearities in hydrologic systems. The number of neurons in the hidden layer is determined by the trial-and-error method. Also, the Levenberg–Marquardt (LM) algorithm, the most common calibration algorithm, was implemented to train the ANN.

3.2.2 Support vector regression (SVR)

Support Vector Machines (SVMs) (Vapnik 1998) are powerful tools for classification and regression tasks, offering a solid theoretical foundation, particularly through structural risk minimization (SRM), which balances model complexity and training error to improve generalization. SVMs aim to find an optimal hyperplane that maximizes the margin between distinct classes, using kernel functions like linear, polynomial, radial basis function (RBF), and sigmoid to handle non-linear reparability by mapping data into higher-dimensional spaces (Dariane and Borhan 2024). In SVR, the goal is to predict continuous outcomes by minimizing errors within a margin of tolerance, rather than aiming for absolute precision. The model parameters are determined by minimizing an empirical risk, with a loss function that allows small deviations from the true values without penalty. Slack variables pro-

vide flexibility in handling small violations of the margin, and a cost function controls the trade-off between margin maximization and error minimization. The regression function is expressed using Lagrange multipliers and kernel functions, enabling efficient computation even for high-dimensional data (Sachindra et al. 2018). More details in Fig. 4.

In this study, radial basis function (RBF) type kernel machine works better than other types of kernels and regularization parameter (C) and Epsilon are determined 1 and 0.1, respectively by the trial-and-error method.

3.2.3 Random forest (RF)

RF is a popular machine learning technique that can be applied to the task of downscaling, particularly in the context of climate modeling, meteorology, or geospatial data analysis. Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. RF leverages the concept of *bagging* (Bootstrap Aggregating), where different subsets of the data are used to train multiple trees, and the final prediction is made by averaging (or voting) across all the trees. It works by selecting random subsets of features at each decision node, which helps to reduce overfitting and increase the model's generalization capabilities. Random Forest can capture complex, non-linear relationships between the input (coarse-scale) and output (fine-scale) variables, which is often essential in environmental modeling (Hutengs and Vohland 2016).

3.2.4 Stacked generalization

Stacked Generalization, originally introduced by Wolpert (1992), is implemented as a two-level ensemble learning framework to improve precipitation downscaling by optimally combining multiple base learners while reducing model bias and variance. If $D = \{(X_{t.yt})\}_{t=1}^{T}$ denote the complete historical dataset, where $X_t = \{x_{t.1}.x_{t.2}.\ldots.x_{t.p}\}$ denote the vector

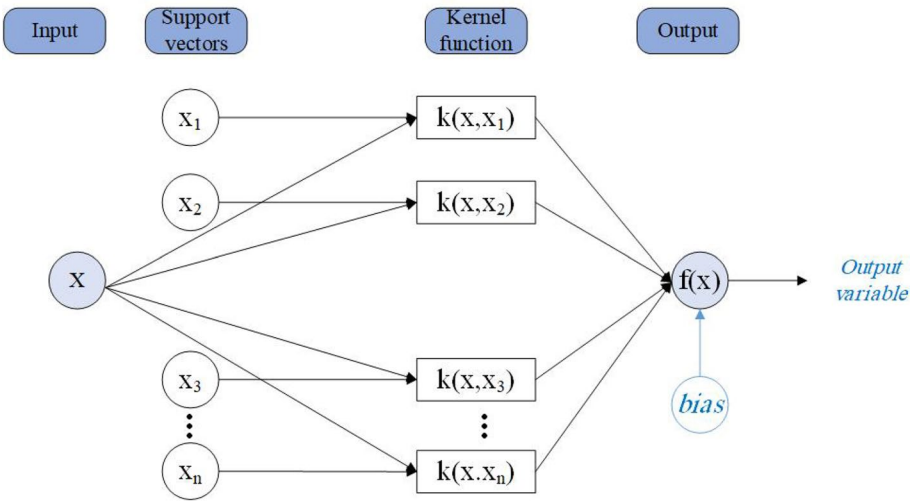

Fig. 4 SVR algorithm

of large-scale predictor variables derived from GCM outputs at time t, and let y_t represent the observed monthly precipitation at the corresponding station.

Base learners

Three machine learning models including MLP, SVR, and RF are employed as base learners. Each base learner $m \in \{1.2.3\}$ approximates a nonlinear mapping between predictors and observations:

$$\widehat{y}_t^{(m)} = f_m\left(X_t\right) \tag{2}$$

where f_m represents the functional form learned by the $m-th$ base model. These models were selected due to their proven effectiveness in statistical downscaling and hydroclimatic applications (Sachindra et al. 2018; Ahmed et al. 2019; Kalu et al. 2024).

Leakage-safe generation of meta-features

To prevent information leakage (Cawley and Talbot 2010), meta-features are generated exclusively using out-of-fold (OOF) predictions. The training dataset D_{train} is partitioned into K temporally ordered folds using an expanding-window cross-validation strategy, which is appropriate for time-dependent hydroclimatic data. For each fold k, base learners are trained on observations up to time t_{k-1} and generate predictions for the subsequent validation block:

$$\widehat{y}_t^{(m)} = f_m^{-k}\left(X_t\right).t \in \text{fold k} \tag{3}$$

where f_m^{-k} denotes the base learner trained without access to fold k.

The OOF predictions are concatenated to construct the meta-feature vector:

$$Z_t = \begin{bmatrix} \widehat{y}_t^{(1)} \\ \widehat{y}_t^{(2)} \\ \widehat{y}_t^{(3)} \end{bmatrix}.t \in D_{train} \tag{4}$$

At no stage are in-sample predictions used in the stacking layer, ensuring a leakage-safe implementation (Tyralis et al. 2019).

Meta-learner specification

A linear regression model is used as the meta-learner to combine base-model predictions:

$$\widehat{y}_t^{SG} = \beta_0 + \sum_{m=1}^{3} \beta_m \widehat{y}_t^{(m)} \tag{5}$$

where β_0 is the intercept and β_m are regression coefficients estimated using ordinary least squares. Linear regression was selected to ensure numerical stability and interpretability when combining correlated base-model outputs, and to minimize overfitting risk (Breiman 1996). Regularization was not applied, as sensitivity analyses did not yield performance improvements.

The meta-learner is trained only on OOF predictions derived from D_{train}. The testing dataset D_{test} is used exclusively for model evaluation and is not involved at any stage of

base-model training or meta-learner fitting, ensuring a temporally independent assessment of predictive performance consistent with real-world forecasting conditions.

3.3 Model evaluation

To compare different downscaling methods and GCMs, statistical metrics including the Nash–Sutcliffe Efficiency (NSE), Mean Squared Error (MSE), and Kling-Gupta Efficiency (KGE) were used. The formulas for the three metrics are shown below:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \tag{6}$$

where, O_i is observed value, P_i is predicted value, $\overline{O}$ is mean of the observed values, and n is number of observations.

If NSE = 1 indicates perfect performance with predictions matching observed values. If NSE > 0 suggests acceptable performance, meaning the model's predictions are better than simply using the mean of the observed data. However, an NSE ≤ 0 reflects poor performance, indicating that the model's predictions are no better, or even worse, than using the observed mean as a baseline (Nash and Sutcliffe 1970).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2 \tag{7}$$

where, O_i and P_i represent the observed and predicted values, respectively, and n is the total number of observations.

A lower Mean Squared Error (MSE) value indicates better model performance, as it reflects smaller average differences between observed and predicted values. Due to the squaring term in its calculation, MSE heavily penalizes larger errors, making it particularly sensitive to outliers in the data.

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \tag{8}$$

where, r is Linear correlation coefficient between observed and predicted values.

$$\beta = \frac{\overline{P}}{\overline{O}} \text{Bias ratio (mean of predicted to mean of observed values)} \tag{9}$$

$$\gamma = \frac{\sigma_P}{\sigma_O} \text{ Variability ratio (standard deviation of predicted to observed values)} \tag{10}$$

A Kling-Gupta Efficiency (KGE) value of 1 represents perfect agreement between observed and predicted data, with values closer to 1 indicating better model performance.

This metric is especially valuable for identifying specific sources of model error, such as bias, variability, or correlation issues, providing a more comprehensive evaluation of model accuracy (Patil and Stieglitz 2015).

3.4 Standardized precipitation index (SPI)

This method, developed by McKee et al. (1993), is designed to assess drought severity across different timescales. The SPI, which is a powerful, flexible and simple index (Habibi et al. 2024), and provides a standardized evaluation of precipitation anomalies over varying periods. The concept of this index is based on cumulative probability of the precipitation data as:

$$G(x) = \frac{1}{\beta_{pro}^{\alpha_{pro}}(\alpha_{pro})} \int_0^x x^{\alpha_{pro}-1} e^{-\frac{x}{\beta_{pro}}} dx \tag{11}$$

where,

$$\alpha_{pro} = \frac{1}{4A}\left(1 + \sqrt{1 + \frac{4A}{3}}\right) \tag{12}$$

$$A = \ln(x_{sr}) - \frac{\sum_{i=1}^{n} \ln(x_i)}{n} \tag{13}$$

$$\beta_{pro} = \frac{x_{sr}}{\alpha_{pro}} \tag{14}$$

And $G(x)$ is the cumulative Gamma distribution function, α_{pro} is the shape parameter, β_{pro} is the scale parameter, x_{sr} is the average precipitation, n is the precipitation measurement number, and x_i is the precipitation in the sequence of data.

If x=0, then $H(x) = q + (1-q)\,G(x)$ is used instead of previous cumulative probability $G(x)$.

Where, q is the probability of no precipitation.

The standard normal distribution, characterized by a mean of zero and a variance of one, was derived through the transformation of the cumulative probability. Subsequently, the values of the SPI were calculated using Eq. (15).

$$SPI = \begin{cases} -\left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\right) & 0 < H(x) \le 0.5 \\ +\left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\right) & 0.5 < H(x) \le 1 \end{cases} \tag{15}$$

And

$$t = \begin{cases} \sqrt{\ln \frac{1}{(H(x))^2}} & 0 < H(x) \le 0.5 \\ \sqrt{\ln \frac{1}{(1-H(x))^2}} & 0.5 < H(x) \le 1 \end{cases} \tag{16}$$

The cumulative probability of the observed precipitation is represented by $H(x)$. The constants c_0, c_1, c_2, d_1, d_2 and d_3 hold specific values: 2.515517, 0.802853, 0.010328, 1.432788, 0.189269, and 0.001308, respectively.

In this research, SPI values for 3-, 6-, and 12-month timescales were calculated using a parametric approach. The aim is to examine the variations in drought indicator trends across these different time scales. SPI-3 is useful for analyzing seasonal droughts, often affecting agriculture and water availability over a few months, while SPI-6 reflects medium-term drought patterns that can influence crop cycles and reservoir levels. SPI-12, on the other hand, represents long-term drought trends, impacting water resources, hydrological systems, and broader climatic conditions. By using SPI across these timescales, the study ensures a comprehensive assessment of drought characteristics, making the comparison between parametric and non-parametric methods robust across varying drought durations and severities. Detailed information about the calculation of this index is given in (Mirdarsoltany et al. 2025).

3.5 Run theory

Within the framework of run theory, drought characteristics are assessed using key indicators such as severity, duration, and intensity. Drought severity represents the cumulative total of SPI values that fall below a designated threshold. Duration is defined as the continuous time span during which the SPI remains under this threshold, and intensity is computed by dividing the severity by the duration (Mesbahzadeh et al. 2020; Mirdarsoltany et al. 2025). For this analysis, drought events are identified when SPI values drop below -1, which is a commonly accepted threshold (Azam et al. 2018). The duration of a drought corresponds to the number of consecutive time steps with SPI values under -1, as illustrated in Fig. 5 (if SPI<-1 is considered drought event). Severity is calculated by summing the SPI values within this period, and intensity is obtained by dividing that sum by the length

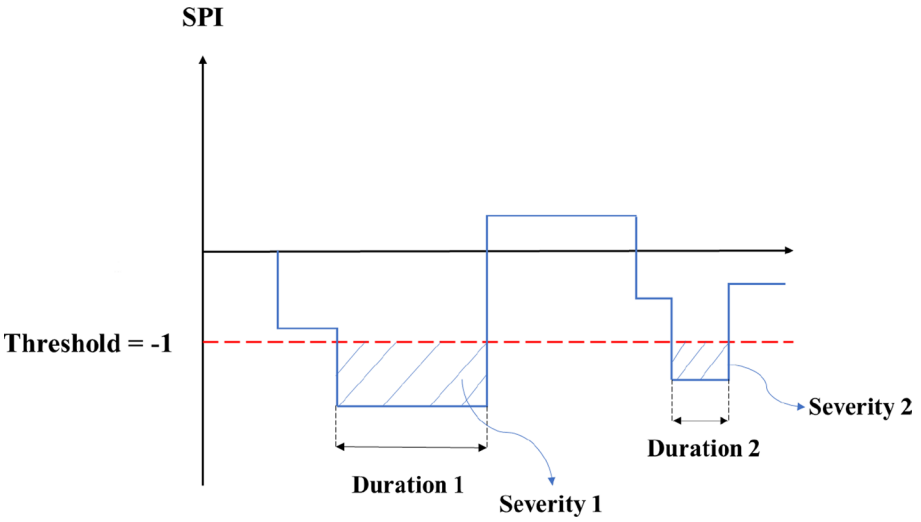

Fig. 5 Schematic illustration of the run theory

of the drought. This approach provides a systematic way to quantify both the strength and persistence of drought episodes.

4 Results and discussion

As mentioned before, initially, the input variables considered for the models included GCM-derived included surface air temperature, near-surface air temperature, geopotential height at multiple pressure levels (1000, 925, 850, 700, 600, 500, 400, and 300 hPa), relative and specific humidity at the surface and pressure levels, near surface wind speed, sea level pressure, surface air pressure. After calculating the Pearson correlation coefficient between each candidate predictor and the observed precipitation series, five predictors from five GCMs were selected for training ML models to perform downscaling. After a series of trial-and-error approach, it was determined that only precipitation, surface temperature, and surface air pressure positively influenced the training process, while relative humidity and sea level pressure had minimal or negative impacts (Table 3). Using these three key variables along with seasonality index, four different ML methods, MLP, SVR, RF and Stacked Generalization, were employed for downscaling monthly precipitation data.

Each model leveraged specific tuning features to enhance its performance. The best tuning features for each model were identified through a trial-and-error approach. Each model was run 1000 times, and the optimal features were selected based on the average performance metrics, including NSE, MSE, and KGE. It should be noted that hyper parameter tuning was used along with the above trial-and-error approach for MLP method. Since in trial-and-error approach you do not exhaustively test many combinations, there is less risk of overfitting the model to the validation data, and therefore the trial-and-error approach was selected for this study.

Additionally, the dataset was divided into three subsets for all ML methods: training, validation, and testing. In total, 80% of the data (1970–2014) was allocated to training and validation, which were randomly split into 70% for training and 10% for validation. The remaining 20% (2015–2023) was reserved for testing. This division was carefully examined to prevent overfitting.

The results presented in Table 4 show the performance comparison of the MLP, SVR, RF, and Stacked Generalization methods across multiple stations and different climate models (ACCESS-CM2, CanESM5, HadGEM3-GC31-LL, and MIROC6). Table 4 indicates that the Stacked method consistently outperforms the individual base methods (MLP, SVR, and RF), showing the highest NSE across all stations and climate models. This is further supported by the MSE results in Table 4, where the Stacked method exhibits the lowest Mean Squared Error (MSE) compared to the other ones. Table 4 shows that the Stacked method also performs the best in terms of KGE, highlighting its superior ability to capture both the correlation and variability of the data. Overall, this table confirms that the Stacked General-

Table 3 Predictors used for training ML methods

Predictor	Unit	Status
Precipitation (pr)	$kg\ m^{-2}\ s^{-1}$	Selected
Surface air pressure (ps)	Pa	Selected
Surface temperature (ts)	K	Selected
Sea level pressure (psl)	Pa	Not Selected
Relative humidity	%	Not Selected

Table 4 Results of statistical metrics for different GCMs and downscaling methods

Down-scaling method	Model	Evaluation metric	Station number							
			1	2	3	4	5	6	7	8
MLP	ACCESS-CM2	NSE	0.45	0.46	0.40	0.49	0.48	0.43	0.48	0.49
		MSE	2025.96	1458.51	2563.94	1804.88	1657.56	2531.72	2050.54	965.58
		KGE	0.49	0.54	0.47	0.53	0.52	0.49	0.53	0.57
	CanESM5	NSE	0.59	0.55	0.55	0.56	0.59	0.57	0.58	0.59
		MSE	1405.4	1282.38	2095.52	1620.91	1405.4	2042.62	1753.69	832.55
		KGE	0.63	0.64	0.60	0.63	0.63	0.62	0.63	0.65
	Had-GEM3-GC31-LL	NSE	0.52	0.52	0.46	0.54	0.52	0.49	0.55	0.51
		MSE	1839.84	1342.42	2387.02	1687.24	1575.52	2322.27	1846.08	938.42
		KGE	0.55	0.59	0.52	0.59	0.58	0.56	0.60	0.57
	MIROC6	NSE	0.55	0.54	0.47	0.54	0.53	0.50	0.52	0.56
		MSE	1766.24	1297.28	2349.68	1665.82	1552.87	2280.87	1931.86	877.29
		KGE	0.61	0.65	0.60	0.66	0.63	0.60	0.64	0.66
SVR	ACCESS-CM2	NSE	0.30	0.36	0.37	0.39	0.35	0.33	0.38	0.39
		MSE	2451.18	1658.45	2668.55	2062.18	1971.7	2865.1	2317.41	1100.8
		KGE	0.32	0.38	0.32	0.39	0.35	0.32	0.37	0.50
	CanESM5	NSE	0.30	0.42	0.38	0.4	0.39	0.34	0.37	0.39
		MSE	2446.98	1552.03	2621.11	2013.71	1867.02	2817.04	2343.82	1100.15
		KGE	0.32	0.51	0.34	0.5	0.39	0.33	0.36	0.39
	Had-GEM3-GC31-LL	NSE	0.34	0.39	0.35	0.40	0.4	0.33	0.40	0.42
		MSE	2326.8	1597.69	2726.57	2020.78	1849.77	2845.83	2274.73	1069.4
		KGE	0.35	0.41	0.30	0.40	0.39	0.33	0.38	0.42
	MIROC6	NSE	0.44	0.45	0.41	0.44	0.44	0.37	0.45	0.48
		MSE	1752.56	1490.61	2531.63	1914.05	1752.56	2713.8	2123.42	985.2
		KGE	0.42	0.43	0.36	0.42	0.42	0.36	0.42	0.45
RF	ACCESS-CM2	NSE	0.43	0.36	0.33	0.55	0.46	0.44	0.43	0.39
		MSE	2100.38	1672.73	2779.36	1653.85	1700.28	2484.01	2192.72	1111.93
		KGE	0.57	0.56	0.55	0.57	0.59	0.60	0.59	0.60
	CanESM5	NSE	0.55	0.55	0.49	0.59	0.54	0.57	0.45	0.48
		MSE	1753.61	1267.8	2280.66	1532.62	1522.21	2068.39	2113.29	984.06
		KGE	0.62	0.70	0.68	0.75	0.63	0.67	0.61	0.63
	Had-GEM3-GC31-LL	NSE	0.45	0.38	0.48	0.5	0.51	0.46	0.47	0.50
		MSE	2039.11	1623.48	2302.19	1782.84	1592.61	2418.06	2058.2	959.37
		KGE	0.58	0.57	0.53	0.64	0.59	0.58	0.60	0.63
	MIROC6	NSE	0.55	0.43	0.48	0.52	0.55	0.55	0.57	0.55
		MSE	1504.84	1522.87	2314.8	1716.82	1504.84	2107.93	1780.41	883.89
		KGE	0.64	0.62	0.67	0.68	0.64	0.67	0.67	0.65

Table 4 (continued)

Down-scal-ing method	Model	Eval-uation metric	Station number							
			1	2	3	4	5	6	7	8
SG	ACCESS-CM2	NSE	0.52	0.49	0.44	0.53	0.54	0.50	0.51	0.47
		MSE	1852.96	1413.97	2450.94	1703.53	1534.34	2282.43	1948.47	990.88
		KGE	0.58	0.54	0.48	0.60	0.61	0.56	0.58	0.53
	CanESM5	NSE	0.52	0.53	0.47	0.53	0.52	0.52	0.53	0.55
		MSE	1836.96	1321.1	2352.81	1693.07	1562.28	2220.37	1905.02	891.04
		KGE	0.59	0.60	0.52	0.60	0.59	0.59	0.60	0.62
	Had-GEM3-GC31-LL	NSE	0.56	0.53	0.44	0.53	0.54	0.53	0.54	0.51
		MSE	1744.73	1327.39	2448.97	1686.11	1513.68	2190.02	1866.4	945.46
		KGE	0.63	0.60	0.48	0.60	0.62	0.60	0.61	0.57
	MIROC6	NSE	0.61	0.58	0.58	0.62	0.61	0.60	0.62	0.60
		MSE	1352.5	1229.51	1994.03	1462	1352.5	1966.56	1649.31	815.37
		KGE	0.70	0.66	0.66	0.70	0.70	0.68	0.70	0.68

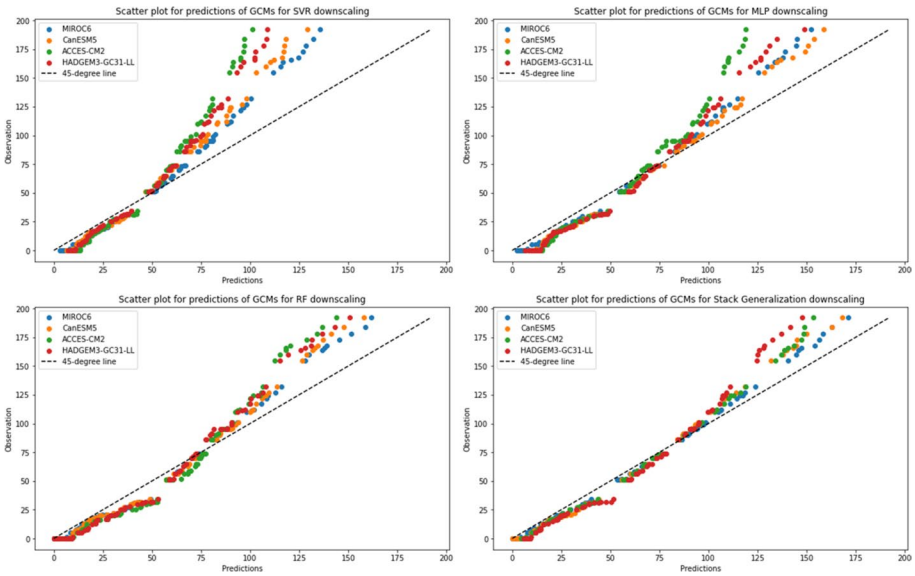

Fig. 6 Scatter plot for predicted values based on different GCMs and MLs and observation data for Afjeh station and according to test data

ization method is the most effective one, providing more accurate and reliable precipitation predictions than the individual MLP, SVR, and RF methods across different stations and climate models.

Figure 6 presents the scatter plot for the Afjeh station, comparing the actual and predicted precipitation values across four different models and four different downscaling methods. It can be seen that predicted values derived by stacked generalization model show the best alignment with 45-degree line, indicating that this model provides the most accurate predic-

tions for Afjeh station. In addition, predictions from all downscaling methods indicate that, except for the MLP model, where CanESM5 performed best, MIROC6 demonstrated the highest performance among the other GCMs when using the SVR, RF, and SG downscaling methods.

Figure 7 presents Taylor diagrams comparing the performance of Stacked generalization model for the sake of downscaling. The results show that Stack Generalization (SG) outperformed other models, exhibiting a higher correlation, and the standardized deviation of the simulated GCMs was closer to that of the observed data.

The superior performance of the Stacked Generalization (SG) approach compared to the individual base learners (MLP, SVR, and RF) can be attributed to its ability to effectively exploit the complementary strengths of heterogeneous models. While individual machine learning algorithms tend to capture specific aspects of the nonlinear relationship between large-scale climate predictors and local precipitation, each model is also subject to distinct structural biases and limitations. The SG framework mitigates these weaknesses by combining the predictions of multiple base learners through a secondary learning stage, allowing

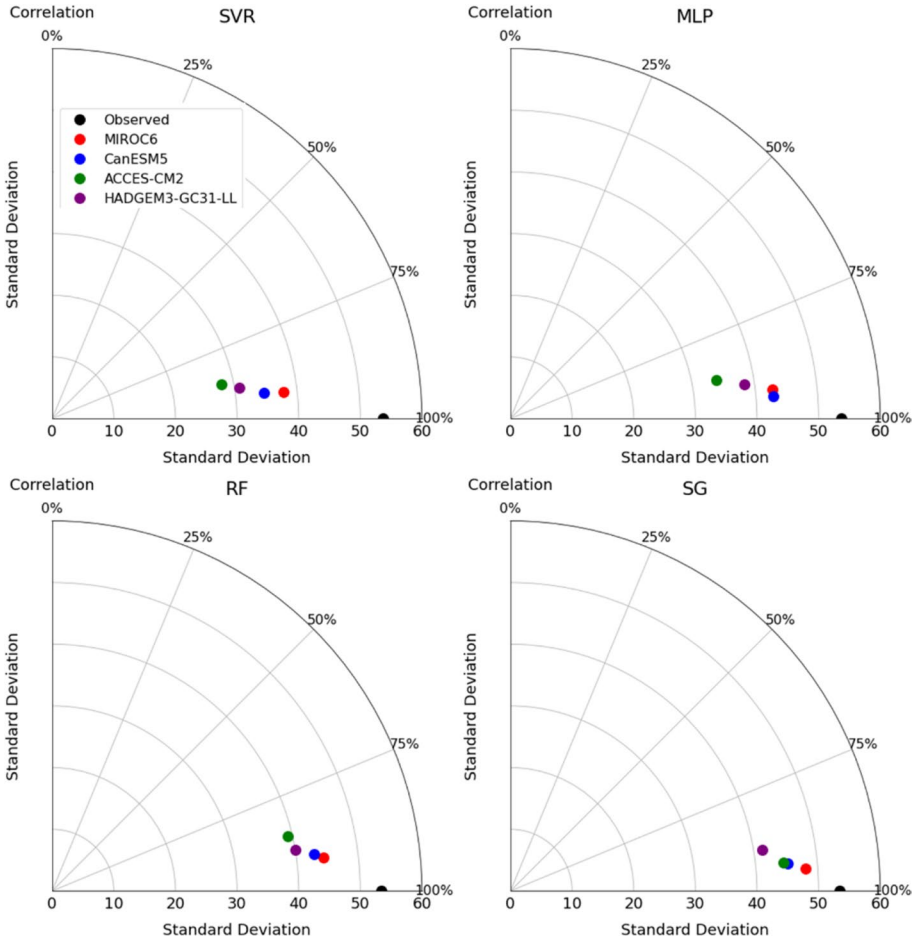

Fig. 7 Taylor diagrams for estimated precipitation by different MLs for testing data

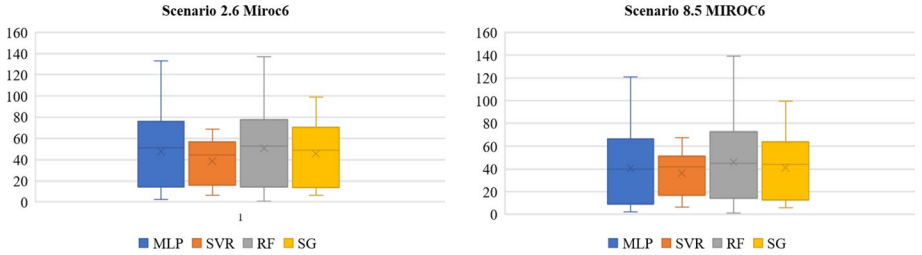

Fig. 8 Comparison of downscaling methods at Afjeh Station under scenarios 2.6 and 8.5

Table 5 Percentage of difference in average precipitation in comparison with historical average precipitation for Afjeh station

Model MIROC6	Scenario 2.6			Scenario 8.5		
	Near	Mid	Far	Near	Mid	Far
	− 16%	− 18%	− 20%	− 20%	− 23%	−32%

the meta-learner to assign optimal weights to each model based on their predictive skill. In addition, the use of out-of-fold predictions in the stacking process reduces the risk of overfitting and improves generalization performance. This explains the consistently higher NSE and KGE values and lower MSE achieved by the SG model across different stations and GCMs. Similar improvements using stacking-based ensemble approaches have been reported in previous hydrological and climate downscaling studies, confirming the robustness of this framework for precipitation modeling.

Additionally, the results demonstrate that MIROC6 was the best GCM across most downscaling methods. However, again an exception was found with the MLP model, where CanESM5 performed better, indicating that different GCMs may lead to varying results depending on the ML technique used. This underscores the importance of selecting the appropriate model for each specific case.

Figure 8 compares the performance of different machine learning-based downscaling methods for monthly precipitation data at the Afjeh station under two climate scenarios, scenario 2.6 and scenario 8.5 based on MIROC6 model. Across both scenarios, the MLP and RF methods show the highest variability, suggesting it captures a wider range of precipitation extremes. The SVR method consistently has the smallest interquartile range (IQR) and lower variability, indicating a more conservative or stable estimate of precipitation.

Overall, the results suggest that the choice of machine learning method significantly influences the range and variability of downscaled precipitation, with scenario 2.6 showing generally higher medians compared to scenario 8.5. As this study used SPI which uses a moving average to measure drought characteristics, variability of predicted values, leave prominent impacts on the results, and therefore selection of the best GCM and downscaling method can help policy-makers to take proper measures to mitigate drought consequences.

For a more thorough analysis of each method, the future period was divided into three distinct categories: near future (2030–2055), mid future (2055–2080), and far future (2080–2100) (Table 5). With respect to the derived precipitation values according to scenario 2.5 and 8.5 and considering MIROC6 results, the average precipitation at Afjeh station is projected to decrease in the near, mid, and far future, respectively. Similar reductions are

observed in other stations as well, though only the results for Afjeh station are presented here for clarity. According to the Table 5, for scenario 8.5, the amount of average precipitation will decrease in far future by 32%.

Non-exceedance probabilities for three precipitation thresholds (20 mm, 50 mm, and 70 mm) were analyzed using historical data and projected future scenarios (2.6 and 8.5). These thresholds correspond to the 25th, 50th, and 75th quantiles, respectively, and were derived from the average results of all climate models and downscaling methods. For the near future, the non-exceedance probability for 20 mm precipitation decreased compared to historical data, with about 30% of the data falling below this threshold for scenario 2.6. These decreases are obvious for the non-exceedance probabilities for 50 mm precipitation as well. However, for 70 mm precipitation, the non-exceedance probability, increased from 65% historically to 70% and 75% for scenarios 2.6 and 8.5, respectively. In the mid future, the non-exceedance probability for 20 mm precipitation continued to decline, and for 50 mm precipitation, it dropped to 52%. Conversely, scenario 8.5 saw an increase in non-exceedance probability for 50 mm precipitation to around 60%. For 70 mm precipitation, both future scenarios exhibited an increase in non-exceedance probabilities, reaching 80% and 85%, respectively. In the far future, non-exceedance probabilities for 20 mm precipitation remained relatively unchanged. However, for 50 mm precipitation, scenario 2.6 showed a decrease in non-exceedance probability compared to historical data. In contrast, for scenario 8.5, the probability for 50 mm precipitation increased from 50% to approximately 70%, and for 70 mm precipitation, the probabilities rose to 70% and 90% in scenarios 2.6 and 8.5, respectively. This figure illustrates that, under both climate scenarios (2.6 and 8.5), the frequency of precipitation events below 70 mm (representing the 75th percentile) is projected to increase in the near, mid, and far future. This trend can leave impacts on duration, severity and intensity of drought events which will be discussed in this study (Fig. 9).

The SPI was calculated using a parametric approach across three temporal scales: 3, 6, and 12 months. This analysis incorporated both observed datasets and outputs from GCMs for the historical period, along with projections under two Shared Socio-economic Pathways (SSP): 2.6 and 8.5. These scenarios, drawn from the IPCC Sixth Assessment Report, represent contrasting climate futures. Scenario 2.6 reflects a low-emission trajectory aimed at limiting global temperature rise to below 2 °C, highlighting the potential outcomes of effective mitigation strategies. Conversely, scenario 8.5 illustrates a high-emission scenario, indicative of limited mitigation efforts and more severe climate impacts.

The dual inclusion of scenarios 2.6 and 8.5 enables a broad examination of potential future conditions, supporting comprehensive risk assessments and policy planning for climate adaptation and mitigation.

To analyze drought variability, the SPI was calculated at 3-, 6-, and 12-month intervals. These different time scales offer insight into varying drought dynamics: SPI-3 is suitable for identifying short-term, seasonal droughts with implications for agriculture and water supply; SPI-6 captures intermediate drought trends affecting crop growth cycles and reservoir storage; and SPI-12 reflects longer-term hydrological droughts with impacts on water systems and regional climate patterns. Through this multi-scale approach, the study provides an in-depth evaluation of drought characteristics in the study area.

Figure 10 displays the time series of the SPI for three different time scales: SPI-3, SPI-6, and SPI-12. Above the threshold (− 1), SPI values represented by blue bars, indicate wet conditions, while SPI values below -1.0 (shown in red) correspond to Moderate Dryness and

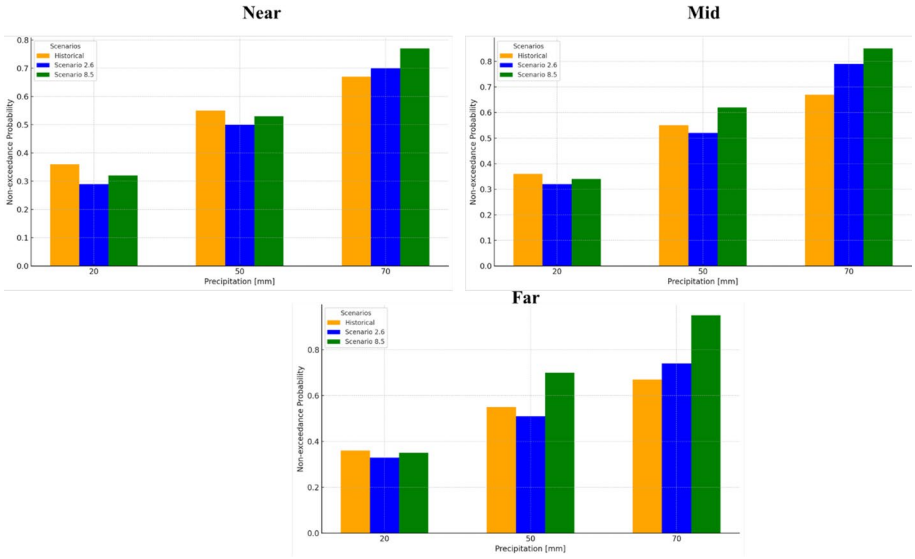

Fig. 9 The non-exceedance probability for 20 mm, 50 mm, and 70 mm precipitation levels under historical data and future climate scenarios (2.6 and 8.5)

Severe Dryness conditions, signaling drought events. Specifically, droughts are considered to occur when SPI values fall below a threshold of -1.0. The figure effectively illustrates how drought intensity and duration vary across different time scales, providing a clear picture of the evolving severity of drought conditions over time. This visualization is key to understanding how droughts develop and how their severity changes depending on the time scale analyzed. It can be seen that the shorter timescale is, the more drought events can be captured. However, drought events are more severe in longer timescales.

The results in Fig. 11 present the drought characteristics for the historical period across different stations based on SPI-3, SPI-6, and SPI-12 indices. The duration analysis shows that SPI-12 exhibits longer drought durations compared to SPI-3 and SPI-6, especially at Latyan, where it reaches its maximum with a duration of about 8 months. In terms of intensity, SPI-3 shows the most intense drought conditions across locations, with values ranging from about − 1.2 to − 1.8, followed by SPI-6 and SPI-12. The severity results further reinforce the impact of longer timescales (SPI-12), with Latyan experiencing the most severe drought conditions among all locations, reaching a severity value of about − 18. The observed differences highlight the influence of timescales on drought characteristics, emphasizing the importance of SPI-12 for capturing prolonged and severe drought events.

Under scenario 2.6, the drought characteristics derived from the Stack Model reveal a clear pattern of increasing drought severity and duration with longer timescales, particularly as represented by the SPI-12 index. In the near future, SPI-12 consistently shows the longest drought durations across all stations, with a peak of approximately 8 months observed in Afjeh. In contrast, SPI-3 and SPI-6 indicate shorter drought durations of about 2 to 3 months. Although the intensity values across all indices remain relatively stable, SPI-3 tends to exhibit the most intense drought events, with values ranging from -1.4 to -1.5 across stations. However, in terms of drought severity, SPI-12 presents the most criti-

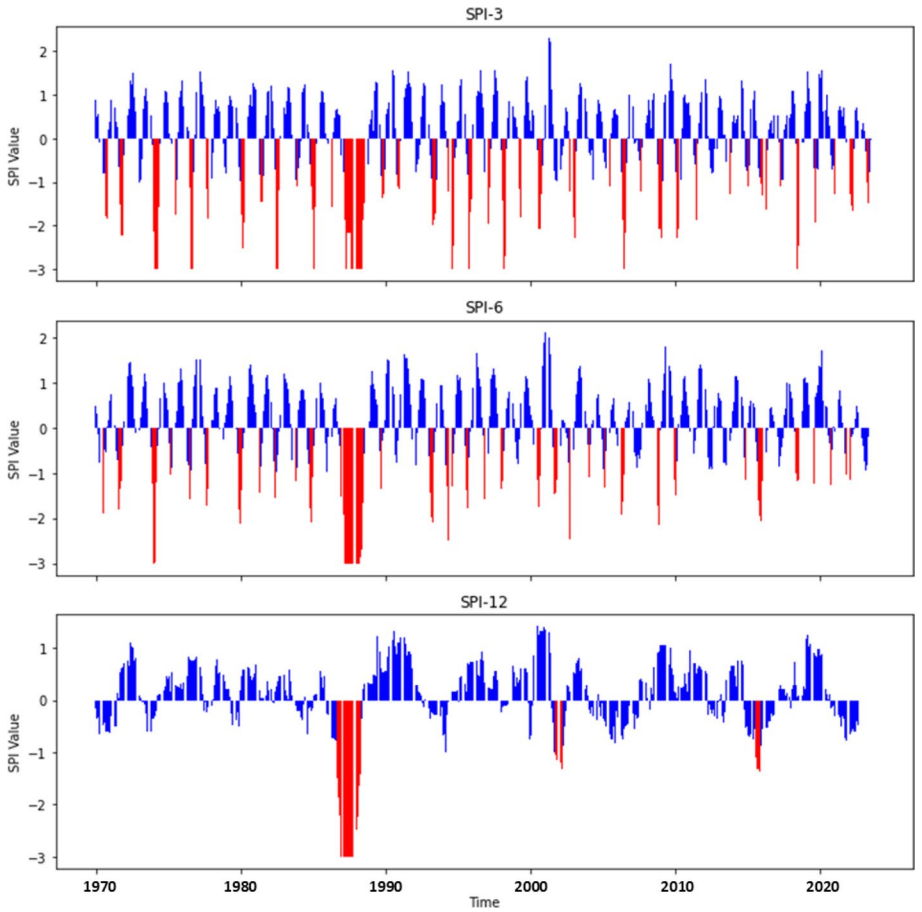

Fig. 10 Time series of the SPI at 3, 6 and 12 timescales for historical precipitation, showing drought (red) and wet (blue) conditions over time for Afjeh station

cal conditions, reaching values close to -12, especially in Afjeh, emphasizing the impact of prolonged drought episodes. Moving into the mid future, slight reductions in SPI-12 drought durations are observed, decreasing to around 5–6 months across most stations, while intensity levels remain consistent with those of the near future. The severity values also show a moderate decline, with maximum values dropping to around -8 in stations like Kandsofla and Lavasanbozorg. In the far future, the duration of SPI-12 droughts increases again, reaching nearly 8 months in areas such as Kandsofla, and severity values return to previous high levels close to -12. SPI-3 and SPI-6 durations remain shorter throughout all periods, typically around 2 months, while intensity values for these indices remain stable and relatively unchanged. These findings suggest that although short-term drought characteristics remain consistent, long-term droughts, as captured by SPI-12, are expected to become more prolonged and severe under scenario 2.6, especially in specific regions. For comparison, results from scenario 8.5 indicate that future droughts could become significantly more severe and longer than those under scenario 2.6. However, drought intensity

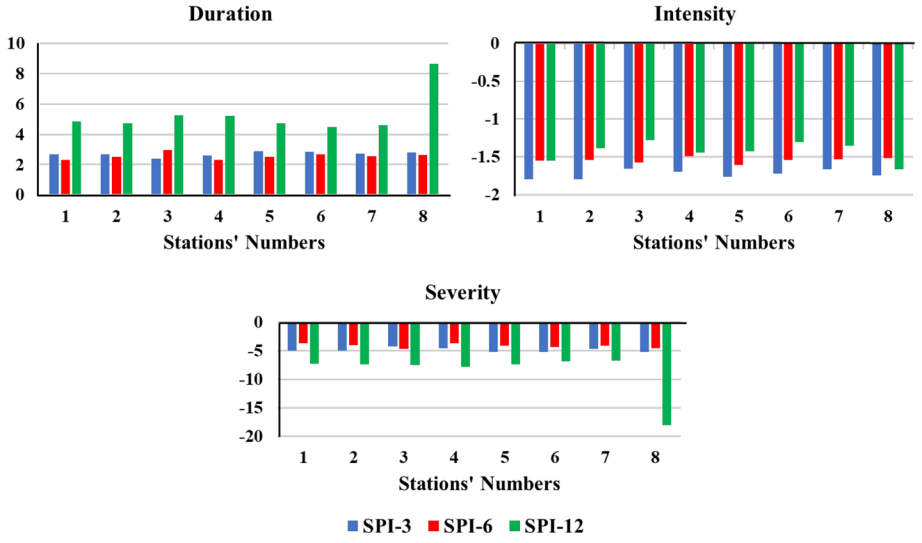

Fig. 11 Drought Characteristics for historical period

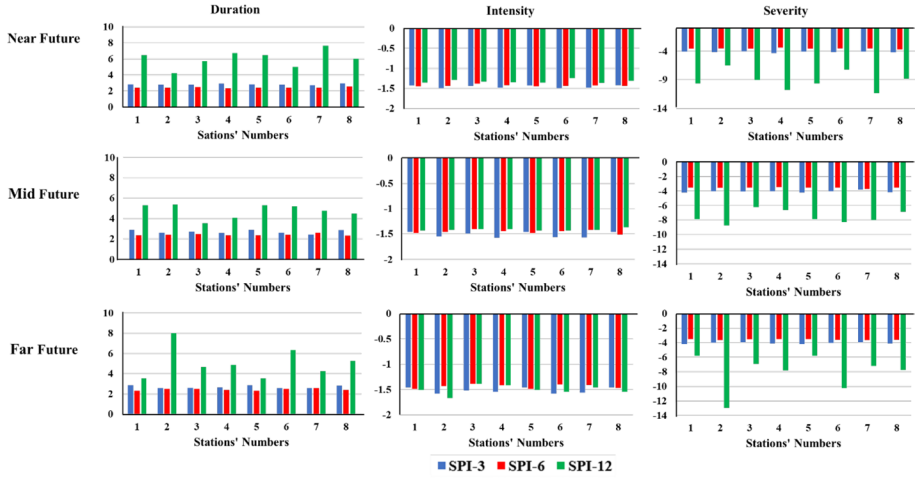

Fig. 12 Drought Characteristics for scenario 2.6 derived by Stack method and MIROC6

remains relatively unchanged between scenarios, as it is defined by the ratio of severity to duration, indicating that the more extreme drought conditions projected under scenario 8.5 result from a combined increase in both duration and severity. Nonetheless, the findings under scenario 2.6 still highlight a clear risk of intensifying drought impacts in the coming decades, particularly for long-term drought events. (Figs. 12 and 13)

The analysis of drought characteristics across different time scales (SPI-3, SPI-6, and SPI-12) and future periods reveals varying model performances. For SPI-3, HadGEM3-GC31-LL predicts the longest drought durations in the near and mid future, while CanESM5 takes the lead in the far future, indicating a shift in model dominance over time. MIROC6 con-

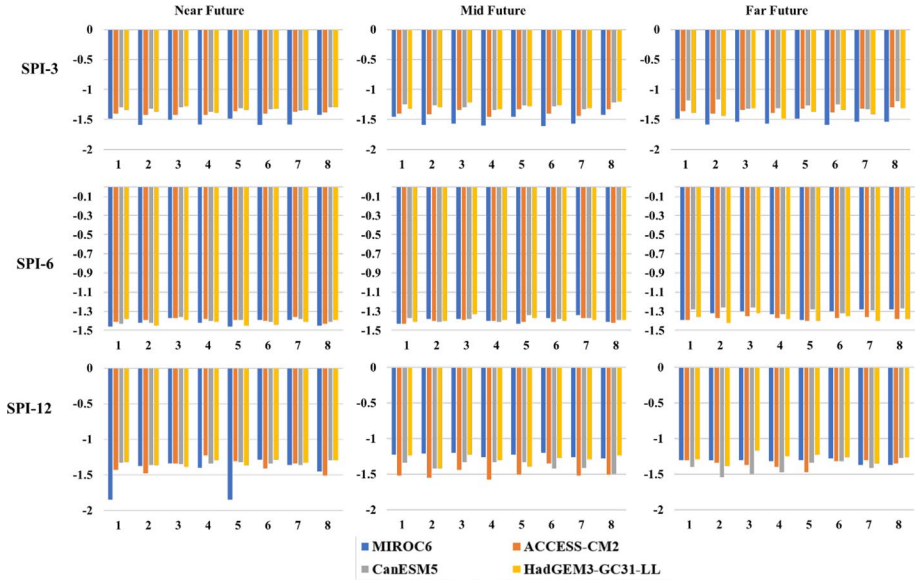

Fig. 13 The results of intensity based on SG method

sistently projects the most intense droughts across all periods for SPI-3, while for SPI-6, it leads in the near and mid future but is surpassed by CanESM5 in the far future. SPI-6 results also show high spatial variability, with no single model consistently outperforming Others in terms of drought duration. For SPI-12, ACCESS-CM2 predicts the longest and most severe droughts in the near future, with mixed results in the mid future, and HadGEM3-GC31-LL dominating in the Far Future in both duration and severity. Overall, the severity and duration of projected droughts vary significantly by model, time period, and SPI scale, highlighting the complexity and spatial dependence of future drought scenarios. The key takeaway is that no single model universally outperforms others across all conditions, underscoring the importance of using a multi-model ensemble approach to capture the range of possible future drought behaviors and to inform robust water resource planning.

5 Conclusion

The Latyan watershed, one of Iran's most significant basins, was selected to enhance the precision of statistical downscaling results. The primary objective was to use the stacked generalization and ML techniques to improve the performance of traditional statistical downscaling methods and assess their impact on drought characteristics. Initially, five predictors from five GCMs, along with a seasonal index, were selected to train ML models for downscaling. Following a series of experiments, it was determined that only precipitation, surface temperature, and surface air pressure positively influenced the training process. Using these three key variables in conjunction with the seasonal index, four different ML methods namely MLP, SVR, RF, and Stacked Generalization were employed to downscale monthly precipitation data. The results indicated that the Stacked method consistently out-

performed the individual base ones (MLP, SVR, and RF), achieving the highest Nash–Sutcliffe Efficiency (NSE) across all stations and climate models, and exhibiting the lowest Mean Squared Error (MSE) compared to the other models.

Furthermore, the findings revealed that MIROC6 was the most effective GCM across most downscaling methods. The performance of various ML-based downscaling methods for monthly precipitation data was compared under two climate scenarios: 2.6 (optimistic) and 8.5 (pessimistic) scenarios, based on the MIROC6 model. The results showed that the choice of ML model significantly influenced the magnitude and variability of downscaled precipitation, with the S2.6 scenario showing higher mean averages compared to 8.5 scenario.

For a more detailed analysis, the future period was divided into three distinct categories: near future (2030–2055), mid-century (2055–2080), and late century (2080–2100). According to the precipitation values derived from scenarios 2.6 and 8.5, and based on MIROC6 results, it is projected that the average precipitation at the Afjeh station will decrease in the near, mid, and late future periods, respectively.

Subsequently, the SPI was calculated using a parametric method at 3-, 6-, and 12-month timescales. This analysis utilized observed data alongside GCMs for the historical period and two future scenarios (2.6 and 8.5). Various timescales, particularly 3-, 6-, and 12-month intervals, were employed to calculate drought indices in the study area to examine how drought index trends change across these timescales. The findings indicated that while short-term drought characteristics remain stable, long-term droughts, as recorded by SPI-12, are expected to become longer and more severe under scenario 2.6, especially in specific regions. In comparison, the results of scenario 8.5 suggest that future droughts could be significantly more severe and prolonged than those under 2.6 scenario. However, drought intensity remains relatively unchanged between scenarios, as it is defined by the ratio of severity to duration, indicating that the more severe drought conditions projected under 8.5 scenario result from a combined increase in duration and severity. Nonetheless, the findings about 2.6 scenario still highlight the apparent risk of intensified drought effects in the coming decades, particularly for long-term drought events.

Author contributions Amirhossein Mirdarsoltani: Applying the methodology, preparing codes, writing the original draft, and editing. Matineh Imani Borhan: Applying the methodology, analyzing the data, writing the original draft, and editing. Leila Rahimi: Methodology, supervision, validating the results, review, and edit the draft. Carlo De Michele: Review and edit the draft.

Funding No funding was used in this research.

Declarations

Conflict of interests The authors declare no competing interests.

Ethical approval All authors have read, understood, and complied as applicable with the statement on "Ethical responsibilities of Authors" as found in the Instructions for Authors.

References

Agatonovic-Kustrin S, Beresford R (2000) Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharm Biomed Anal 22:717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

Ahmed K, Shahid S, Haroon SB, W X-J (2015) Multilayer perceptron neural network for downscaling rainfall in arid region: a case study of Baluchistan, Pakistan. J Earth Syst Sci 124:1325–1341

Ahmed K, Sachindra DA, Shahid S et al (2019) Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. Hydrol Earth Syst Sci 23:4803–4824

Anandhi A, Srinivas VV, Nanjundiah RS, Nagesh Kumar D (2008) Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. Int J Climatol 28:401–420

Azam M, Maeng SJ, Kim HS, Murtazaev A (2018) Copula-based stochastic simulation for regional drought risk assessment in South Korea. Water 10:359

Baghanam AH, Eslahi M, Sheikhbabaei A, Seifi AJ (2020) Assessing the impact of climate change over the northwest of Iran: an overview of statistical downscaling methods. Theor Appl Climatol 141:1135–1150. https://doi.org/10.1007/s00704-020-03271-8

Banfi F, Cammalleri C, De Michele C (2024) A joint spatio-temporal characterization of the major meteorological droughts in Europe. Environ Res Lett 19:94041

Behfar N, Sharghi E, Nourani V, Booij MJ (2024) Drought index downscaling using AI-based ensemble technique and satellite data. Theor Appl Climatol 155:2379–2397

Bhardwaj K, Mishra A, Khedun CP (2025) Global droughts in a warming climate: Evaluation of SPI and SPEI under 1.5°, 2°, and 3 °C global warming. J Hydrol 659:133309. https://doi.org/10.1016/j.jhydrol.2025.133309

Breiman L (1996) Stacked regressions. Mach Learn 24:49–64. https://doi.org/10.1007/BF00117832

Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079–2107

Chen S-T, Yu P-S, Tang Y-H (2010) Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. J Hydrol 385:13–22

Coulibaly P (2004) Downscaling daily extreme temperatures with genetic programming. Geophys Res Lett 31:

Dai A (2011) Drought under global warming: a review. Wires Clim Change 2:45–65. https://doi.org/10.1002/wcc.81

Dariane AB, Borhan MI (2024) Comparison of classical and machine learning methods in estimation of missing streamflow data. Water Resour Manag. https://doi.org/10.1007/s11269-023-03730-7

Dariane AB, Borhan MI (2025) Comparative assessment of classical and machine learning approaches for rainfall data restoration. Environ Earth Sci 84:272. https://doi.org/10.1007/s12665-025-12255-8

Duhan D, Pandey A (2015) Statistical downscaling of temperature using three techniques in the Tons River basin in Central India. Theor Appl Climatol 121:605–622

Fiorillo F, Guadagno FM (2010) Karst spring discharges analysis in relation to drought periods, using the SPI. Water Resour Manag 24:1867–1884. https://doi.org/10.1007/s11269-009-9528-9

George J, Athira P (2024) A model output statistic-based probabilistic approach for statistical downscaling of temperature. Theor Appl Climatol 155:3871–3890

Ghosh S, Mujumdar PP (2008) Statistical downscaling of GCM simulations to streamflow using relevance vector machine. Adv Water Resour 31:132–146. https://doi.org/10.1016/j.advwatres.2007.07.005

Guo H, Bao A, Chen T et al (2021) Assessment of CMIP6 in simulating precipitation over arid Central Asia. Atmos Res 252:105451. https://doi.org/10.1016/j.atmosres.2021.105451

Guria R, Dwivedi S, Nayak P et al (2025) A comprehensive 120-year assessment of drought dynamics and climate teleconnections in Odisha, India (1901–2020): insights from SPI and trend evaluation. Nat Hazards 121:13811–13845. https://doi.org/10.1007/s11069-025-07336-7

Habibi B, Meddi M, Emre T et al (2024) Drought assessment and characterization using SPI, EDI and DEPI indices in northern Algeria. Nat Hazards 120:5201–5231. https://doi.org/10.1007/s11069-024-06408-4

He Q, Wang M, Liu K et al (2023) Spatiotemporal analysis of meteorological drought across China based on the high-spatial-resolution multiscale SPI generated by machine learning. Weather Clim Extrem 40:100567. https://doi.org/10.1016/j.wace.2023.100567

Hertig E, Jacobeit J (2008) Downscaling future climate change: temperature scenarios for the Mediterranean area. Glob Planet Change 63:127–131. https://doi.org/10.1016/j.gloplacha.2007.09.003

Huang YF, Ang JT, Tiong YJ et al (2016) Drought forecasting using SPI and EDI under RCP-8.5 climate change scenarios for Langat River Basin, Malaysia. Procedia Eng 154:710–717. https://doi.org/10.1016/j.proeng.2016.07.573

Hutengs C, Vohland M (2016) Downscaling land surface temperatures at regional scales with random forest regression. Remote Sens Environ 178:127–141. https://doi.org/10.1016/j.rse.2016.03.006
Kalu I, Ndehedehe CE, Ferreira VG, Kennard MJ (2024) Machine learning assessment of hydrological model performance under localized water storage changes through downscaling. J Hydrol 628:130597. https://doi.org/10.1016/j.jhydrol.2023.130597
Kazemzadeh M, Malekian A (2016) Spatial characteristics and temporal trends of meteorological and hydrological droughts in northwestern Iran. Nat Hazards 80:191–210
Kazemzadeh M, Noori Z, Alipour H et al (2022) Detecting drought events over Iran during 1983–2017 using satellite and ground-based precipitation observations. Atmos Res 269:106052. https://doi.org/10.1016/j.atmosres.2022.106052
Kim T-W, Valdés JB (2003) Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. J Hydrol Eng 8:319–328
Kumar YP, Maheswaran R, Agarwal A, Sivakumar B (2021) Intercomparison of downscaling methods for daily precipitation with emphasis on wavelet-based hybrid models. J Hydrol 599:126373
Laddimath RS, Patil NS (2019) Artificial neural network technique for statistical downscaling of global climate model. Mapan 34:121–127
Laimighofer J, Laaha G (2022) How standard are standardized drought indices? Uncertainty components for the SPI & SPEI case. J Hydrol 613:128385. https://doi.org/10.1016/j.jhydrol.2022.128385
Lorenzo MN, Pereira H, Alvarez I, Dias JM (2024) Standardized precipitation index (SPI) evolution over the Iberian Peninsula during the 21st century. Atmos Res 297:107132. https://doi.org/10.1016/j.atmosres.2023.107132
McKee TB, Doesken NJ, Kleist J (1993) The relationship of drought frequency and duration to time scales. In: Proceedings of the 8th Conference on Applied Climatology. California, pp 179–183
Mekanik F, Imteaz MA, Gato-Trinidad S, Elmahdi A (2013) Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. J Hydrol 503:11–21. https://doi.org/10.1016/j.jhydrol.2013.08.035
Mesbahzadeh T, Mirakbari M, Mohseni Saravi M et al (2020) Meteorological drought analysis using copula theory and drought indicators under climate change scenarios (RCP). Meteorol Appl 27:e1856. https://doi.org/10.1002/met.1856
Mirdarsoltany A, Dariane AB, Borhan MI (2025) Comprehensive GIS-driven evaluation of drought severity and duration: comparative assessment of parametric and non-parametric SPI methodologies. Theor Appl Climatol 156:225. https://doi.org/10.1007/s00704-025-05448-5
Nabaei S, Sharafati A, Yaseen ZM, Shahid S (2019) Copula based assessment of meteorological drought characteristics: regional investigation of Iran. Agric for Meteorol 276–277:107611. https://doi.org/10.1016/j.agrformet.2019.06.010
Naresh Kumar M, Murthy CS, Sesha Sai MVR, Roy PS (2009) On the use of Standardized Precipitation Index (SPI) for drought intensity assessment. Meteorol Appl A J Forecast Pract Appl Train Tech Model 16:381–389
Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I — a discussion of principles. J Hydrol 10:282–290. https://doi.org/10.1016/0022-1694(70)90255-6
Nishant N, Hobeichi S, Sherwood S et al (2023) Comparison of a novel machine learning approach with dynamical downscaling for Australian precipitation. Environ Res Lett 18:94006
Nourani V, Razzaghzadeh Z, Baghanam AH, Molajou A (2019) ANN-based statistical downscaling of climatic parameters using decision tree predictor screening method. Theor Appl Climatol 137:1729–1746
Okkan U, Inan G (2015) Bayesian learning and relevance vector machines approach for downscaling of monthly precipitation. J Hydrol Eng 20:4014051
Patil SD, Stieglitz M (2015) Comparing spatial and temporal transferability of hydrological model parameters. J Hydrol 525:409–417. https://doi.org/10.1016/j.jhydrol.2015.04.003
Pour SH, Shahid S, Sammen SS (2023) Chapter 25 - Runoff modeling using group method of data handling and gene expression programming. In: Eslamian S, Eslamian FBT-H of H (eds). Elsevier, pp 353–377
Sachindra DA, Ng AWM, Muthukumaran S, Perera BJC (2016) Impact of climate change on urban heat island effect and extreme temperatures: a case-study. Q J R Meteorol Soc 142:172–186
Sachindra DA, Ahmed K, Rashid MM et al (2018) Statistical downscaling of precipitation using machine learning techniques. Atmos Res 212:240–258. https://doi.org/10.1016/j.atmosres.2018.05.022
Saemian P, Tourian MJ, AghaKouchak A et al (2022) How much water did Iran lose over the last two decades? J Hydrol Reg Stud 41:101095. https://doi.org/10.1016/j.ejrh.2022.101095
Shayeghi A, Ziveh AR, Bakhtar A et al (2024) Assessing drought impacts on groundwater and agriculture in Iran using high-resolution precipitation and evapotranspiration products. J Hydrol 631:130828. https://doi.org/10.1016/j.jhydrol.2024.130828
Shirvani A, Landman WA (2022) Probabilistic prediction of SPI categories in Iran using sea surface temperature climate indices. Meteorol Atmos Phys 134:92. https://doi.org/10.1007/s00703-022-00931-4

Sobral BS, Oliveira-Júnior JF de, de Gois G, et al (2019) Drought characterization for the state of Rio de Janeiro based on the annual SPI index: trends, statistical tests and its relation with ENSO. Atmos Res 220:141–154. https://doi.org/10.1016/j.atmosres.2019.01.003

Sunusi N, Auliana NH (2025) Assessing SPI and SPEI for drought forecasting through the power law process: A case study in South Sulawesi, Indonesia. MethodsX 14:103235. https://doi.org/10.1016/j.mex.2025.103235

Svoboda M, Hayes M, Wood D (2012) Standardized precipitation index: user guide

Tabari H, Paz SM, Buekenhout D, Willems P (2021) Comparison of statistical downscaling methods for climate change impact analysis on precipitation-driven drought. Hydrol Earth Syst Sci 25:3493–3517. https://doi.org/10.5194/hess-25-3493-2021

Takayabu I, Kanamaru H, Dairaku K et al (2016) Reconsidering the quality and utility of downscaling. J Meteorol Soc Jpn Ser II 94A:31–45. https://doi.org/10.2151/jmsj.2015-042

Tirivarombo S, Osupile D, Eliasson P (2018) Drought monitoring and analysis: standardised precipitation evapotranspiration index (SPEI) and standardised precipitation index (SPI). Phys Chem Earth Parts A B C 106:1–10. https://doi.org/10.1016/j.pce.2018.07.001

Tyralis H, Papacharalampous G, Burnetas A, Langousis A (2019) Hydrological post-processing using stacked generalization of quantile regression algorithms: large-scale application over CONUS. J Hydrol 577:123957. https://doi.org/10.1016/j.jhydrol.2019.123957

Vapnik V (1998) Statistical learning theory Wiley. N Y 1:2

Vasiliades L, Loukas A, Patsonas G (2009) Evaluation of a statistical downscaling procedure for the estimation of climate change impacts on droughts. Nat Hazards Earth Syst Sci 9:879–894

Wilby RL, Charles SP, Zorita E, et al (2004) Guidelines for use of climate scenarios developed from statistical downscaling methods. Support Mater Intergov Panel Clim Chang available from DDC IPCC TGCIA 27:

Wilby RL, Wigley TML (1997) Downscaling general circulation model output: a review of methods and limitations. Prog Phys Geogr Earth Environ 21:530–548. https://doi.org/10.1177/030913339702100403

Wolpert DH (1992) Stacked generalization. Neural Netw 5:241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Yazdian H, Salmani-Dehaghi N, Alijanian M (2023) A spatially promoted SVM model for GRACE downscaling: using ground and satellite-based datasets. J Hydrol 626:130214

Yeganeh-Bakhtiary A, EyvazOghli H, Shabakhty N et al (2022) Machine learning as a downscaling approach for prediction of wind characteristics under future climate change scenarios. Complexity 2022:8451812. https://doi.org/10.1155/2022/8451812

Yerdelen C, Abdelkader M, Eris E (2021) Assessment of drought in SPI series using continuous wavelet analysis for Gediz Basin, Turkey. Atmos Res 260:105687. https://doi.org/10.1016/j.atmosres.2021.105687

Zareian MJ, Dehban H, Gohari A, Torabi Haghighi A (2024) Assessment of CMIP6 models performance in simulation precipitation and temperature over Iran and surrounding regions. Environ Monit Assess 196:701. https://doi.org/10.1007/s10661-024-12878-7

Zeleňáková M, Abd-Elhamid HF, Soľáková T et al (2025) Drought risk assessment in East Slovakia using SPI and RDI at different time scales. Nat Hazards 121:12095–12122. https://doi.org/10.1007/s11069-025-07274-4

Zhang Q, Li YP, Huang GH et al (2022) A novel statistical downscaling approach for analyzing daily precipitation and extremes under the impact of climate change: application to an arid region. J Hydrol 615:128730. https://doi.org/10.1016/j.jhydrol.2022.128730

Authors and Affiliations

Amirhossein Mirdarsoltany[1] · Matineh Imani Borhan[2] · Leila Rahimi[3] · Carlo De Michele[4]

✉ Leila Rahimi
rahimi.leila0093@gmail.com

[1] Department of Civil Engineering and Geodetic Science, Leibniz University, Hannover, Germany

[2] Department of Civil Engineering, K.N. Toosi University of Technology, Tehran, Iran

[3] Department of Civil and Environmental Engineering, Florida State University, Tallahassee, FL, USA

[4] Department of Civil and Environmental Engineering, Politecnico Di Milano, Milan, Italy