Water Harvesting Research Vol. 5, No. 2, Autumn & Winter 2022, Original Paper p. 241-256

University of Birjand

10.22077/JWHR.2023.6527.1090

Simulating Snow Cover Extent by Combined Principal Component Analysis and Artificial Intelligence Approaches Using Climatic Parameters

Amin Amini Rakan[a], Keivan Khalili[a*], Hossein Rezaei[a], Nasrin Fathollahzadeh Attar[b]

[a]Department of Water Engineering, Urmia University, Urmia, Iran
[b]Department of Statistics, University of Padua, Padua, Italy

[*]Corresponding Author, E-mail address: k.khalili@urmia.ac.ir
Received: 26 June 2023/ **Revised**: 09 August 2023/ **Accepted**: 15 August 2023

Abstract

Snow cover holds significant importance in hydrology as it plays a vital role in the water cycle and water resource management. Acting as a natural reservoir, snow stores water during winter and gradually releases it as it melts. This process contributes to streamflow, groundwater recharge, and overall water availability. Main goal of this study is the modeling and prediction of the changes in snow cover extent in Baranduz River basin, in Iran. Accurate modeling of snow cover area is crucial in hydrology as it enables precise predictions and assessments of water resources. These models incorporate snow accumulation, melt rates, and distribution, allowing informed decision-making for water management, agriculture, and ecosystem preservation. Therefore, the snow cover extent of the basin was extracted from MODIS 8-day maximum snow extent production from 2000 to 2019. Forty meteorological parameters, 20 satellite based and 20 surface stationary collected data, were used as the independent variables. The PCA was performed to parameters, and the PCA6 vector was used as input to the machine learning models. ANN, SVM, CART, and RF machine learning approaches were performed in this study. The results showed, all machine learning models had satisfactory performance and efficiency in modeling and predicting the snow cover extent. The PCA-RF model showed the highest accuracy. The RMSE and R^2 values for the PCA-RF model were 0.345 and 0.895, respectively, in the testing phase. Despite the fact that models have not been able to predict some of the boundary points accurately, they have still demonstrated acceptable performance.

Keywords: ANN, Baranduz River, CART, PCA, RF, Snow Cover Extent, SVM.

1. Introduction

Snow cover has a pivotal role in the hydrological processes of a river basin, making it essential for water resource management and climate studies. The depth and extent of snow cover directly influence hydrological processes, such as water availability, streamflow generation, and water storage. Acting as a natural reservoir, snow accumulates water during winter and gradually releases it during the melting season, ensuring a steady water supply throughout the year and generating streamflow (Karimi et al., 2016). However, snow cover significantly impacts the earth's energy balance due to its high albedo by reflecting a considerable amount of solar radiation. However, snow cover affects regional climate patterns, including temperature and precipitation distribution, which implicates the ecosystems and human activities (Li et al., 2021). Modeling snow cover in a river basin is important to studies and predictions of water availability, flood risks, climate change impacts. By incorporating factors such as snow depth, density, and distribution, these models can simulate snowmelt processes, estimate the timing and amount of runoff, and provide valuable insights for water resource planning, reservoir operations, and flood control strategies (Boudhar et al., 2020). Additionally, modeling snow cover allows for the evaluation of changes in snow accumulation and melt patterns under multiple climate scenarios, that

helps the development of adaptation strategies to the climate change impacts (Cohen and Rind, 1991).

Modelling and simulations, especially with machine learning and Artificial Intelligence approaches, are widely applicable in hydrologic studies. Artificial intelligence (AI) models have gained significant attention and proven themselves to be valuable in the hydrology. These models with the help of evolutionary algorithms, improve understanding of hydrological processes and enhance prediction accuracy. These models are used in hydrology to predict various parameters, such as flood and streamflow (Ahmed et al., 2021; Niu and Feng, 2021), droughts (Abbasi et al., 2019; Zhu et al., 2021), evapotranspiration (Abghari et al., 2012; Rezaverdinejad, 2016), precipitation (Khalili and Nazeri Tahroudi, 2016; Mehdizadeh et al., 2017; Nakhaei et al., 2023), air temperature (Yakut and Süzülmüş, 2020), soil temperature (Behmanesh and Mehdizadeh, 2017), dew point temperature (Attar et al., 2018), sediment transportation (Gupta et al., 2021), groundwater quality and levels (Abou Zakhem et al., 2017; Sahu et al., 2020). Snow cover modeling plays an important role in understanding the dynamics of snow in various climates. In recent years, machine learning techniques have emerged as valuable tools for improving snow cover estimation and mapping. One innovative approach proposed by (Hou et al., 2021) combines machine learning techniques with the Common Land Models to enhance the estimation of snow depth and fractional snow cover. Their study focuses on the northern Xinjiang region in China and employs MODIS fractional snow cover data. Snow has various parameters and indices to measure and all of them are usable in different studies for different purposes. The AI models are widely used in almost all of these areas and studies. The machine learning techniques have been applied in studies to estimating and modeling, snow depth (Kazama et al., 2021), snow cover (Kuter et al., 2021; Liu et al., 2020), snow density (Lee and Park, 2021), snow water equivalent (Kim et al., 2021), snow run-off (Duan et al., 2020), snow evaporation (Lin et al., 2020; Milly and Dunne, 2020) and the snow cover and storages relation (Bahrami et al., 2020a; Zhang et al., 2020). In conclusion, using machine learning techniques, especially in snow cover modeling, such as the integration of ML techniques into data mining schemes and the use of algorithms like ANN and SVR, holds promise for improving the performance of estimation. These approaches have demonstrated enhanced accuracy and consistency in snow studies.

Estimating snow cover area in a watershed, requires on-site snow measurement and snowfall monitoring, which is a challenging task and is practically not carried out by any government or private organization in any country. Nowadays, with the advancement and availability of remote sensing technology, snow cover data has become accessible in great areas like watersheds and river basis. Therefore, this study is based on remote sensing data. In many studies, MODIS sensor data has been suggested for snow cover extraction because of its accuracy and availability (Saavedra et al., 2018; Wu et al., 2021). A product of this sensor, using the Normalized Difference Snow Index (NDSI) algorithm, is the 8-day maximum snow extent data. This data essentially represents the maximum snow cover area over the past eight days (Riggs et al., 2015). The calculation details and the method for extracting snow cover area in the watershed will be fully described in the following section.

Although snow cover data is available to researchers through remote sensing, extracting and using this data requires technical knowledge and significant amount of time for satellite image classification. Since snow cover and snow water equivalent are input parameters in many hydrological processes and models, modeling and predicting snow cover extent as a time series can eliminate the time and complexities of remote sensing approach. It can also make this data accessible to individuals who have no specialized knowledge of remote sensing or have limited technical expertise. One study that considered snow cover area in a watershed as a monthly time series is the (Karimi et al., 2016). In this study, the snow cover area of the Haraz River basin in Iran was extracted from MODIS and modeled as a monthly time series. Stochastic models from the ARIMA and ARCH families were utilized in this study and the RMSE and

AIC values 0.878 and 2246.125, respectively were calculated for models. It appears that the performance of snow cover time series models can be significantly improved by utilizing machine learning algorithms. According to the background, despite the importance of the prediction of changes in snow cover area in the watersheds, very limited studies have been done in this field, also If the study and evaluation of snow cover changes would be carried out with respect to the spatial and temporal dimensions, it would be led to more accurate conclusions in hydrological studies. The mathematical modeling of the snow cover extent could help the hydrologists to skip the remote sensing steps and helps researchers, who have limited knowledge of RS and GIS. Therefore, the main purpose of this study is to fit or train machine learning models for estimating the monthly snow cover extent in Baranduz River basin, located in West Azerbaijan province at North West of Iran. The snow extent is extracted from MODIS 8-day maximum snow extent data from 2000 to 2019. The models are trained by 40 monthly climatology parameters including both satellite and surface stationary data.

2. Materials and Methods

2.1. Study Area

The Baranduz River basin is located west of Lake Urmia in Iran. Its upstream originates near the borders of Turkey and Iran, and after flowing for a distance of 56 km, it merges with the Balanej River and finally after another 11 km, flows into Lake Urmia. The total area of the Baranduz River basin measures 1151 km^2, and has an average annual discharge of 165 MCM. The average annual precipitation of this basin is 268.3 mm. The lowest elevation of the basin lies adjacent to Lake Urmia, starting at approximately 1272 meters and gradually rising to 3453 meters in the mountainous regions. Notably, this area is renowned for its apple farms that cultivate premium apple varieties. During the summer season, irrigation for the majority of these apple farms relies on the Baranduz River streamflow, with the river's discharge primarily dependent on snow presence and the extent of snow cover in the higher elevations of the region. Fig. 1 illustrates the map of the study area, providing a comprehensive depiction of the Baranduz River basin and the geographic characteristics of the study area and surface data stations.

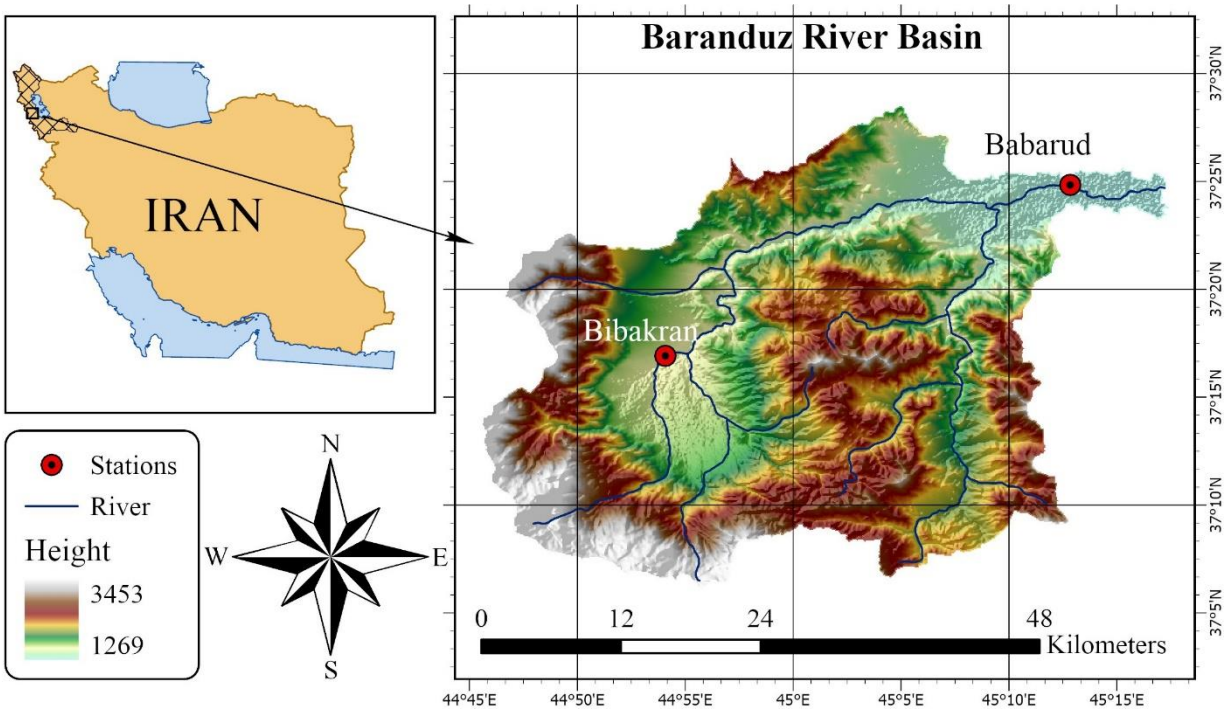

Fig. 1. Baranduz River basin characteristics and hydro-climatology stations.

2.2. Data and Preprocessing

2.2.1. Climatology Data

This research incorporates with four different datasets. Two of these datasets consist of observational data obtained from the

Bibakran and Babarud Hydro-climatology stations, whose geographic coordinates are illustrated in Fig. 1. Additionally, we utilized two datasets from the NASA (LaRC) Power Project, which utilizes the MERRA-2 model archive to generate meteorological data. These datasets encompass global coverage with a resolution of 0.05 decimal degrees. These data were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program. This data is available from (https://power.larc.nasa.gov). Table 1 provides an overview of the datasets employed in this study, including the specific parameters utilized from each dataset. The snow-covered modeling area within the basin was trained and tested using a collection of 40 monthly observation parameters, from February 2000 to October 2019.

Table 1. Meteorological parameters collected from Babarud and Bibakran ground hydrometeorological stations and satellite base data from the NASA LaRC project.

Hydrometeorological stations Data		MERRA-2 (LaRC) Data	
Parameters	Unit	Parameters	Unit
Evaporation	mm	Average temperature	°C
Precipitation	mm	Maximum average temperature	°C
Minimum temperature	°C	Minimum average temperature	°C
Minimum average temperature	°C	Sun insolation	MJ/m^2.day
Average temperature	°C	Infrared insolation	MJ/m^2.day
Maximum average temperature	°C	Relative humidity	%
Maximum temperature	°C	Dew point temperature	°C
Relative humidity 12:30	%	Wind speed in 2meter	m/s
Relative humidity 06:30	%	Precipitation	Mm
Relative humidity 18:30	%	Surface pressure	kPa

2.2.2. Snow Cover Data

The snow cover area analysis in this study utilized from Moderate Resolution Imaging Spectroradiometer (MODIS) 8-day maximum snow extent data, as described by (Hall and Riggs, 2016). The MODIS data employed in this study utilizes the Normalized Difference Snow Index (NDSI) to determine the maximum snow extent over eight-day intervals. This approach effectively mitigates the impact of cloud cover, a significant challenge encountered in remotely sensed snow cover data (Hall et al., 2002). The specific MOD10A2 data series, accessible online through the international snow and ice center website, was used for this analysis. Various amounts of worldwide satellite snow cover data are accessible from the national snow and ice data center. These data are all available from (https://nsidc.org) website. For more information about the snow cover extent data which used in this research, you could also see:

https://doi.org/10.5067/MODIS/MOD10A2.006).

ArcGIS software was employed to extract the snow cover area of the 8-day MODIS layers from February 2000 to October 2019, which were cropped to match the extent of the Baranduz River basin. At the end, from each month's layers within this timeframe, the monthly average snow cover extent was derived out. The main statistical information of the monthly snow cover extent in the Baranduz River basin is shown in Table 2.

Table 2. Main statistical information of the monthly snow cover extent in Baranduz River basin

Max	Mean	Standard Deviation	Skew	Kurtosis
1229.18	312.83	372.23	0.95	2.42

2.3. Box-Cox Transformation

Box-Cox transformation, introduced by Box and Cox (Box and Cox, 1964), offers a powerful method to address anomalies such as non-additivity, non-normality, and heteroscedasticity. In data analysis, it is often assumed that the series or samples are normally distributed (Sakia, 1992). This technique employs a simple linear regression between all variables and the objective variable, using a maximum likelihood test to evaluate the fitness of the objective variable to a normal distribution curve.

The transformation technique determines the optimal power to be applied to the objective series, aiming to align the sample data series with a normal distribution. Additional information and in-depth details can be found in the original paper (Box and Cox, 1964) and subsequent studies (Osborne, 2010). To assess the fit of the sample data to a statistical distribution, the Q-Q plot is a valuable tool. The Q-Q plot, short for Quantile-Quantile plot, is a scatter plot that compares two sets of data: the theoretical distribution and the samples being tested. If the sample data and the distribution align perfectly, the Q-Q plot will show a straight line. Any deviations between the two sets of data indicate disparities between the samples and the theoretical distribution (Box and Cox, 1964; Osborne, 2010).

To avoid any unwanted effects of the high standard deviation of a parameter to the training phases, all parameters are centered to an average of zero and a standard deviation of 1. The Box-Cox method is employed in this study to evaluate the normality of parameters and identify the optimal transformation for snow cover area, aiming to achieve a normal distribution and alleviate the influences of skewness. In order to randomize the data and reducing the unwanted errors of statistical distribution, all the parameters are standardized, and the monthly snow cover extent time series, transformed by the λ value of 0.275 using Eq. 1.

$$Y^N = \left(Y^\lambda - 1\right) \Big/ \lambda \qquad (1)$$

In this equation, Y^N is the normalized series and Y is the original series (Osborne, 2010).

2.4. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used statistical technique that has found application in various fields, including hydrology. It was first introduced by Karl Pearson in 1901, making it over a century old (Pearson, 1901). PCA aims to simplify the analysis of complex datasets by transforming the original variables into a new set of uncorrelated variables called principal components. These principal components are linear combinations of the original variables and are ordered based on the amount of variation they capture in the data. By selecting a subset of the principal components, PCA allows for a reduction in the dimensionality of the dataset while retaining the most important information.

In the field of hydrology, PCA has proven to be a valuable tool for analyzing and interpreting multivariable hydrological series. Hydrological processes involve numerous variables such as precipitation, streamflow, evapotranspiration, and water quality parameters, which are often correlated (Jehn et al., 2020). PCA helps in identifying the dominant patterns of variation in these variables and provides insights into the underlying processes. By extracting the principal components, hydrologists can reveal the common modes of variability and characterize the spatial and temporal patterns of hydrological phenomena. Especially in groundwater quality analysis, PC Analysis was used to cluster the observation wells and determine the most important pollutant in the studied groundwater aquifer (Bahrami et al., 2020b). PCA has been applied in hydrological studies to investigate various aspects such as drought analysis, flood prediction, water quality assessment, and climate change impacts on hydrological processes (Abou Zakhem et al., 2017).

In summary, Principal Component Analysis is a statistical technique that has been employed in hydrology for over a century. It works by transforming original variables into uncorrelated principal components, capturing the dominant patterns of variation in multivariable hydrological data. By reducing the dimensionality of the dataset, PCA aids in the interpretation and understanding of hydrological processes and has found applications in diverse areas within hydrology research (Song et al., 2010; Westra et al., 2007).

2.5. Artificial Neural Network (ANN)

Artificial Neural Networks are popular methods in the field of artificial intelligence used for solving various problems such as regression, prediction, classification, and information retrieval (Yakut and Süzülmüş, 2020). These networks are inspired by the neural cells of living organisms and consist of

multiple layers of neural cells that are interconnected and trained as a unit to provide results. In these networks, information is transmitted from the input layer to the hidden layers, where different features of the data are extracted and processed (Braspenning et al., 1995).

The layers of an artificial neural network are stacked one after another, and each layer passes the output information to the next layer. The first layer is typically the input layer, responsible for receiving the data and passing it to the next layer. The subsequent layers are the hidden layers, which process the information using weights or fixed values defined between the neurons and ultimately produce the output. The last layer is the final output of the neural network, adjusted based on the desired output and the number of output variables (Hassoun, 1995).

To build an artificial neural network, the number of layers and the number of neurons in each layer need to be determined. Then, the inputs and desired outputs are defined, and the fixed values between the different layers are specified. After these steps, the network is trained to produce the desired output given the provided inputs. For this purpose, a set of data is fed into the network, and the weights between the layers are adjusted in a way that minimizes the error in the output (Yegnanarayana, 2009). Various algorithms are used to adjust the weights, and one of them is the Back-Propagation algorithm. This algorithm adjusts the weights of the fixed values between the layers based on the comparison between the original and generated data. This process continues until a convergence threshold is reached, and the algorithm stops (Riedmiller and Braun, 1993). The convergence threshold in this algorithm is the error between the observed data and the network's output, which in this study is set to 0.05. Additionally, the present study's artificial neural network has three hidden layers with 7, 5, and 3 neurons and an initial fixed value of 1 for each intermediate layer. Fig. 2 illustrates the trained artificial neural network in the current study. The six blue neurons are the input neurons of the PCA6 vectors and the last neuron of the network is the output or the snow cover extent value.

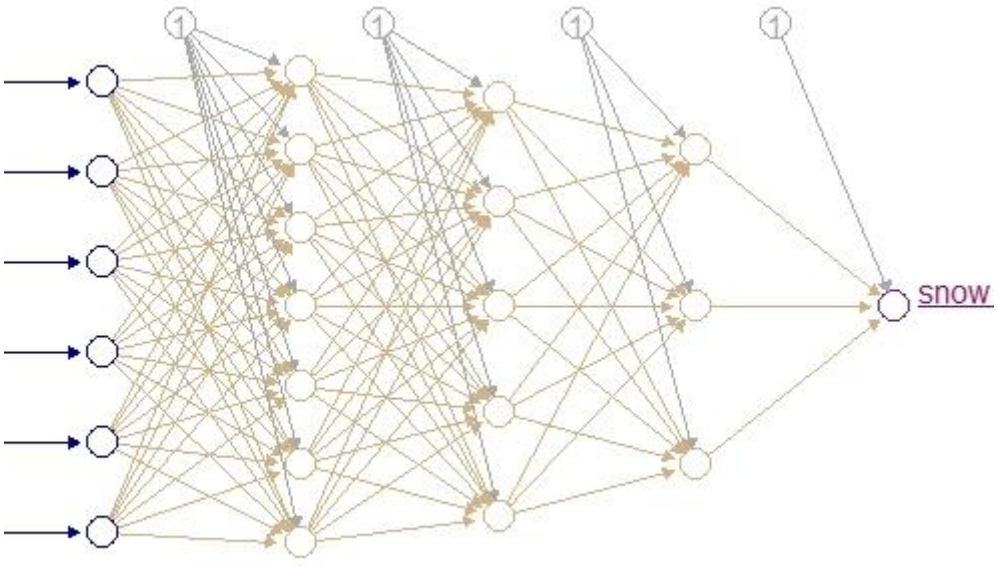

Fig. 2. Structure of the trained neural network model. Bias or constant weights are one. Three hidden (beige) layers with 7, 5 and 3 neurons. Six input neurons (blue) for PCA1 to PCA6. One output neuron (snow).

At first, the network has been trained by training data, including 80% of the PCA6 vectors and the snow cover time series, after the training, the test snow cover time series were calculated from the PCA6 test data, using the trained artificial neural network. All computational stages related to the ANN model in this study were conducted using the NeuralNet package in the R programming software.

2.6. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression analysis. The main idea behind SVM is to find a hyperplane or a higher-dimensional space that provides the best separation between different data points (Vapnik, 1998). This algorithm was initially introduced by Vapnik in 1963. In 1995, Vapnik and Cortes extended this method to solve multi-dimensional classification and regression problems. In this algorithm, after fitting the hyperplane, the model seeks to find a boundary that maximizes the margin between the two classes. This boundary is determined by the algorithm in a way that optimally maximizes the distance between the optimal margin and the same-class data points, while minimizing the distance between the optimal margin and the different-class data points (Cortes and Vapnik, 1995; Vapnik and Lerner, 1963).

The best differentiating hyperplane is the one with the maximum distance between the two classes, so it should assign the minimum value to itself. In non-linear classification tasks where the data points are not linearly separable, various mappings (kernel functions) have been defined in SVM. These mappings transfer the problem to a new space where the data points can be linearly separated. These kernel functions, also known as kernels, are convex functions that have a unique real solution under optimization conditions (Steinwart and Christmann, 2008).

To understand the precise functioning of kernel functions, the equations, and the computation methods of Support Vector Machines in various regression and classification problems, references such as (Gunn, 1998), (Hearst et al., 1998) and (Steinwart and Christmann, 2008) can be consulted. Furthermore, the application and fitting of SVM on hydrological problems have been extensively described in studies such as (Niu and Feng, 2021), (Yakut and Süzülmüş, 2020), and (Zhu et al., 2021). All computational stages related to the SVM model in this study were conducted using the E1071 package in the R programming software.

2.7. Classification and Regression Tree (CART)

The Classification and Regression Trees (CART) model is a powerful machine learning approach used for data classification and prediction. This model operates based on a hierarchical tree structure and was first introduced by (Breiman, 1984). At each node of the tree, the data is divided into two distinct groups based on specific features, and each group can be further subdivided into binary subsets (parent and child) based on other features. This process continues until the algorithm reaches a threshold regarding the number of tree branches or the observations present in each tree branch (Nakhaei et al., 2023). As the trees grow larger, produce more branches and nodes, they incorporate more information into the model.

After constructing a tree with a maximum size, tree pruning is performed using one of the pruning methods, starting from the leaves and moving towards the root (from child to parent). The goal of the CART model is not to create just one pruned tree; instead, it aims to generate a sequence of pruned trees, each of which represents candidate options for the final optimal tree (Wang et al., 2023). To identify a good tree, its performance is evaluated on independent test data. Finally, the constructed decision tree is validated and assessed on a separate set of test data (Sharma and Kumar, 2016). All computational stages related to the CART model in this study were conducted using the rpart package in the R programming software.

2.8. Random Forrest (RF)

The RF method is a machine learning technique developed based on the Classification and Regression Tree (CART) approach. CART was initially introduced by Breiman in 1984 as a solution for classification problems. In 2001, Breiman further extended this method to develop RF, which offers improved results in addressing regression problems and is currently regarded as one of the most powerful machine learning techniques (Breiman, 2001).

In this methodology, a random set of decision trees is constructed based on the bootstrap method, each independently executed on the training data. Subsequently,

each tree within the ensemble independently fits a prediction model for the test data using the features extracted from the training data. These prediction models undergo evaluation using various metrics and are optimized based on input variables. Ultimately, the predictions from all trees are aggregated through averaging, forming the final prediction and serving as the system's output (Biau, 2012). The Random Forest (RF) method is recognized as a robust and widely employed technique in machine learning problems, typically applied to tasks involving large and diverse datasets. In such ensemble learning techniques, it is assumed that the collective accuracy of group training surpasses that of individual algorithmic models (Vapnik, 1998).

To perform classification or regression using the Random Forest (RF) method, two parameters need to be provided by the user. One of these parameters is the size of the subsets or, in other words, the number of features for each tree (M), while the other parameter is the number of subsets or the number of trees (T). The values of M and T are chosen based on the machine or computer's hardware performance and memory. Typically, the value of M ranges from a few hundred to several thousand, and the value of T is suggested to be the square root of the number of variables. It is evident that as the values of M and T increase, the final ensemble model will be closer to reality, but the algorithm becomes larger, and its learning time increases (Biau and Scornet, 2016). In the present study, the value of M is set to 500, and the value of T is set to 3. All computational stages related to the RF model in this study were conducted using the Random Forrest package in the R programming software.

2.9. Evaluation Indicators

An evaluation of model performance entails the utilization of a set of dimensionless criteria applied to both actual observations and model-derived data. The purpose of employing these criteria is to enable a comprehensive comparison between the models under study. Based on the conducted researches, it has been determined that the best criteria for comparing the models and choosing the best models are the root mean square error (RMSE) and the mean absolute error (MAE) for evaluating the model error, and the coefficient of determination (R^2) and the correlation coefficient (R) between the observed and calculated data. The equations and calculation methods for each criterion are discussed below:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i}^{N}(Y_i - X_i)^2}{N}} \quad (2)$$

$$\mathrm{MAE} = \frac{\sum_{i}^{N}|Y_i - X_i|}{N} \quad (3)$$

$$R = \frac{\sum(Y_i - \bar{Y})\,(X_i - \bar{X})}{\sqrt{\sum(Y_i - \bar{Y})^2 \sum(X_i - \bar{X})^2}} \quad (4)$$

$$R^2 = 1 - \frac{\sum(Y_i - X_i)^2}{\sum(Y_i - \bar{Y})^2} \quad (5)$$

In these equations, Y_i and X_i are the observational series and calculated from the model, respectively. Also, $\bar{Y}$ and $\bar{X}$ are the average of the series and N is the number of observations.

Another evaluation criterion utilized in this study is the commonly used Akaike Information Criterion (AIC). The AIC is calculated for different models using Eq. 6. A smaller AIC value indicates a better fit between the observation values and the values calculated from the model. Therefore, any model with a lower AIC is selected as a more suitable model.

$$\mathrm{AIC} = N * \mathrm{Ln}\left(\frac{SS_E}{N}\right) + 2K \quad (6)$$

In this equation, SS_E represents the sum of squared errors of the model, N denotes the number of observations, and K is the number of variables used in the model plus one (Akaike, 1974).

3. Results and Discussion

As mentioned in previous sections, in this study, monthly time series of snow cover area in the Baranduz River basin in Iran were extracted from 8-day MODIS data and modeled using 40 meteorological parameters. The data were first standardized and normalized, and then dimension reduction applied using Principal Component Analysis (PCA). Since the meteorological data in this study, exhibited high intercorrelations, which is one of the characteristics of hydrological data, dimension reduction methods needed to be employed in order to prevent computational complexity and avoid collinearity issues in the

models. In this study, PCA was utilized as the dimension reduction method and 80% of data are used for training models and the 20% of data are used for testing the trained models. All of the results of PCA and subsequent modeling are described in the following section.

3.1. PCA Analysis

In this method, new orthogonal vectors are generated using the study data, each capturing a portion of the total variance of the variables. Fig. 3 illustrates the percentage of variance coverage by each PCA vector in this study. To involve a higher percentage of variance in the model, a greater number of PCA vectors should be used for modeling.

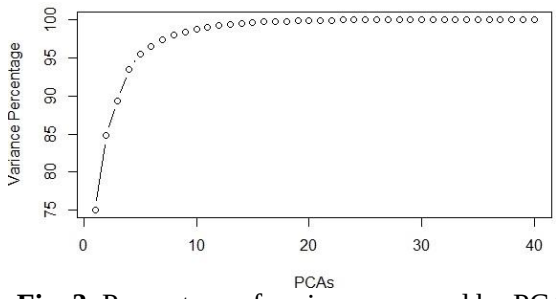

Fig. 3. Percentage of variance covered by PCA orthogonal vectors.

As shown in Fig. 3, the firs vector covers almost 75% of the variance and more than 96% of the variance is covered by 6 PCA vectors. This highlights the significant dimensionality reduction achieved by the PCA method and emphasizes its effectiveness in capturing the essential information from the original dataset. From PCA 7 to PCA 40, the coverage percentage of variance is rising too slowly and covers approximately 4% or 5%, therefor these PCA vectors are excluded from consideration. Consequently, by employing the PCA dimensionality reduction method, the number of independent variables, originally 40, has been reduced to 6. These 6 variables cover approximately 96% of the variance, which makes them suitable inputs for machine learning models.

3.2. Modeling Results

The 6 PCA vectors described in the previous section have been utilized as inputs for machine learning models. In the following section, the results, performance and efficiency of the ANN, SVM, CART, and RF models are presented, discussed and examined. Also, following the model fitting process, evaluation indices have been calculated for all models during the training and testing phases. The values of these comparative indices are presented in Table 3 calculated from all fitted models.

Table 3. Performance and Error evaluation of the models during both training and testing phases.

Model	Phase	RMSE	MAE	R^2	R	AIC
PCA- ANN	Train	0.122	0.088	0.984	0.992	-794.9
	Test	0.381	0.318	0.873	0.934	-71.02
PCA- SVM	Train	0.195	0.143	0.963	0.981	-616.6
	Test	0.351	0.263	0.886	0.941	-77.85
PCA- CART	Train	0.319	0.245	0.898	0.947	-428.6
	Test	0.352	0.288	0.882	0.939	-77.56
PCA- RF	Train	0.146	0.117	0.982	0.991	-727.0
	Test	0.345	0.276	0.895	0.946	-79.20

As shown in Table 3, the ANN model demonstrates the best performance among the other machine learning models for train phase. The root mean square error (RMSE) is calculated 122 for this model, which is the lowest among the other models, indicating the least amount of error. Additionally, the calculated R-squared (R^2) value for this model is 0.99, indicating a highly suitable and significant result. Subsequently, the R^2 values calculated for the RF, SVM, and CART models are 0.98, 0.96, and 0.89, respectively. The best model in the training phase was the ANN model. ANN models are known for the best training accuracy, these models can train from complex phenomena's behavior very accurately but the time of training can take longer than other models. On the other hand, in test phase, the performance of the ANN model did not yield the best results. During the testing phase, the SVM and RF models demonstrated the best performance. RMSE, MAE, R^2, R and AIC for SVM model in testing phase were 0.351, 0.263, 0.886, 0.944 and -77.85 and for

the RF model were 0.345, 0.276, 0.895, 0.946 and -79.20, respectively. The calculated comparative indices for both models were very close to each other as shown in Table 3, but despite these indices and RMSE, the MAE for the PCA-SVM model were slightly lower than the PCA-RF model. This indicates that the SVM model has committed more significant errors in terms of magnitude, although its overall computations have resulted in lower error than the RF model. However, the R^2, R, and AIC values of PCA-RF model, were better than the PCA-SVM model. Also, based on the results presented in Table 3, the poor performance of the CART model is clearly evident. All models in the current study showed significantly better performance than the stochastic models of the (Karimi et al., 2016) study.

Fig. 4 illustrates the observed and calculated data from the models during the testing and training phases. In Fig. 4, regression lines are also visible, indicating the relationship between the observed and calculated data.

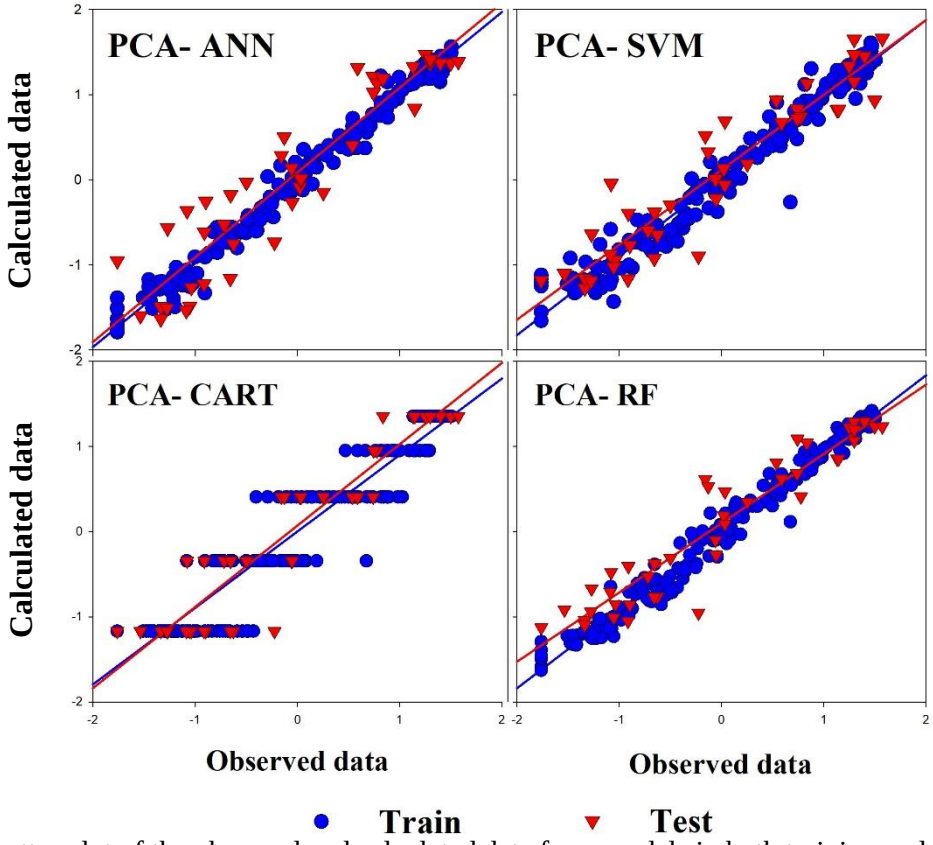

Fig. 4. Scatter plot of the observed and calculated data from models in both training and testing phases.

As shown in Fig. 4, it is clear that the PCA-CART model is not a suitable model for regression problems and the data has been classified using a classification approach but the modified version of it, the Random Forrest approach showed the best regression performance in this study. The classification tree of the PCA-CART model is shown in Fig. 5. As shown in Fig. 5, the first branch of the classification tree is divided by the 0.322 value of the PCA1, which is the most important component of the Principal Component Analysis and in this study cover the 75% of the variance Fig. 3. As shown in Fig. 4 and Table 3, all three models, PCA-RF, PCA-SVM, and PCA-ANN, have successfully predicted more than 90% of the observations accurately. However, all three models have encountered difficulties in predicting minimum values or values corresponding to zero snow cover in the watershed area, which mostly occurs in summer times. These models have tended to overestimate minimum values, predicting them to be greater than zero. The red triangles represent the performance of the models during the testing phase. In the PCA-ANN model shown in Fig. 4, it is evident that the test data demonstrates higher errors and deviated significantly from the regression line.

This observation further validates the results presented in Table 3, which indicate a substantial decrease in the performance of the PCA-ANN model during the testing phase. The structure of artificial neural networks (ANN) is designed in such a way that they can effectively learn from data and understand the relationships between variables. However, in some cases, this training phase with high accuracy can lead to model overfitting, primarily due to the presence of interdependencies or among variables (Braspenning et al., 1995). Fig. 6, Fig. 7, Fig. 8 and Fig. 9 display the observed and calculated data derived from PCA-ANN, PCA-SVM, PCA, CART, and PCA-RF models. Additionally, the corresponding error values calculated from each model during both the training and testing phases are presented.

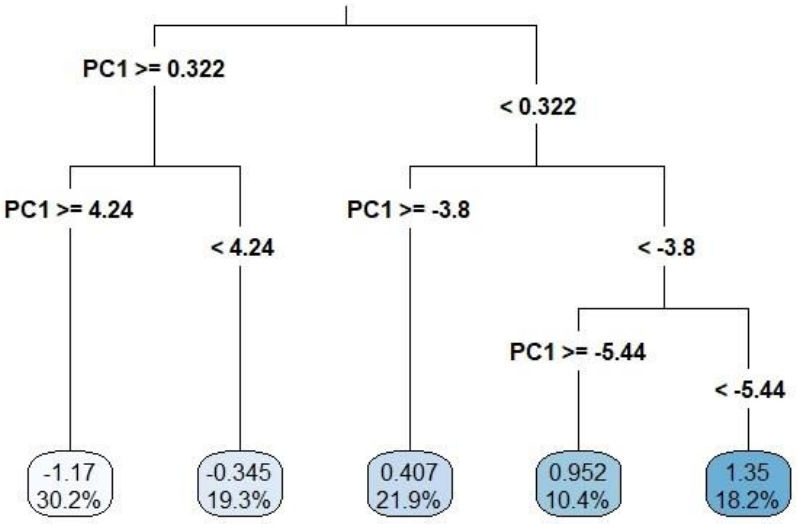

Fig. 5. Classification tree of the PCA-CART model.

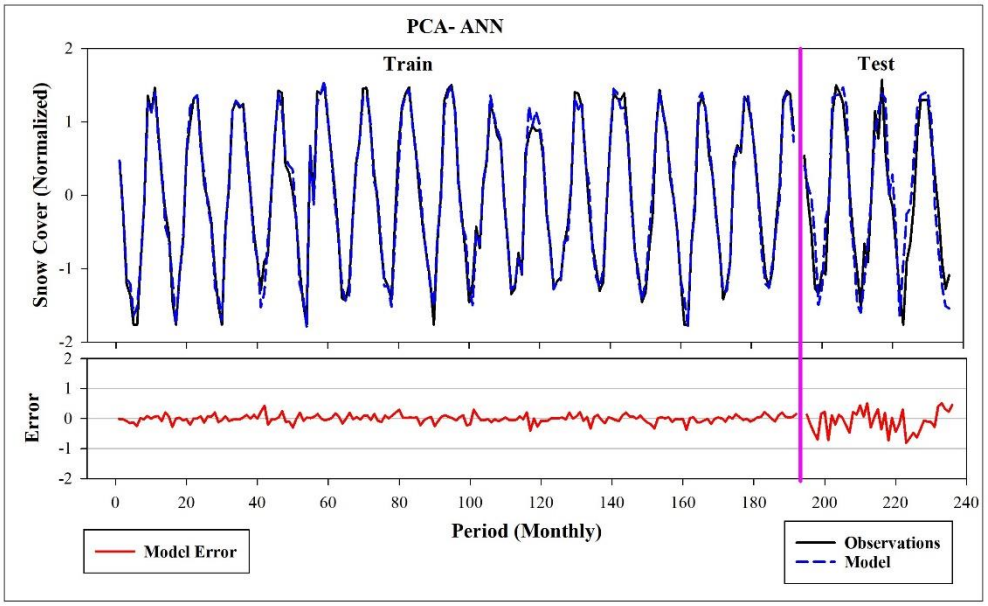

Fig. 6. Observed and calculated values from PCA-ANN model with the error values calculated from both training and testing phases.

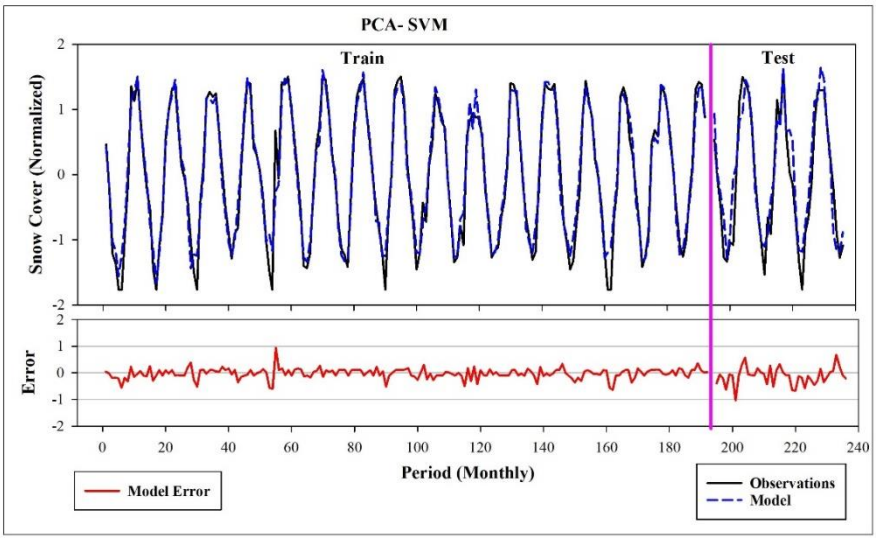

Fig. 7. Observed and calculated values from PCA-SVM model with the error values calculated from both training and testing phases.

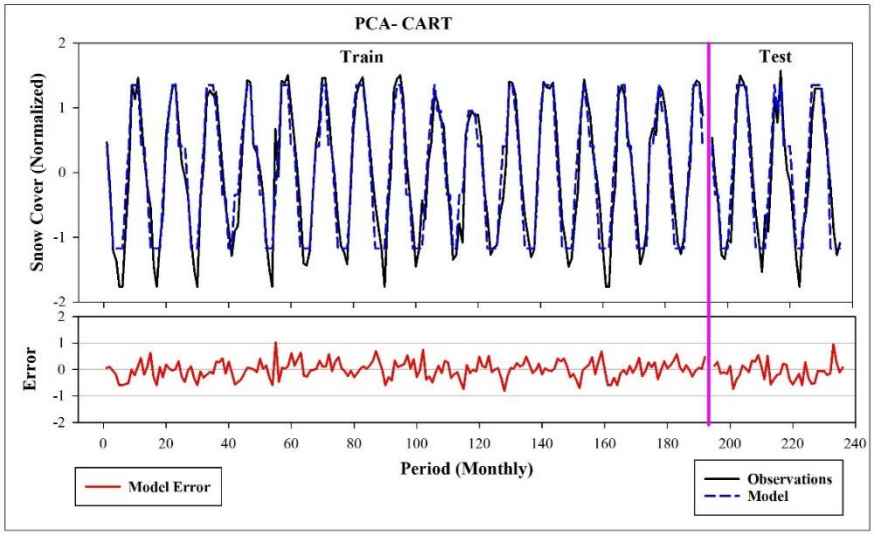

Fig. 8. Observed and calculated values from PCA-CART model with the error values calculated from both training and testing phases.

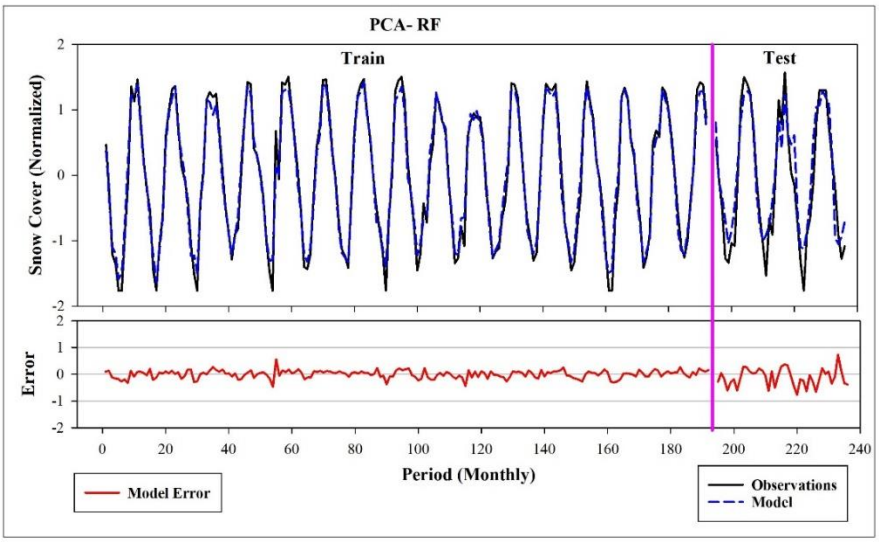

Fig. 9. Observed and calculated values from PCA-RF model with the error values calculated from both training and testing phases.

All models have successfully predicted changes in snow cover extent within the Baranduz River basin. As observed, all models have performed well in predicting maximum snow cover points and demonstrated significant accuracy (Fig. 6, Fig. 7, Fig. 8, and Fig. 9). Additionally, the PCA-ANN model has performed perfect in predicting minimum points, both during the testing and training phases. However, other models, especially the PCA-CART model, showed poor performance in minimum points prediction, as they appear to overestimate the minimum snow surface levels compared to the actual values.

Moreover, an increase in prediction error values during the testing phase is evident for all models, as expected from the results in Table 3. Fig. 6 to Fig. 9, clearly illustrate the increasing error values during the testing phase. The error plots for all models in the testing phase indicate higher values, except for the PCA-CART model, which exhibits higher errors in both the training and testing phases.

Overall, the performance of all models has been acceptable, and the results show relative superiority compared to the (Karimi et al., 2016) study. However, the PCA-SVM model has demonstrated lower error rates compared to other models, and the correlation between the actual and predicted data from the PCA-RF model has been higher than all of the other models. Choosing between the PCA-SVM and PCA-RF models is challenging due to the small differences in the model's performance. Nevertheless, based on the Akaike Information Criterion (AIC), the PCA-RF model can be selected as the best model among the all studied models.

4. Conclusion

In this study, the snow cover extent and its changes in the Baranduz River basin have been modeled using machine learning models and 40 meteorological parameters. Initially, the time series of snow cover was extracted from MODIS sensor data for the years 2000 to 2019 and sorted on a monthly average basis. Additionally, 20 meteorological parameters from ground hydro climatology stations in Babarud and Bibakran, located within the basin area, and 20 meteorological parameters from the NASA LaRC project, were used as independent variables on a monthly basis. After standardizing and normalizing the data using the Box-Cox transformation, PCA (Principal Component Analysis) was applied to variables for dimension reduction purposes in models. The PCA1 to PCA6 orthogonal vectors were formed and used as inputs for the ANN (Artificial Neural Networks), SVM (Support Vector Machine), CART (Classification and Regression Trees), and RF (Random Forrest) models.

The modeling results demonstrated that the machine learning models used in this study performed remarkably well and were capable of accurately predicting the behavior and changes of the snow cover in the Baranduz River basin. The R^2 values for all models, both in the testing and training phases, exceeded 0.87, indicating high performances of the models. The PCA-ANN model showed the best performance in the training phase, with the R^2 value of 0.99, while in the testing phase, the PCA-SVM and PCA-RF models showed the best performance. The prediction error for the PCA-SVM model (MAE=0.263) was lower than the PCA-RF model (RMSE=0.276), but the correlation coefficient and coefficient of determination in the PCA-RF model were higher (R=0.946, R^2=0.895). Choosing the best model between these two models is challenging due to their slight performance differences. However, considering the lower value of the calculated Akaike information criterion (AIC=-79.20) and Root Mean Square Error (RMSE=0.345), the PCA-RF model demonstrated the best performance in modeling the snow cover extent in the Baranduz River basin.

In conclusion, each model brings its strengths – SVM excels in capturing complex relationships, ANN handles intricate patterns, CART offers interpretable decision-making, and RF provides ensemble accuracy. The use and development of these models can significantly improve our understanding of snow cover dynamics, benefiting fields like water resource management, disaster preparedness, and ecological research.

5. Disclosure Statement

No potential conflict of interest was reported by the authors.

6. References

Abbasi, A., Khalili, K., Behmanesh, J., & Shirzad, A. (2019). Drought Prediction Using GEP-GARCH Hybrid Model (Case Study: Salmas Synoptic Station). *Iranian Journal of Soil and Water Research*, *50*(6), 1317-1329. https://doi.org/10.22059/ijswr.2019.271596.668069

Abghari, H., Ahmadi, H., Besharat, S., & Rezaverdinejad, V. (2012). Prediction of daily pan evaporation using wavelet neural networks. *Water Resources Management*, *26*(12), 3639-3652.

Abou Zakhem, B., Al-Charideh, A., & Kattaa, B. (2017). Using principal component analysis in the investigation of groundwater hydrochemistry of Upper Jezireh Basin, Syria. *Hydrological Sciences Journal*, *62*(14), 2266-2279.

Ahmed, A. M., Deo, R. C., Feng, Q., Ghahramani, A., Raj, N., Yin, Z., & Yang, L. (2021). Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *Journal of Hydrology*, *599*, 126350.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Attar, N. F., Khalili, K., Behmanesh, J., & Khanmohammadi, N. (2018). On the reliability of soft computing methods in the estimation of dew point temperature: The case of arid regions of Iran. *Computers and electronics in agriculture*, *153*, 334-346.

Bahrami, A., Goïta, K., & Magagi, R. (2020a). Analysing the contribution of snow water equivalent to the terrestrial water storage over Canada. *Hydrological Processes*, *34*(2), 175-188.

Bahrami, M., Khaksar, E., & Khaksar, E. (2020b). Spatial variation assessment of groundwater quality using multivariate statistical analysis (Case Study: Fasa Plain, Iran). *Journal of Groundwater Science and Engineering*, *8*(3), 230-243.

Behmanesh, J., & Mehdizadeh, S. (2017). Estimation of soil temperature using gene expression programming and artificial neural networks in a semiarid region. *Environmental Earth Sciences*, *76*(2), 76.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063-1095.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197-227.

Boudhar, A., Ouatiki, H., Bouamri, H., Lebrini, Y., Karaoui, I., Hssaisoune, M., Arioua, A., & Benabdelouahab, T. (2020). Hydrological Response to Snow Cover Changes Using Remote Sensing over the Oum Er Rbia Upstream Basin, Morocco. In *Mapping and Spatial Analysis of Socio-economic and Environmental Indicators for Sustainable Development* (pp. 95-102). Springer.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211-243.

Braspenning, P. J., Thuijsman, F., & Weijters, A. J. M. M. (1995). *Artificial neural networks: an introduction to ANN theory and practice* (Vol. 931). Springer Science & Business Media.

Breiman, L. (1984). *Classification and Regression Trees* (1st Edition ed.). Routledge.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Cohen, J., & Rind, D. (1991). The effect of snow cover on the climate. *Journal of climate*, *4*(7), 689-706.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

Duan, Y., Liu, T., Meng, F., Yuan, Y., Luo, M., Huang, Y., Xing, W., Nzabarinda, V., & De Maeyer, P. (2020). Accurate simulation of ice and snow runoff for the mountainous terrain of the kunlun mountains, China. *Remote Sensing*, *12*(1), 179.

Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report*, *14*(1), 5-16.

Gupta, D., Hazarika, B. B., Berlin, M., Sharma, U. M., & Mishra, K. (2021). Artificial intelligence for suspended sediment load prediction: a review. *Environmental Earth Sciences*, *80*(9), 1-39.

Hall, D., & Riggs, G. A. (2016). *MODIS/Terra Snow Cover 8-Day L3 Global 500m SIN Grid, Version 6*. Boulder, Colorado USA. , NASA National Snow and Ice Data Center Distributed Active Archive Center.

Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGirolamo, N. E., & Bayr, K. J. (2002). MODIS snow-cover products. *Remote sensing of environment*, *83*(1-2), 181-194.

Hassoun, M. H. (1995). *Fundamentals of artificial neural networks*. MIT press.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.

Hou, J., Huang, C., Chen, W., & Zhang, Y. (2021). Improving snow estimates through assimilation of MODIS fractional snow cover data using machine learning algorithms and the common land model. *Water Resources Research*, *57*(7), e2020WR029010.

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., & Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, *24*(3), 1081-1100.

Karimi, S., Niksokhan, M. H., & Karimi, S. (2016). Modeling snow cover area and predicting its changes in Haraz catchment. *Imaging*, *2*(4), 450-455.

Kazama, S., Sakamoto, K., Salem, G. S. A., & Kashiwa, S. (2021). Improving the Accuracy of Snow and Hydrological Models Using Assimilation by Snow Depth. *Journal of Hydrologic Engineering*, *26*(1), 05020043.

Khalili, K., & Nazeri Tahroudi, M. (2016). Performance evaluation of ARMA and CARMA models in modeling annual precipitation of Urmia synoptic station. *Water and Soil Science*, *26*(2-1), 13-28.

Kim, R. S., Kumar, S., Vuyovich, C., Houser, P., Lundquist, J., Mudryk, L., Durand, M., Barros, A., Kim, E. J., & Forman, B. A. (2021). Snow Ensemble Uncertainty Project (SEUP): Quantification of snow water equivalent uncertainty across North America via ensemble land surface modeling. *The Cryosphere*, *15*(2), 771-791.

Kuter, S., Aksu, C., Bolat, K., & Akyurek, Z. (2021). *An Alternative Machine Learning-Based Methodology for H-SAF H35 Fractional Snow Cover Product*.

Lee, E., & Park, S. K. (2021). *Optimization of snow density parameter of Noah Land Surface Model using micro-genetic algorithm for estimating snow depth*.

Li, G., Zhao, Y., Zhang, W., & Xu, X. (2021). Influence of snow cover on temperature field of frozen ground. *Cold Regions Science and Technology*, *192*, 103402.

Lin, Y., Cai, T., & Ju, C. (2020). Snow evaporation characteristics related to melting period in a forested continuous permafrost region. *Environmental Engineering & Management Journal (EEMJ)*, *19*(3).

Liu, C., Huang, X., Li, X., & Liang, T. (2020). MODIS fractional snow cover mapping using machine learning technology in a mountainous area. *Remote Sensing*, *12*(6), 962.

Mehdizadeh, S., Behmanesh, J., & Khalili, K. (2017). A comparison of monthly precipitation point estimates at 6 locations in Iran using integration of soft computing methods and GARCH time series model. *Journal of Hydrology*, *554*, 721-742.

Milly, P. C., & Dunne, K. A. (2020). Colorado River flow dwindles as warming-driven loss of reflective snow energizes evaporation. *Science*, *367*(6483), 1252-1255.

Nakhaei, M., Mohebbi Tafreshi, A., & Saadi, T. (2023). An evaluation of satellite precipitation downscaling models using machine learning algorithms in Hashtgerd Plain, Iran. *Modeling Earth Systems and Environment*, 1-15.

Niu, W.j., & Feng, Z.-k. (2021). Evaluating the performances of several artificial intelligence methods in forecasting daily streamflow time series for sustainable water resources management. *Sustainable Cities and Society*, *64*, 102562.

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, *15*(1), 12.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), 559-572.

Rezaverdinejad, V. (2016). Evaluation and Comparison of GRNN, MLP and RBF Neural Networks for Estimating Cucumber, Tomato and Reference Crops' Evapotranspiration in Greenhouse Condition. *Water and Soil Science*, *25*(4/2), 123-136.

Riedmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP

algorithm. IEEE international conference on neural networks,

Riggs, G. A., Hall, D. K., & Román, M. O. (2015). MODIS snow products collection 6 user guide. *National Snow & Ice Data Center*.

Saavedra, F. A., Kampf, S. K., Fassnacht, S. R., & Sibold, J. S. (2018). Changes in Andes snow cover from MODIS data, 2000–2016. *The Cryosphere*, *12*(3), 1027-1046.

Sahu, R. K., Müller, J., Park, J., Varadharajan, C., Arora, B., Faybishenko, B., & Agarwal, D. (2020). Impact of Input Feature Selection on Groundwater Level Prediction From a Multi-Layer Perceptron Neural Network. *Frontiers in Water*, *2*, 46.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *41*(2), 169-178.

Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, *5*(4), 2094-2097.

Song, F., Guo, Z., & Mei, D. (2010). Feature selection using principal component analysis. 2010 international conference on system science, engineering design and manufacturing informatization,

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Vapnik, V. (1998). Statistical learning theory Wiley. *New York*, *1*(624), 2.

Vapnik, V., & Lerner, A. Y. (1963). Recognition of patterns with help of generalized portraits. *Avtomat. i Telemekh*, *24*(6), 774-780.

Wang, W., Zhao, Y., Tu, Y., Dong, R., Ma, Q., & Liu, C. (2023). Research on Parameter Regionalization of Distributed Hydrological Model Based on Machine Learning. *Water*, *15*(3), 518.

Westra, S., Brown, C., Lall, U., & Sharma, A. (2007). Modeling multivariable hydrological series: Principal component analysis or independent component analysis? *Water Resources Research*, *43*(6).

Wu, Y., Duguay, C. R., & Xu, L. (2021). Assessment of machine learning classifiers for global lake ice cover mapping from MODIS TOA reflectance data. *Remote sensing of environment*, *253*, 112206.

Yakut, E., & Süzülmüş, S. (2020). Modelling monthly mean air temperature using artificial neural network, adaptive neuro-fuzzy inference system and support vector regression methods: A case of study for Turkey. *Network: Computation in Neural Systems*, *31*(1-4), 1-36.

Yegnanarayana, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.

Zhang, X., Wang, R., Yao, Z., & Liu, Z. (2020). Variations in glacier volume and snow cover and their impact on lake storage in the Paiku Co Basin, in the Central Himalayas. *Hydrological Processes*, *34*(8), 1920-1933.

Zhu, Q., Luo, Y., Zhou, D., Xu, Y.-P., Wang, G., & Tian, Y. (2021). Drought prediction using in situ and remote sensing products with SVM over the Xiang River Basin, China. *Natural Hazards*, *105*(2), 2161-2185.