


Comparison of different statistical downscaling models and future projection of areal mean precipitation of a river basin under climate change effect

Aytac Guven * and Abdulhadi Pala 

Civil Engineering Department, Gaziantep University, Gaziantep, Turkey

*Corresponding author. E-mail: aguven@gantep.edu.tr

 AG, 0000-0002-1935-1375; AP, 0000-0003-3090-2171

ABSTRACT

Investigation of the hydrological impacts of climate change at the local scale requires the use of a statistical downscaling technique. In order to use the output of a Global Circulation Model (GCM), a downscaling technique is used. In this study, statistical downscaling of monthly areal mean precipitation in the Gökşun River basin in Turkey was carried out using the Group Method of Data Handling (GMDH), Support Vector Machine (SVM) and Gene Expression Programming (GEP) techniques. Large-scale weather factors were used for the basin with a monthly areal mean precipitation (P_M) record from 1971 to 2000 used for training and testing periods. The R^2 -value for precipitation in the SVM, GEP and GMDH models are 0.62, 0.59, and 0.6 respectively, for the testing periods. The results show that SVM has the best model performance of the three proposed downscaling models, however, the GEP model has the lowest AIC value. The simulated results for the Canadian GCM3 (CGCM3) A1B and A2 scenarios show a similarity in their average precipitation prediction. Generally, both these scenarios anticipate a decrease in the average monthly precipitation during the simulated periods. Therefore, the results of the future projections show that mean precipitation might decrease during the period of 2021–2100.

Key words: artificial intelligence, climate change, GCMs, precipitation, statistical downscaling

HIGHLIGHTS

- An integrated hydrological model for prediction of areal mean precipitation of a river basin under climate change effect is proposed.
- Statistical downscaling and GCM are utilized to estimate the climate change effects on the basin.
- Gokşun River basin is used as the case-study area with three precipitation stations.
- Simulated results anticipate a decrease in the average monthly precipitation during the period of 2021–2100.

1. INTRODUCTION

Climate change and its impact, increasing concentration of ‘greenhouse gases’ in the atmosphere causing climate changes, have gained serious consideration in hydrology. General circulation models (GCMs) are analytical models representing physical processes in the atmosphere, oceans, ice and land surface, and are the primary tools that maintain reasonable accurate climate information at global scale, simulating the response of the global climate system to increasing greenhouse gas concentrations (Tofiq & Guven 2014, 2015).

GCM outputs cannot generate local climate details at finer spatial resolution due to the uncertainty in the spatial resolution between GCM and hydrological models (Shahid *et al.* 2013). Subsequently, downscaling methods have been developed to transform GCM outputs from coarse spatial resolution to a finer spatial resolution that can be directly used for forecasting climate change at the local scale (Hashmi *et al.* 2011). Based on its working principle, it can be broadly categorized as statistical downscaling and dynamic downscaling. Statistical downscaling is more widely used in hydrology studies because it requires less computational demand while dynamic downscaling requires high computational resources and expertise. Statistical downscaling establishes the empirical relationship between the large-scale climatic parameters such as mean sea level pressure, wind speed, zonal velocity (predictors) and local parameters (predictands) such as temperature, precipitation, and discharge (Chen *et al.* 2010a, 2010b).

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/>).

Statistical downscaling can be roughly divided into four categories: regression methods, weather pattern-based approaches, stochastic weather generators, and limited-area modeling (Wilby & Wigley 1997). A regression method is preferred among these approaches because it is easy to implement and it has low computation requirements. However, when the variable of interest is precipitation, the input-output relationship is often very complex and linear regression-based methods may not work very well. Therefore, a number of non-linear regression based downscaling techniques have been offered and successfully applied (Mpelasoka *et al.* 2001; Haylock *et al.* 2006). The scientific literature of the past decade includes predictions of runoff based on precipitation and rainfall-runoff models using statistical downscaling and different GCM scenarios (Schmidli *et al.* 2007; Yonggang *et al.* 2007; Chen *et al.* 2012).

The Group Method of Data Handling (GMDH) method was developed by A. G. Ivakhnenko (Ivakhnenko 1971). GMDH is described as a nonparametric (self-organizing) learning algorithm. This means that it makes no prior assumptions about the 'true' form of the model that it derives. It is a learning algorithm in the sense that it uses a 'hill-climbing' approach and remembers its successes during a systematic trial and error search (Green *et al.* 1988). It is also called a 'polynomial model' in which the predicted value of the output will be as close as possible to the actual value of output by constructing successive layers with simple polynomial terms such as polynomial, harmonic, square root, inverse polynomial, logarithmic, exponential, rounded polynomial, etc. (Muttli & Liong 2012). These polynomials with two variables can be expressed as:

$$y = p_0 + p_1x_1 + p_2x_2 + p_3x_1^2 + p_4x_2^2 + p_5x_1x_2 \quad (1)$$

where, p_i is the model coefficient, x_i is the base function.

The GMDH model automatically selects the most accurate input variables and the optimal network structure in order to minimize the difference between the structure output and desired output. GMDH is therefore generalizable and can be adapted to the complexity of non-linear systems with a relatively simple and numerically stable network (Assaleh *et al.* 2013). The advantage of the GMDH algorithm is that it is reasonable for a system that has a large input dataset. There are a lot of applications of GMDH in environmental, ecological, and social systems (Duffy & Franklin 1975; Josef & Petr 2011). GMDH has become a popular method used in water resources studies and has been successfully applied in the prediction of hydro meteorological variables (Saburo *et al.* 1976; Muttli & Liong 2012; Zahraie *et al.* 2017).

Gene Expression Programming (GEP) is a new variant of genetic programming (GP) that was developed by Ferreira (Ferreira 2001). GEP is a genotype/phenotype system that evolves computer programs of different sizes and shapes (the phenotype) encoded in linear chromosomes of fixed length (the genotype). GEP is an evolutionary algorithm, using symbolic regression to fit the data to get an optimum form of mathematical function. GEP automatically determines the statements for the study solution by coding the expression as a tree formation with nodes and terminals. The link between symbols of function and the chromosome is represented in the tree by a one on one relation. Combination of genes or one gene create the chromosome and each gene produces a smaller sub-program. The linking function is utilized to connect to the genes when the chromosome has more than one gene (Ferreira 2001). GEP has been widely applied in hydrologic engineering, particularly in the last decade. It has especially been applied in the prediction of hydrometeorological variables (Güven 2009; Güven & AYTEK 2009; Güven & Talu 2010; Seckin & Güven 2012; Traore & Güven 2013; Al-Juboori & Güven 2016).

The support vector machine (SVM) technique was introduced by Vapnik in 1995, and gained popularity in data-driven study fields because of its many alluring features and promising empirical performance (Vapnik 1995). SVM is essentially a classification and regression method that has been obtained from statistical learning theories. The regression version of the SVM framework follows the principle of Structural Risk Reduction (SRM), which is proven to be superior to the traditional empirical risk reduction (ERM) principle used by traditional modeling techniques (Vapnik 1999). Traditional ERM concentrates on minimizing training data error, whereas SRM minimizes the upper limit of expected risk, hence providing SVM with the ability to generalize and this is the essential aim of statistical learning (Vapnik 1999). There are four fundamental advantages of SVM. Firstly, it has a regularization parameter, which allows the user to reduce or avoid problems associated with over-fitting. Secondly, SVM is decided by a convex optimization problem for which there are accurate methods (e.g., sequential minimal optimization (SMO)). Thirdly, it is estimated to be bound by the test error rate, and there is a significant body of theory behind it which suggests it should be a good idea. The last and main advantage of SVM is that it utilizes the kernel trick (function) to construct expert knowledge about the investigated phenomenon so that the model complexity together with estimation error is concurrently minimized (Karimi *et al.* 2017). SVM has been used during the last decade for rainfall-runoff modeling. Many SVM-based models related to the field of hydrology were

investigated by Raghavendra & Deka 2014 who concluded that regression models were ineffective compared to SVM models in many cases as analyzed in the literature. The statistical downscaling of daily precipitation using SVM and multivariate analysis has been investigated. The SVM model results were compared with a statistical downscaling model (SDSM) and it was found that the statistical performance of the SVM model is better than the SDSM in predicting daily precipitation (Chen *et al.* 2010a, 2010b). Daneshfaraz *et al.* (2021a, 2021b) investigated the application of SVM for predicting hydraulic parameters of a vertical drop equipped with horizontal screens and the results showed that the application of SVM performs high R^2 value (0.991 for testing and training modes). Daneshfaraz *et al.* (2021a, 2021b) also examined the application of SVM for estimating vertical drop hydraulic parameters in the presence of dual horizontal screens. The results show that this method can accurately predict the hydraulic performance of these systems.

GMDH, GEP and SVM techniques have been applied separately and results compared. An inclusive multiple modelling (IMM) strategy can be used to decrease residual errors. Sadeghfam *et al.* (2021) investigated the hydrological impact of climate change in terms of downscaling of monthly precipitation by producing an IMM approach. IMM strategies handle multiple models at two levels. The model at Level 2 merges outputs of those at Level 1 and produces Level 2 results, which are enhanced compared with those of the Level 1 model in terms of dispersion of residual errors. In this way, IMM maintains a more defensible modelling approach for application in the projection stage.

In this study, GMDH, GEP and SVM methods are applied to develop areal mean precipitation downscaling models and to find the future change pattern for precipitation. Finally, a multiple linear regression (LR) model is analyzed for comparison. The novelty of the paper is contained in the application of three different downscaling techniques (GEP, GMDH, and SVM), comparison of those techniques according to statistical indicators, and projection of downscaling techniques under CGCM3 A1B and A2 emission scenarios.

2. STUDY AREA AND DATA USED

The study area was the Göksun River basin which is a sub-basin of the Ceyhan River basin. This basin area is located 100 km from Kahramanmaraş city on the boundary of the Göksun region on the eastern Mediterranean. Figure 1 shows the Digital Elevation Model (DEM) of Göksun River Basin. This basin surrounds an area of 2,307 km², and its elevation ranges from 1,170 m to 2,800 m a.s.l. There are three local precipitation gauge stations in or near the study area. Summary of the rainfall data used in this study for each station is given in Table 1.

The mean precipitation data of these three stations were converted to the first dataset (predictand set) used in this study, the areal mean precipitation (PM) of the basin, which were calculated by using the Thiessen Polygons method. Figure 2 shows the Thiessen Polygons of the basin and location of the precipitation gauge stations. Different coloured lines represent each Thiessen polygon line to find the fraction of each station in order to determine areal mean precipitation. The centre of the basin is located at latitude 38° 6', longitude 36° 48'. The second dataset is the large-scale predictor variables data for grid box (Box_ 11X_14Y) representing the study area, obtained from the Canadian Global Climate Model (CGCM) which is available at their website (www.cccsn.ec.gc.ca/index.php). Each grid cell represents a mesh surrounding the corresponding model grid points. The dimension of each grid cell is approximately 3.75° latitude and 3.75° longitude (Gaussian grid). This process takes place by input decimal latitudinal and longitudinal coordinates of any location, and in this study the coordinate of the centroid of the basin is used. The CGCM3 variables are used as predictors in this study, in as much as they are broadly used in several climate changes and downscaling studies. Table 2 utilizes the description of CGCM3 variables which contains 26 variable data sets. These data sets consist of 26 variables for each scenario (Tofiq & Guven 2014, 2015; Singh *et al.* 2015).

Figure 3 shows the location of Göksun basin and precipitation stations. Figure 4 shows a map of Turkey with the Göksun basin centroid.

3. METHODOLOGY

3.1. Selection of the most effective predictors

Predictor selection is one of the most critical steps in the climate downscaling process. In this study, the correlation between the predictors and predictands was evaluated. To improve the correlation with predictor variables and predictands, the predictands were modified based on a number of normalization methods. These modifications are: taking natural logarithm (Ln), min-max normalization (Mn) which normalizes the data to the range from 0 to 1, and standardization (Stand) of the predictand variables. To select the most effective predictors and to see the linear relationship between inputs and outputs,

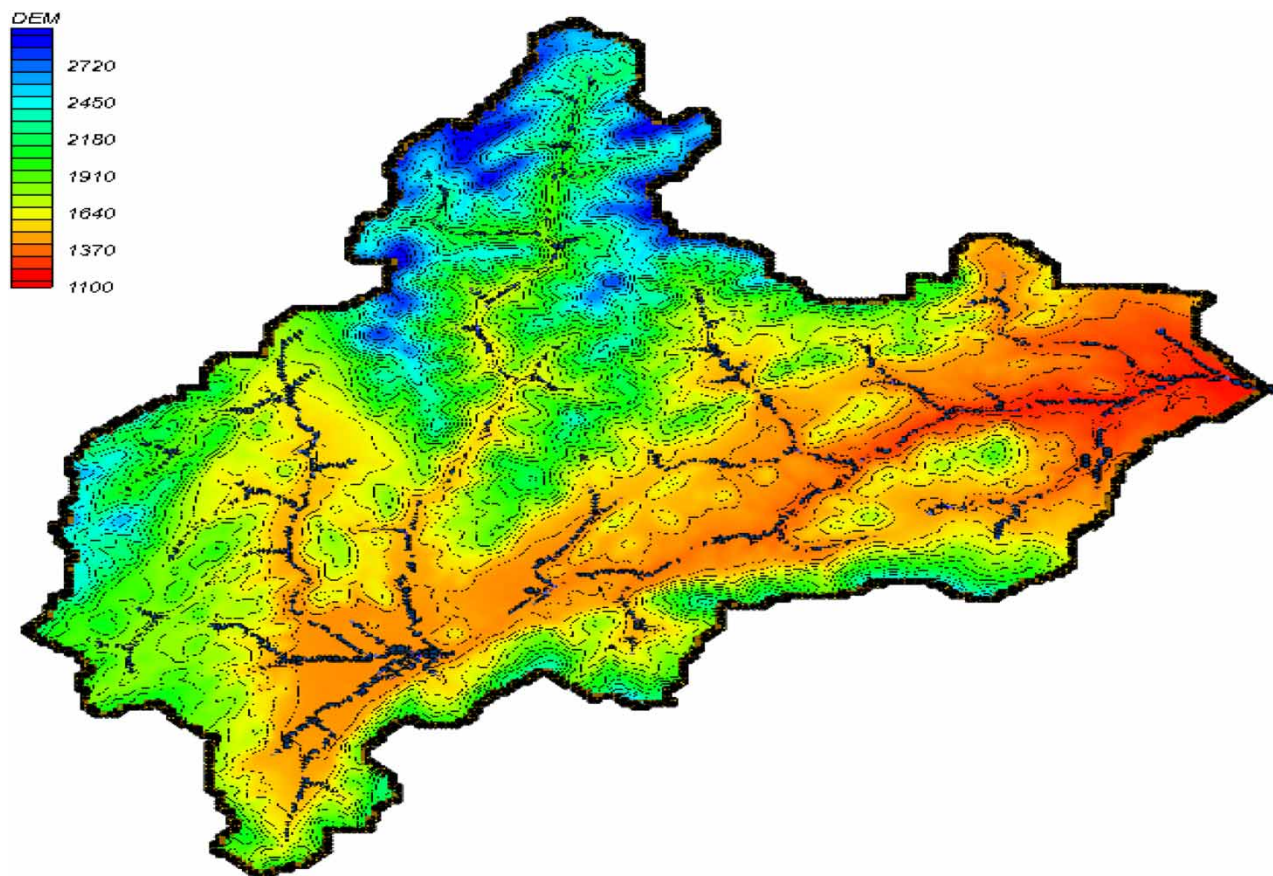


Figure 1 | Digital Elevation Model (DEM) of Göksun River Basin.

Table 1 | Statistical summary of the data used in this study for each station

Item	Afsin	Göksun	Elbistan
Mean Annual Rainfall (mm)	426.50	597.20	385.88
Mean Monthly Rainfall (mm)	35.54	49.77	32.16
Maximum Annual Rainfall (mm)	560.30	869.40	548.50
Minimum Annual Rainfall (mm)	273.80	354.20	247.40
Maximum Daily Rainfall (mm)	65.50	85.40	46.10
Duration of Data (year)	41	42	42
Number of Data set (days)	14,965	15,330	15,330

correlation analysis was undertaken using the Pearson rank correlation coefficient method. Results illustrated that most large-scale weather factors were statically correlated with local precipitation under a confidence level of 99%. Factors with a higher correlation coefficient were selected as the predictors for downscaling models.

The set of predictors was divided into three categories according to the value of the correlation coefficient (R) between the predictand and the predictors. These categories are: $R > 0.3$, $R > 0.4$ and $R > 0.5$. Predictors with a correlation of $R > 0.5$ were used for natural logarithm normalization of predictor and predictand variables. The highest correlating GCM variables, with $R > 0.5$ (10 of them in this case), were selected and presented in [Table 3](#).

Those with input correlation bigger than 0.5 were Mean sea level pressure (V1), 1,000 hPa Meridional velocity (V4), 1,000 hPa Divergence (V7), 850 hPa Wind speed (V15), 850 hPa Meridional velocity (V16), 850 hPa Meridional velocity (V17), 850

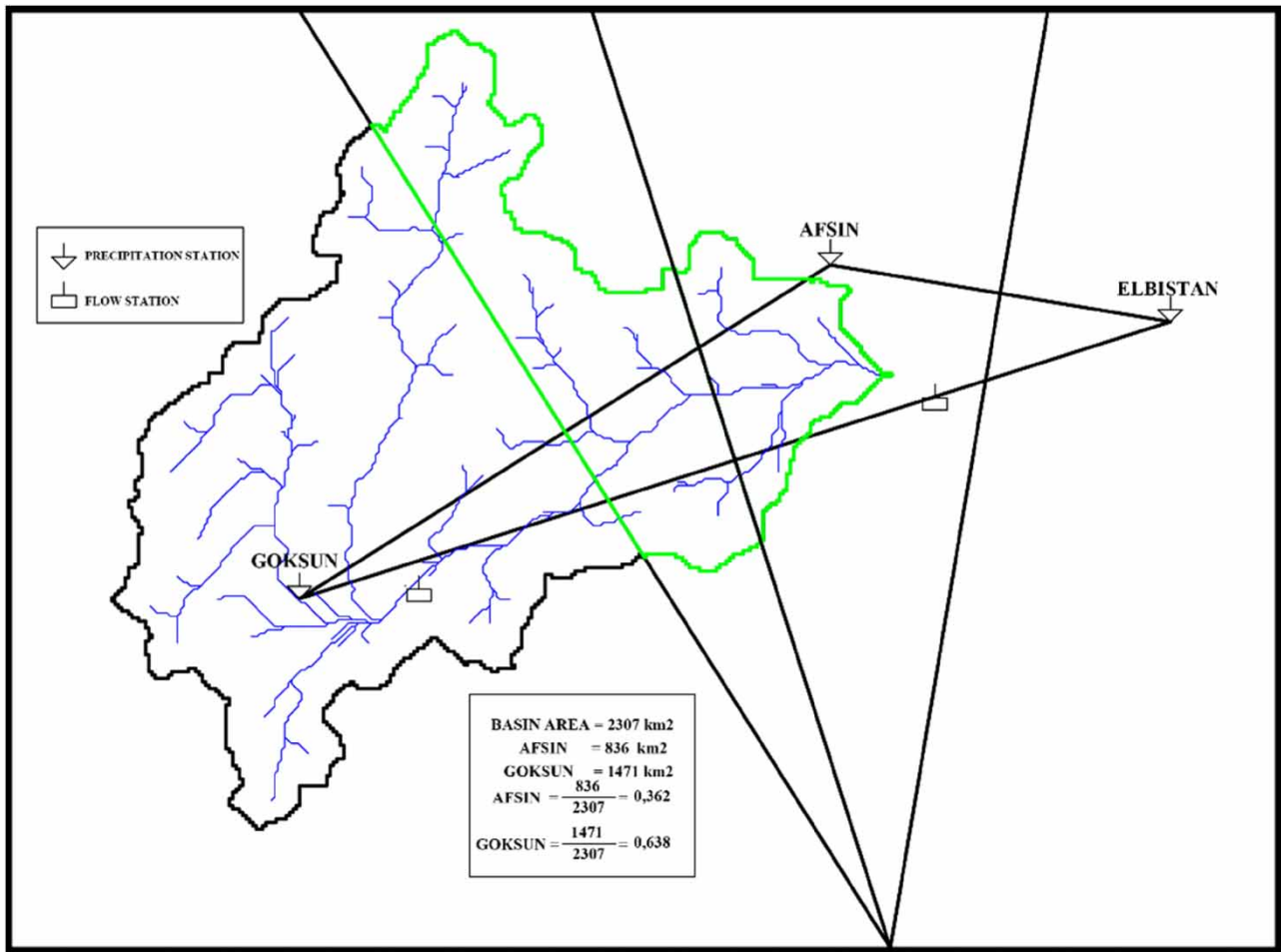


Figure 2 | Thiessen polygons of the basin and location of precipitation gauge stations.

hPa Vorticity (V18), 850 hPa Geopotential (V19), 850 hPa Wind direction (V20), 1,000 hPa Specific humidity (V25), Screen air temperature (2 m) (V26).

3.2. Downscaling using GMDH, GEP and SVM techniques

There are several steps to solve a problem using the GMDH and GEP programs (Mpelasoka *et al.* 2001; Haylock *et al.* 2006).

There are four crucial steps to get the solution to the problem in GEP. In this study, the first critical step is to determine a set of functions. A set of functions (+, -, *, /, power, Ln) were selected from the set of functions in the GEP program. The second critical step for the GEP program is to determine the chromosome structure which includes the number of genes per chromosome and the size of the gene. The best model results were obtained by using four genes per chromosome for each GEP model. The third critical step is to select the linking function. In this study, the linking function selected was '+' (addition). The final critical step is the fitness measure. In this study, the root relative squared error (RRSE) of the training set has been employed as a fitness function.

The GEP model also developed an equation, given in Equation (2), for the calibration process to generate the validation process and to predict future projections. The definition of each variable is given in Table 3.

$$\text{LnP}_M = \left(\frac{\left(V1 \times \left(\left(\frac{V26}{V15} \right) + \left(\frac{V4}{V20} \right) \right) \right)}{V18} \right) + 2.4824 + \left(\frac{((V1 - (V26 + 2.4924)) \times V26)}{V18} \right) + (0.0365 \times (27.383 - V26)) \quad (2)$$

Table 2 | Description of the 26 CGCM3 variables (predictors)

NO	CGCM3 Predictors	Description	Short name
1	c3a2mslpgl	Mean sea level pressure	v1
2	c3a2p_fgl	1,000 hPa Wind speed	v2
3	c3a2p_ugl	1,000 hPa Zonal velocity	v3
4	c3a2p_vgl	1,000 hPa Meridional velocity	v4
5	c3a2p_zgl	1,000 hPa Vorticity	v5
6	c3a2p_thgl	1,000 hPa Wind direction	v6
7	c3a2p_zhgl	1,000 hPa Divergence	v7
8	c3a2p5_fgl	500 hPa Wind speed	v8
9	c3a2p5_ugl	500 hPa Zonal velocity	v9
10	c3a2p5_vgl	500 hPa Meridional velocity	v10
11	c3a2p5_zgl	500 hPa Vorticity	v11
12	c3a2p5thgl	500 hPa Geopotential	v12
13	c3a2p5zhgl	500 hPa Wind direction	v13
14	c3a2p8_fgl	500 hPa Divergence	v14
15	c3a2p8_ugl	850 hPa Wind speed	v15
16	c3a2p8_vgl	850 hPa Zonal velocity	v16
17	c3a2p8_zgl	850 hPa Meridional velocity	v17
18	c3a2p8thgl	850 hPa Vorticity	v18
19	c3a2p8zhgl	850 hPa Geopotential	v19
20	c3a2p500gl	850 hPa Wind direction	v20
21	c3a2p850gl	850 hPa Divergence	v21
22	c3a2prcpgl	Accumulated precipitation	v22
23	c3a2s500gl	500 hPa Specific humidity	v23
24	c3a2s850gl	850 hPa Specific humidity	v24
25	c3a2shumgl	1,000 hPa Specific humidity	v25
26	c3a2tempgl	Screen air temperature (2 m)	v26

The main step is to determine the set of network parameters in order to obtain a solution from the GMDH model. In this study, the maximum network layer is set to 15, the maximum polynomial order is 12, and the convergence tolerance factor 0.001 is used for the GMDH model. Network layer connection is selected as the previous layer with original input variables. The number of neurons per layer is fixed as 15 neurons. The final main step is to decide suitable functions that will be used in the GMDH network. We got the best results by using Linear 1, 2, and 3 variables and Quadratic 1, and 2 variables.

Epsilon Support Vector Regression (SVR) was used to predict monthly mean precipitation for the SVM model. Radial basis function was used as a kernel function. Parameter optimization search control is chosen to get optimal parameter for SVM models. Then a grid search and pattern search is done. The SVM model has three parameters: C , γ , and ϵ . Table 4 shows parameters which are acquired via the SVM to predict the monthly areal mean precipitation.

The statistical measures used to evaluate the accuracy performance of GMDH, GEP and SVM models and LR models are the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE).

4. RESULTS AND DISCUSSION

Large-scale weather factors (26 input sets) were used as predictors which were obtained from the Third Generation Coupled Global Circulation Model (CGCM3). The areal mean precipitation of the basin (P_M) was used as predictand (output). All data sets are based on a monthly time scale. Both data sets were divided into two subsets (calibration and

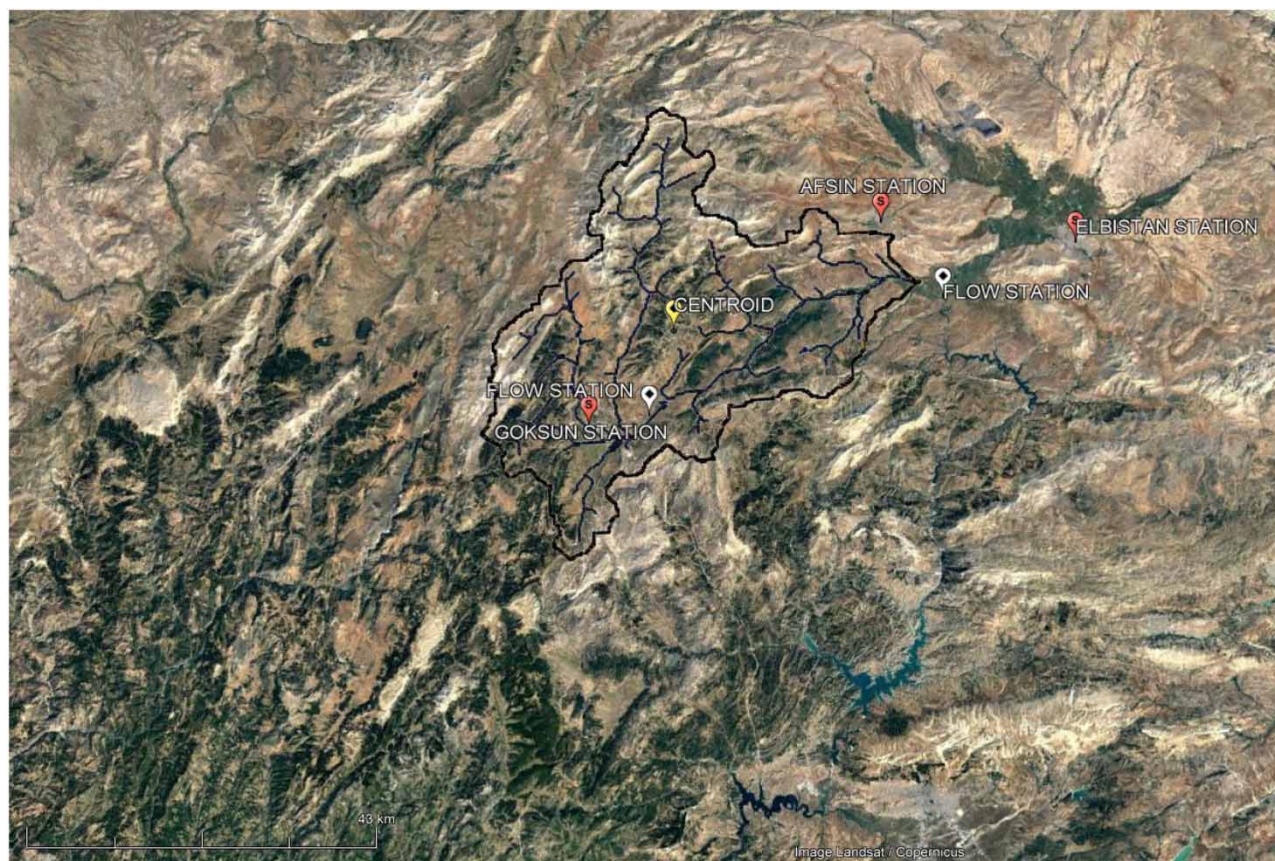


Figure 3 | Location of the Göksun basin and location of precipitation gauge stations.

validation). The first one is the calibration period ranging between 1971 and 1990. The second one is the validation period ranging between 1991 and 2000. GMDH, GEP, and SVM downscaling techniques were utilized to predict the downscaled P_M of the basin. Furthermore, the prediction results of downscaling techniques were compared with each other to examine the best model performance. Also, a multiple linear regression model was used for comparison with the proposed nonlinear models.

The statistical results of the best model of monthly areal mean precipitation P_M in the Göksun River Basin for the validation period are shown in Table 5. It is seen in Table 5 that the simulated mean monthly areal mean precipitation by SVM, GEP, GMDH models, and LR are lower than the observed mean P_M , numerically 10, 10.9, 11.6, and 14.5 mm respectively. This indicates that the means were underestimated by all models. But, the SVM model yielded better results than the other models to simulate the mean P_M . The SVM and GMDH models perform well in predicting the minimum P_M while LR performs poorly to simulate the minimum P_M . Moreover, there is a significant difference in the maximum value of simulated P_M compared to observed data. The GMDH model records the minimum difference of 9% for maximum P_M , and the SVM model records the maximum difference of 41% for maximum P_M . As for the mean, all models underestimated the standard deviation, but the SVM model had the lowest underestimating difference of 36% of all the models.

As seen by the values of statistical indicators given in Table 4, the MAEs for the SVM model (22.26 mm) and the RMSE (32.36 mm) were lower than for the other proposed models. The MAEs for the GEP, GMDH, and LR models were approximately 8%, 8.8% and 19.5% greater than that of the SVM, respectively. The RMSE is approximately 5.5%, 7.5%, and 19.5% greater for the GEP, GMDH, and LR models, respectively, than for the SVM.

Figure 5 represents the scatter plot of the observed P_M versus the predicted ones of the Göksun River Basin for GMDH, GEP, SVM and LR for the validation period. The SVM ($R^2 = 0.62$) and GMDH ($R^2 = 0.61$) models have almost the same coefficient of determination (R^2) and the GEP model's R^2 (0.59) is close to them. However, the SVM model has the highest R^2 while the LR model has the lowest R^2 (0.53).



Figure 4 | Map of Turkey with centroid of Göksun basin.

Table 3 | Most effective GCM set predictors

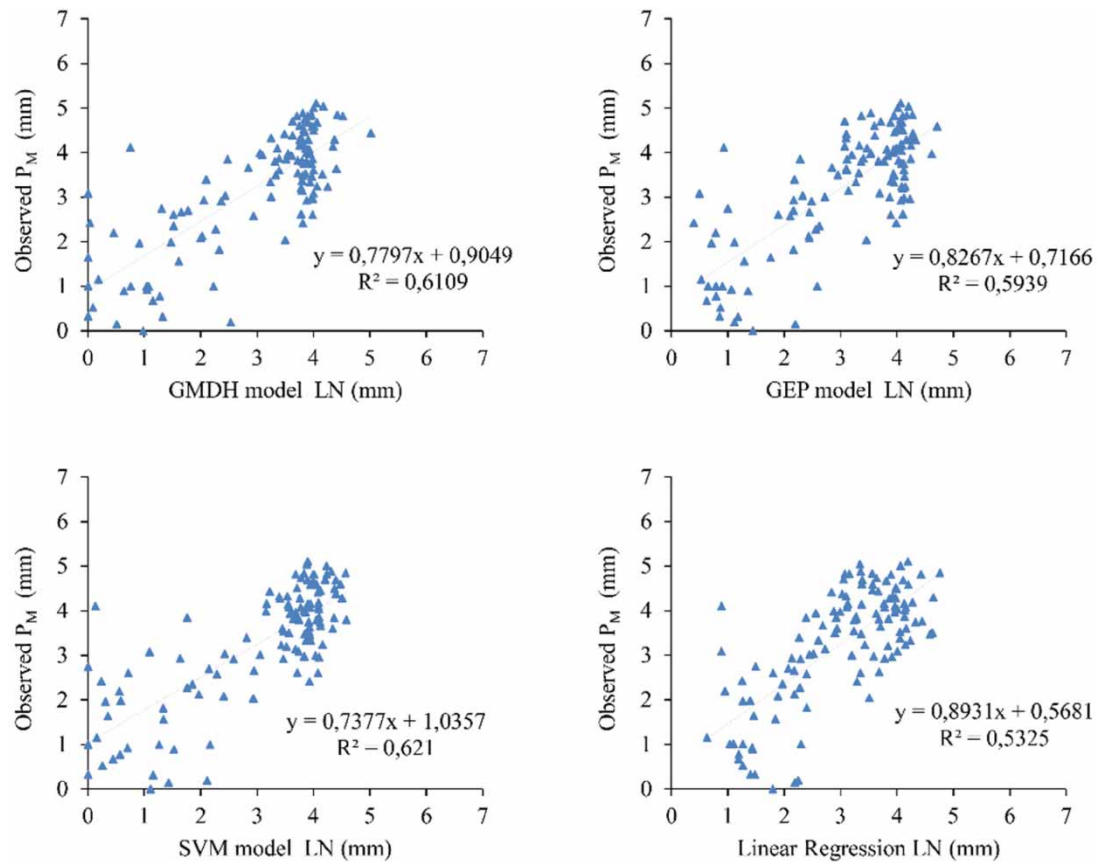
No	GCM Predictors	Description	R	No	GCM Predictors	Description	R
V1	c3a2mslpgl	Mean sea level pressure	0.46	V14	c3a2p8thgl	850 hPa Vorticity	-0.55
V2	c3a2p_vgl	1,000 hPa Meridional velocity	0.53	V15	c3a2p8zhgl	850 hPa Geopotential	-0.58
V3	c3a2p_zhgl	1,000 hPa Divergence	-0.59	V16	c3a2p500gl	850 hPa Wind direction	-0.65
V4	c3a2p8_ugl	850 hPa Wind speed	0.54	V17	c3a2shumgl	1,000 hPa Specific humidity	-0.67
V5	c3a2p8_vgl	850 hPa Zonal velocity	0.57	V18	c3a2tempgl	Screen air temperature (2 m)	-0.68

Table 4 | SVM model parameters

c	γ	ϵ
9.03	3.39	0.01

Table 5 | Comparison of statistical results of observed and simulated P_M during the validation period (1991–2000)

Data	Mean (mm)	Std. Deviation (mm)	MAE (mm)	RMSE (mm)	R^2
Observed	45.94	40.53	–	–	–
GMDH Model	34.37	25.25	24.23	34.14	0.61
GEP Model	35.07	25.59	24.05	34.78	0.59
SVM Model	35.94	25.94	22.26	32.36	0.62
Linear Regression (LR)	31.43	25.86	26.60	38.67	0.53

**Figure 5** | Scatter plot of the observed versus predicted P_M for the Gökşun River Basin for GMDH, GEP, SVM and LR for the validation period.

The linear equations comprising the relation between the observed and the predicted P_M for the four models (GMDH, GEP, SVM, LR) were obtained to be:

$$y = 0.780x + 0.905 \quad (\text{GMDH model}) \quad (3)$$

$$y = 0.827x + 0.717 \quad (\text{GEP model}) \quad (4)$$

$$y = 0.738x + 1.036 \quad (\text{SVM model}) \quad (5)$$

$$y = 0.893x + 0.568 \quad (\text{LR model}) \quad (6)$$

where y represents the observed P_M and x represents the predicted P_M .

Some of the predicted P_M values of the proposed downscaling models were observed to be negative values. Under this situation, those negative values were adjusted to zero. The percentage of negative values over the total number of monthly areal mean precipitations were 3.32% in the GMDH, 2.49% in the GEP, 0.83% in the SVM, and 10% in LR models. LR has the biggest percentage correction of negative values adjusted to zero, while GMDH, GEP, and SVM needed less correction. These corrections affect the statistical properties of the models and cause the tendency to accuracy of the predicted value when the observed value is equal to 0.

Also, the Akaike Information Criterion (AIC) values of each proposed models were calculated and given in Table 5. AIC, introduced by Akaike (1974), is employed to evaluate the generalization capacity of the proposed models.

$$AIC = N \times \ln(MSE) + 2k \tag{7}$$

where N is the number of data, MSE is the mean square error, and k is the number of fitting parameters. AIC is used to evaluate the exchange between calibration performance and network size. The aim is to get a smaller AIC to acquire a network with the best generalization. The number of data used in the validation period is 120. It can be seen that AIC increases when the number of fitting parameters (k) increases. Table 6 indicates that the SVM model has the lowest MSE value (1047.3 mm^2), however the SVM model has the biggest AIC value (3,078) because of having a high number of fitting parameters. Consequently, the lowest AIC (860) and k (4) values of the GEP model present its best robustness and generalization capability.

Figure 6 displays the simulated averaged monthly mean precipitation, P_M , of GMDH, GEP, SVM, and LR models and the observed data during the validation period. It is clear that the SVM model has slightly underestimated the P_M for each month. In January, the LR model overestimated the P_M while other models underestimated. In February, the LR model and the GEP model overestimated the P_M while other models underestimated it. In July and October, the GMDH model overestimated the P_M while other models underestimated it.

In March, April, May, Jun, August, September, and October all models underestimated the P_M . In April, May, and June SVM predicted the best averaged mean precipitation compared to the observed data. The mean underestimating percentage

Table 6 | AIC values of proposed models during validation (1991–2000)

Model	MSE (mm ²)	k	AIC
GMDH	1,165.7	490	1,827
GEP	1,209.7	4	860
SVM	1,047.3	1,122	3,078
LR	1,495.1	11	899

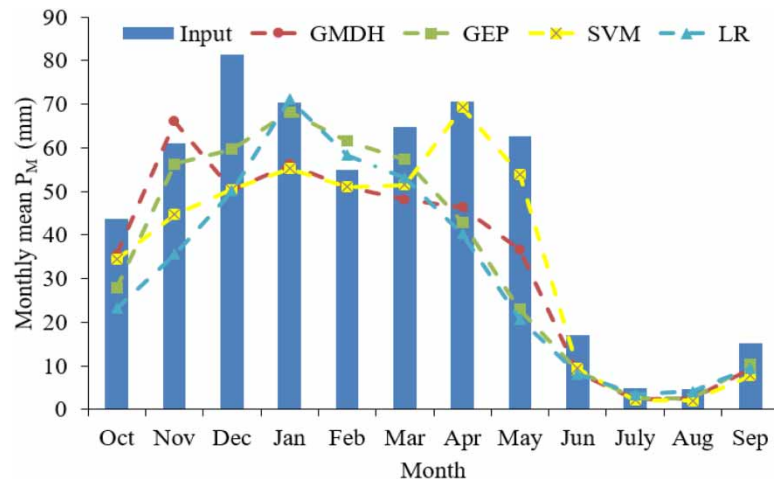


Figure 6 | Observed and predicted averaged monthly P_M for the validation period 1991–2000 by all models.

in April are 33%, 38%, and 41.8% smaller for the GMDH, GEP, and LR models, respectively, than that of the SVM model, as shown in Figure 3. In March, September, November, and December the GEP model performs best in predicting P_M . The mean underestimated percentages in March are 7.5%, 10.4%, and 16.2% smaller for the GMDH model, LR and SVM model, respectively, than that of the GEP model. In July and October, the GMDH model performs best in predicting P_M . The mean underestimating percentage in October are 3%, 21.4%, and 34.6% smaller for the SVM model, the GEP model, and LR model, respectively, than that of the GMDH model. In January, February, July, and August, the LR model performs best in predicting P_M . In February, SVM and GMDH models predicted the same value of P_M .

Consequently, the SVM model outperformed over the other models in the validation period with the highest coefficient of determination (R^2) and least RMSE and MAE values. However, GEP has the lower AIC values than the other models.

5. FUTURE PROJECTION OF P_M UNDER DIFFERENT EMISSION SCENARIOS AND DOWNSCALING MODELS

The downscaling performance of the GCM scale simulated under both CGCM3 A1B and A2 emission scenarios for P_M for the study area has been explored and analyzed. These scenarios are defined by IPCC (2007). A1B scenario describes the future of the world with very rapid economic growth, a world population that peaks in the middle of the century and then declines, and the rapid introduction of new and more efficient technologies and technological focus is balanced between all energy sources. A2 scenario describes a very diverse world, with a growing global population and regional economic growth. This scenario is more disconnected and slower than other scenarios. It also represents a high emissions scenario and illustrates the worst case scenario (Andersen *et al.* 2006).

CGCM3A2 and CGCM3A1B variables were utilized to predict the pattern of future change for the period of 2021–2100 for the P_M in the study area. The downscaled results from both scenarios (A1B and A2) were divided into four time periods with 20 years range, namely: 2020s (2021–2040), 2040s (2041–2060), 2060s (2061–2080), and 2080s (2081–2100), and compared with the baseline period (1971–2000) to examine the future change in P_M for the basin area.

Figure 7 displays the projection of averaged monthly P_M change under CGCM3A1B scenario for different time periods by the SVM model. From Figure 7, it is clear that the SVM model projected that the amount of precipitation for each different time periods will be decreased compared to baseline P_M . A noticeable decrease occurs in May and it projected a decrease in the averaged monthly P_M of about 18%, 31%, 35.5%, and 48.2% for the 2020s, 2040s, 2060s, and 2080s, respectively, under the A1B scenario. From Figure 7, it is observed that a constant decrease in P_M occurs during the four periods (2020s, 2040s, 2060s, and 2080s) for May, June, August, and September. From November to April, it is observed that a constant decrease in P_M does not occur during the four periods (2020s, 2040s, 2060s, and 2080s). In April, there is a increase in P_M from 2040s to 2060s. In February, it is observed that there is a noticeable increase in P_M after 2080s.

Figure 8 shows the projection of averaged monthly P_M change under CGCM3A2 scenario for different time periods by the SVM model. From Figure 8, it is clear that the SVM model projected the amount of precipitation for each different time period will decrease when compared with baseline mean precipitation. However, in April the projected mean precipitation

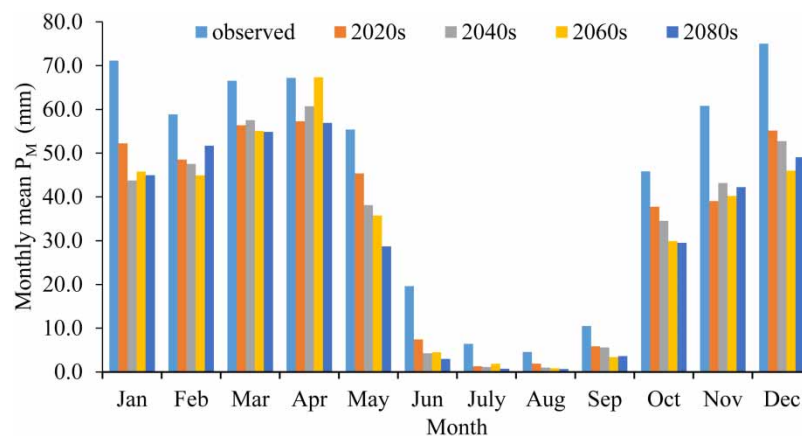


Figure 7 | The projection of averaged monthly P_M change under CGCM3A1B scenario for different time periods by the SVM model.

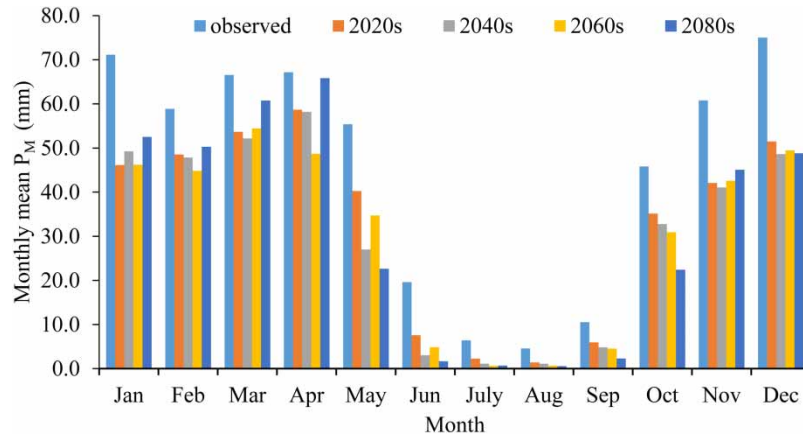


Figure 8 | The projection of averaged monthly P_M change under CGCM3A2 scenario for different time periods by the SVM model.

is close to the observed baseline mean precipitation. A noticeable decrease occurs in May and a decrease is projected in the averaged monthly P_M of about 27.3%, 51.3%, 37.3%, and 59.1% for the 2020s, 2040s, 2060s, and 2080s, respectively, under the A2 scenario. In Figure 8, it is observed that a constant decrease in P_M does not occur during the four periods (2020s, 2040s, 2060s, and 2080s). It is seen that a tendency to decrease in P_M occurs from May to October. It is observed that there is a noticeable increase in P_M after 2080s from January to April.

The changes in max precipitation predicted by both scenarios are different from each other in magnitude, but almost similar in their patterns as seen in Figure 7 and 8. Both scenarios show an average annual decrement with respect to the baseline period (1971–2000) in these figures. Under the A1B scenario, the SVM model predicts a decrease in the averaged annual P_M by 24.7%, 28%, 30.7%, and 32.5% for the 2020s, 2040s, 2060s, and 2080s, respectively. It also projects a decrease in the averaged annual P_M of about 27.5%, 32.3%, 33.1%, and 31.1% for the 2020s, 2040s, 2060s, and 2080s, respectively, under the A2 scenario.

Figure 9 displays the projection of averaged monthly P_M change under CGCM3A1B scenario for different time periods by the GMDH model.

Figure 10 displays the projection of averaged monthly P_M change under CGCM3A2 scenario for different time periods by the GMDH model.

Figure 11 displays the projection of averaged monthly P_M change under CGCM3A1B scenario for different time periods by the GEP model.

Figure 12 displays the projection of averaged monthly P_M change under CGCM3A2 scenario for different time periods by the GEP model.

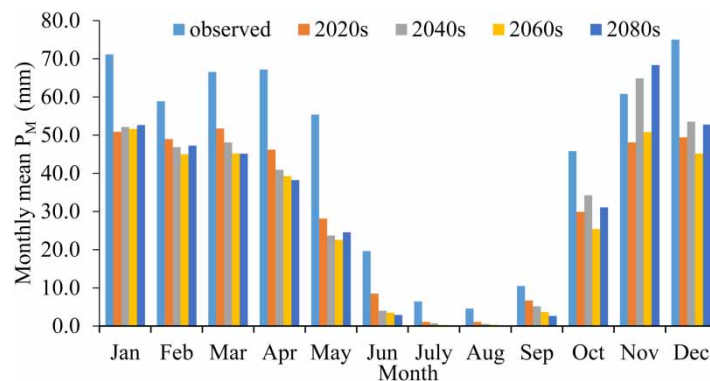


Figure 9 | The projection of averaged monthly P_M change under CGCM3A1B scenario for different time periods by the GMDH model.

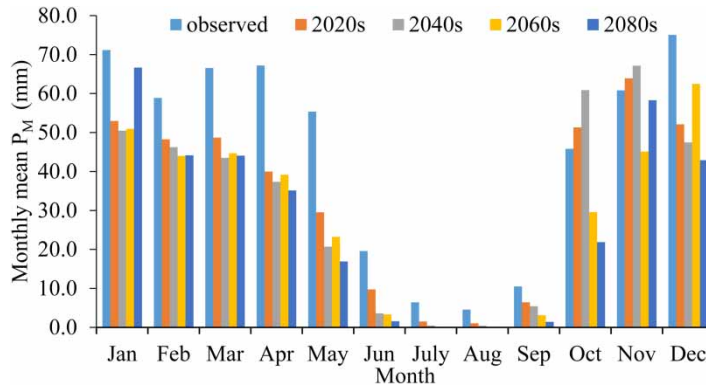


Figure 10 | The projection of averaged monthly P_M change under CGCM3A2 scenario for different time periods by the GMDH model.

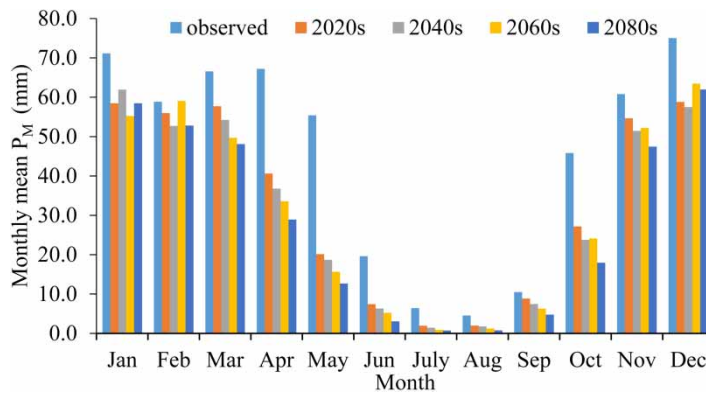


Figure 11 | The projection of averaged monthly P_M change under CGCM3A1B scenario for different time periods by the GEP model.

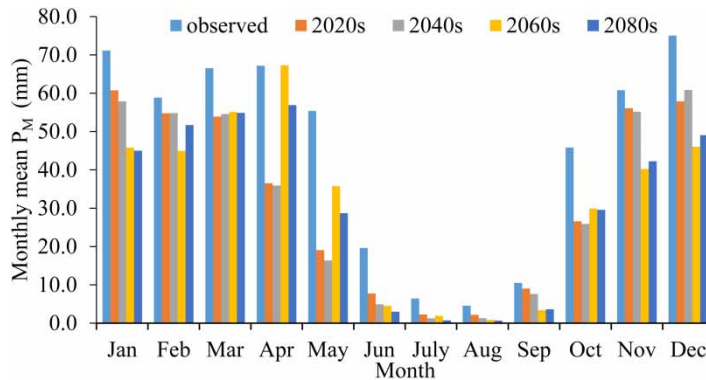


Figure 12 | The projection of averaged monthly P_M change under CGCM3A2 scenario for different time periods by the GEP model.

Figures 13 and 14 represent the projected annual P_M under A1B and A2 scenarios for different time periods by the proposed downscaled models. Figure 6 shows that GMDH model forecasted the lowest annual precipitation, and SVM model projected the highest one. Hollow symbols in Figure 13 and 14 represent the mean annual precipitation downscaled by the models in the baseline period 1971–2000. Figure 14 exhibits that the GMDH model projected the highest annual precipitation, and SVM projected the lowest from 2021–2049 to 2055–2084 periods while the SVM model projected the highest annual precipitation, and GMDH projected the lowest from 2055–2084 to 2070–2099 periods.

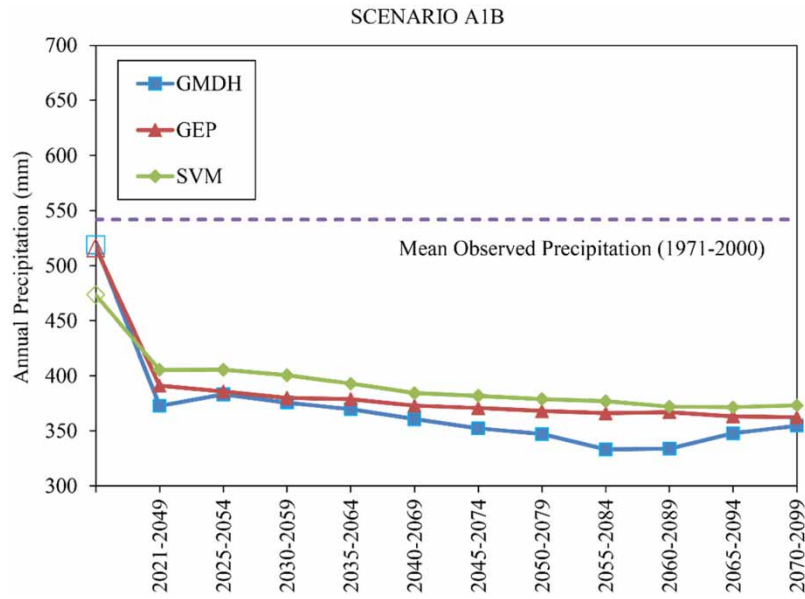


Figure 13 | The projection of annual PM change under CGCM3A1B scenario for different time periods by the SVM, GMDH, and GEP models.

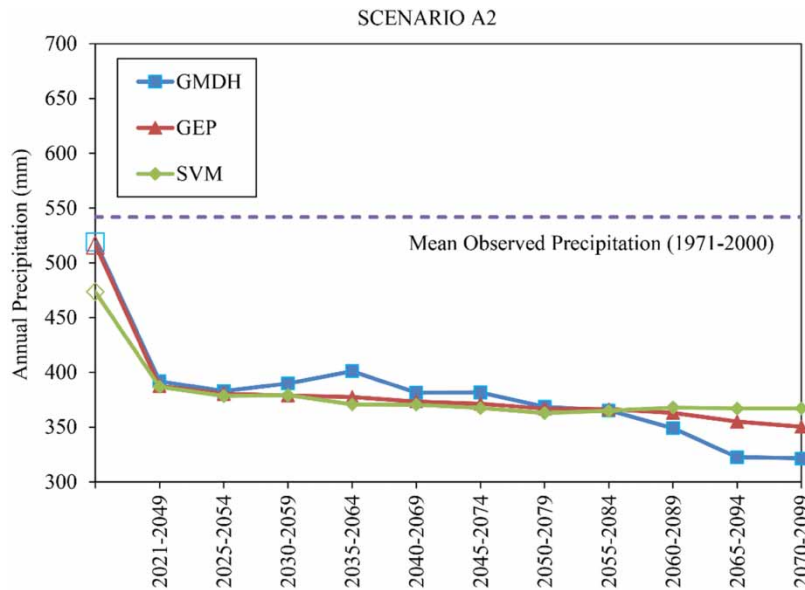


Figure 14 | The projection of annual PM change under CGCM3A2 scenario for different time periods by the SVM, GMDH, and GEP model.

6. CONCLUSION

The purpose of this study was to evaluate the performance of different artificial intelligence techniques in the statistical downscaling process. GMDH and GEP select the most accurate predictor variables automatically, while SVM uses the whole input set. Statistical downscaling of precipitation is very difficult, as the relation between the GCM variables and the local variables are often complicated. In this study, statistical downscaling of areal mean precipitation of a basin was carried out by using GMDH, GEP and SVM techniques comparing the results of LR. Results show that the SVM model outperformed the other models in the validation period with the highest coefficient of determination (R^2) and the lowest RMSE and MAE

values. However, GEP has lower AIC values than the other models. This proves the highest generalization capacity of the GEP model.

The downscaling performance of these techniques under both CGCM A1B and A2 emission scenarios for the areal mean precipitation of the study area was also analyzed. For future projection of different scenarios, all methods performed in a similar manner in their mean precipitation. Likewise, both scenarios forecast a decrease in the average monthly precipitation for future projections, and both of these future projections have simulated that precipitation might decrease during the period of 2021–2100.

The outcomes of this study are believed to be a guide for decision makers in governmental agencies and also a reference for hydrologists who deal with estimation of water resources of river basins.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

- Al-Juboori, A. M. & Guven, A. 2016 [Hydropower plant site assessment by integrated hydrological modeling, gene expression programming and visual basic programming](#). *Water Resour Manage* **30**, 2517–2530.
- Akaike, H. 1974 A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.
- Andersen, H. E., Brian, K., Søren, E. L., Carl, C. H., Torben, S. J. & Erik, K. R. 2006 [Climate change impacts on hydrology and nutrients in a Danish lowland river basin](#). *Science of the Total Environment* **365**, 223–237.
- Assaleh, K., Shanableh, T. & Kheil, Y. A. 2013 Group Method of Data Handling for modelling magnetor heological dampers. *Scientific Research* **4** (1), 70–79.
- Chen, S.-T., Pao-Shan, Y. U. & Tang, Y.-H. 2010a [Statistical downscaling of daily precipitation using support vector machines and multivariate analysis](#). *Journal of Hydrology* **385**, 13–22.
- Chen, S.-T., Yu, P.-S. & Tang, Y.-H. 2010b [Statistical downscaling of daily precipitation using support vector machine and multivariate analysis](#). *Journal of Hydrology* **385**, 13–22.
- Chen, H., Xu, C. Y. & Shenglia, G. 2012 [Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff](#). *Journal of Hydrology* **434–435**, 36–45.
- Daneshfaraz, R., Aminvash, E., Ghaderi, A., Abraham, J. & Bagherzadeh, M. 2021a [SVM performance for predicting the effect of horizontal screen diameters on the hydraulic parameters of a vertical drop](#). *Applied Sciences* **11** (9), 4238.
- Daneshfaraz, R., Bagherzadeh, M., Esmaeeli, R., Norouzi, R. & Abraham, J. 2021b [Study of the performance of support vector machine for predicting vertical drop hydraulic parameters in the presence of dual horizontal screens](#). *Water Supply* **21** (1), 217–231.
- Duffy, J. J. & Franklin, M. 1975 [A learning identification algorithm and its application to an environmental system](#). *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-5** (2), 226–240.
- Ferreira, C. 2001 Gene expression programming: a New adaptive algorithm for solving problems. *Complex Systems* **13** (2), 87–129.
- Green, R. E., Reichelt, R. E. & Bradbury, R. H. 1988 [Statistical behaviour of the GMDH algorithm](#). *Biometrics* **44**, 49–69.
- Guven, A. 2009 [Linear genetic programming for time-series modeling of daily flow rate](#). *Journal of Earth System* **118** (2), 137–146.
- Guven, A. & Aytok, A. 2009 [New approach for stage discharge relationship: gene expression programming](#). *Journal of Hydrologic Engineering* **14**, 812–820.
- Guven, A. & Talu, N. E. 2010 [Gene expression programing for estimating suspended sediment yield in Middle Euphrates Basin, Turkey](#). *CLEAN-Soil Air Water* **38** (12), 1159–1168.
- Hashmi, M. Z., Shamseldin, A. Y. & Melville, B. W. 2011 [Statistical downscaling of watershed precipitation using gene expression programming](#). *Environmental Modelling & Software* **26** (12), 1639–1646.
- Haylock, M. R., Cawley, G. C., Harpham, C., Wilby, R. L. & Goodess, C. M. 2006 [Downscaling heavy precipitation over the UK: a comparison of dynamical and statistical methods and their future scenarios](#). *International Journal of Climatology* **26** (10), 1397–1415.
- IPCC 2007 [Climate Change Synthesis Report](#). Intergovernmental Panel on Climate Change, Geneva, Switzerland.
- Ivakhnenko, A. G. 1971 [Polynomial theory of complex systems](#). *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-1** (4), 364–378.
- Josef, T. & Petr, B. 2011 [Exchange rate predictions in international financial management by enhanced GMDH algorithm](#). *Prague Economic Papers* **2011** (3), 232–249.
- Karimi, S., Shiri, J., Kisi, O. & Xu, T. 2017 [Forecasting daily streamflow values: assessing heuristic models](#). *Hydrology Research* **49** (3), 658–669.
- Mpelasoka, F. S., Mullan, A. B. & Heerdegen, R. G. 2001 [New Zealand climate change information derived by multivariate statistical and artificial neural networks approaches](#). *International Journal of Climatology* **21**, 1415–1433.
- Muttill, N. & Liong, Y. 2012 [Improving Runoff forecasting by input variable selection in Genetic Programming](#).
- Raghavendra, S. N. & Deka, P. C. 2014 [Support vector machine applications in the field of hydrology: a review](#). *Applied Soft Computing* **19**, 372–386.

- Saburo, I., Mikiko, O. & Yoshikazu, S. 1976 Sequential GMDH algorithm and its application to river flow prediction. *IEEE Transactions on Systems, Man, and Cybernetics SMC-6* (7), 473–479.
- Sadeghfam, S., Khatibi, R., Moradian, T. & Daneshfaraz, R. 2021 Statistical downscaling of precipitation using inclusive multiple modelling (IMM) at two levels. *Journal of Water and Climate Change*. <https://doi.org/10.2166/wcc.2021.106>.
- Schmidli, J., Goodess, C. M., Frei, C., Haylock, M. R., Hundscha, Y., Ribalaygua, J. & Scmith, T. 2007 Statistical and dynamical downscaling of precipitation; an evaluation and comparison of scenarios for the European Alps. *Journal of Geophysical Research* **112** (D4), 1–20.
- Seckin, N. & Guven, A. 2012 Estimation of peak flood discharges at ungauged sites across Turkey. *Water Resour Manage* **26**, 2569–2581.
- Shahid, S., Hadipour, S. & Harun, S. B. 2013 Genetic Programming for Downscaling Extreme Rainfall Events. In: *First International Conference on Artificial Intelligence, Modelling & Simulation*.
- Singh, D., Jain, S. K. & Gupta, R. 2015 Statistical downscaling and projection of future temperature and precipitation change in middle catchment of Sutlej River Basin, India. *Journal of Earth System Science* **124**, 843–860.
- Tofiq, F. A. & Guven, A. 2014 Prediction of design flood discharge by statistical downscaling and General Circulation Models. *Journal of Hydrology* **517**, 1145–1153.
- Tofiq, F. A. & Guven, A. 2015 Potential changes in inflow design flood under future climate projections for Darbandikhan Dam. *Journal of Hydrology* **528**, 45–45.
- Traore, S. & Guven, A. 2013 New algebraic formulations of evapotranspiration extracted from gene-expression programming in the tropical seasonally dry regions of West Africa. *Irrigation Science* **31**, 1–10.
- Vapnik, V. N. 1995 *The Nature of Statistical Learning Theory*. Springer Verlag, New York, p. 314.
- Vapnik, V. 1999 An overview of statistical learning theory. *IEEE Transactions on Neural Networks* **10** (5), 988–999.
- Wilby, R. L. & Wigley, T. M. L. 1997 Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography* **21**, 530–548.
- Yonggang, M., Huang, Y., Chen, X., Li, Y. & Bao, A. 2007 Modeling snowmelt runoff under climate change scenarios in an ungauged mountainous watershed, Northwest China. *Mathematical Problems in Engineering* **2013**, 1–9.
- Zahraie, B., Nasser, M. & Nematizadeh, F. 2017 Exploring spatiotemporal meteorological correlations for basin-scale meteorological drought forecasting using data mining methods. *Arabian Journal of Geosciences* **10**, 419.

First received 5 August 2021; accepted in revised form 14 October 2021. Available online 27 October 2021