Evaluation of data-driven models to downscale rainfall parameters from global climate models outputs: the case study of Latyan watershed

Reza Haji Hosseini, Saeed Golian and Jafar Yazdi

ABSTRACT

Assessment of climate change in future periods is considered necessary, especially with regard to probable changes to water resources. One of the methods for estimating climate change is the use of the simulation outputs of general circulation models (GCMs). However, due to the low resolution of these models, they are not applicable to regional and local studies and downscaling methods should be applied. The purpose of the present study was to use GCM models' outputs for downscaling precipitation measurements at Amameh station in Latyan dam basin. For this purpose, the observation data from the Amameh station during the 1980–2005 period, 26 output variables from two GCM models, namely, HadCM3 and CanESM2 were used. Downscaling was performed by three data-driven methods, namely, artificial neural network (ANN), nonparametric K-nearest neighborhood (KNN) method, and adaptive network-based fuzzy inference system method (ANFIS). Comparison of the monthly results showed the superiority of KNN compared to the other two methods in simulating precipitation. However, all three, ANN, KNN, and ANFIS methods, showed satisfactory results for both HadDCM3 and CanESM2 GCM models in downscaling precipitation in the study area.

Key words | artificial intelligence, CanESM2, climate change, downscaling, GCM, HadCM3

Reza Haji Hosseini
Saeed Golian (corresponding author)
Department of Civil Engineering,
Shahrood University of Technology,
Shahrood,
Iran
E-mail: *s.golian@shahroodut.ac.ir*

Jafar Yazdi
Faculty of Civil, Water and Environmental Engineering,
Shahid Beheshti University,
Tehran,
Iran

INTRODUCTION

According to the Intergovernmental Panel on Climate Change (IPCC), the climate of the planet is changing (IPCC 2007). Assessments and research show that the reason for this change is the increase of greenhouse gas emissions, especially CO_2. Some studies have estimated an average global temperature increase between 0.76 and 6.4 °C until 2100 under the A2 emission scenario (IPCC 2008). Also, Table 1 (Xu & Xu 2012) contains temperature changes under the RCP scenarios over the entire globe for the Fifth Assessment Report (AR5) scenarios.

The multimodel ensemble of the Coupled Model Intercomparison Project, Phase 5 (CMIP5) and its predecessors provide critical inputs to the assessment reports produced within the IPCC framework and are also used as input for further investigations of climate change and its impacts (Bring *et al.* 2015).

General circulation models (GCMs) are commonly applied in climate change studies. Although GCMs are capable of representing the primary features of global atmospheric circulation very well, their resolution is not high enough to reproduce regional climatic details (Syed *et al.* 2012). To provide an appropriate logical relationship between GCM outputs and the requirements for climate impact studies, a variety of downscaling methods and regional climate models have been developed. In these methods, statistical relationships are explored between the

doi: 10.2166/wcc.2018.191

Table 1 | Temperature changes under the RCP scenarios over the globe (AR5)

	RCP 2.6	RCP 4.5	RCP 8.5
2011–2040	0.75	0.78	0.88
2041–2070	1.07	1.44	2.07
2070–2100	1.06	1.8	3.55

variables simulated by GCMs that make empirical–statistical relationships between independent variables (Predictor) and dependent variables (Predictant).

Although linear regression has been most widely used (Mahani 2015; Campozano *et al.* 2016), recently, nonlinear methods have emerged (Shahverdi *et al.* 2017). The interest in nonlinear regression methods, for example artificial neural networks (ANNs), is increasing because of their high capability to simulate the complex, nonlinear, and time-varying characteristics of atmospheric variables at different scales (Duhan & Pandey 2015; Tue Vu *et al.* 2016). In addition to the ANN, we have used two other nonlinear methods, namely, KNN and ANFIS, to compare the quality of downscaling methods.

Mahani (2015) used ANN for evaluating the effects of climate change on Polrud River using two GCM models, namely, HadCM3 and CGCM3, on three hydroclimatologic variables: temperature, precipitation, and peak discharge. The results showed an increase in all three parameters, but the results of the CGCM3 model showed regular and more increase compared with HadCM3.

Tue Vu *et al.* (2016) applied ANN as a statistical downscaling model (SDSM) on GCMs during the rainy season at some meteorological gauges in Bangkok, Thailand. The predictors were first selected over different grid boxes surrounding the Bangkok region and then screened by using principal component analysis (PCA) to filter the best correlated predictors for ANN training. The reanalysis downscaled results of the present day climate showed good agreement against station precipitation with a correlation coefficient of 0.8 and a Nash–Sutcliffe efficiency of 0.65.

Campozano *et al.* (2016) presented the downscaling of monthly precipitation estimates of the NCEP/NCAR reanalysis 1 applying the SDSM, ANNs, and the least squares support vector machines (LS-SVM) approach. Downscaled monthly precipitation estimates after bias and variance correction were compared to the median. A preliminary comparison revealed that both artificial intelligence methods, ANN and LS-SVM, performed equally. Results disclosed that the ANN and LS-SVM methods depict, in general, better skills in comparison to SDSM.

Wu *et al.* (2010) studied the application of K-nearest neighbor (KNN) to derive local precipitations based on NCEP Climate Forecast System (CFS) seasonal forecasts and historic rainfall observations. Their study focused on the semiarid area along the southeastern Mediterranean coast. This region is strongly influenced by the Mediterranean climate and complex terrain. This study constructed 60 ensemble members for probabilistic estimates. The KNN algorithm demonstrated its robustness when validated with NCEP/DOE reanalysis from 1981 to 2009 as hind casts before being applied to downscale CFS forecasts. The downscaled predictions show fine-scale information, such as station-to-station variability. The verification against observations shows improved skills of this downscaling utility relative to the CFS model.

Emamgholizadeh *et al.* (2014) investigated the potential of two intelligence models, namely, ANN and adaptive neuro-fuzzy inference system (ANFIS) in estimating the groundwater level of the Bastam Plain in Iran. The results showed that the ANN and ANFIS models can estimate GWL accurately. Also, it was found that the ANFIS model (with root-mean-square-error (RMSE) 0.02 m and determination coefficient (R2) of 0.96) performed better than the ANN model with RMSE = 1.06 m and R2 = 0.83. Djamil & Aldrian (2008) investigated the use of multi-variable ANFIS in assessing daily rainfall using several surface weather parameters as predictors. The data used in that study came from automatic weather station data collected in Timika airport from January until July 2005 with a 15-minute time interval. Talei *et al.* (2010) investigated the effect of inputs used on event-based runoff estimating by ANFIS. Fifteen ANFIS models were compared, differentiated by the choice of rainfall and/or discharge inputs used.

In this study, we assess the precipitation changes and climatic parameters simulated from two GCMs, namely, HadCM3 and CanESM2, and compare them with parameters derived from observed precipitation in the study area. Three data-driven methods, namely, ANN, KNN, and

ANFIS, will be applied to downscale the outputs of GCM models. According to the conducted literature survey, the ANFIS method has not been used so far for precipitation/temperature downscaling, although it has been one of the widely used data-driven models for estimating purposes in other applications of hydrology. The performance of the downscaling method on outputs of two widely used GCM models from two different IPCC modeling exercises, i.e., CMIP4 and CMIP5, will be assessed at a study area in Iran. From the fourth generation, the HadCM3 model has been shown to perform satisfactorily for many parts of Iran (e.g., Samadi *et al.* 2010; Farzaneh *et al.* 2012), while from the fifth IPCC modeling exercise, CanESMs2 is one of the most widely used models over Iran (e.g., Hesami & Zeynolabedini 2016; Rouhani *et al.* 2016).

The performance of three widely used data-driven models, namely, KNN, ANN, and ANFIS in downscaling GCM outputs will be evaluated to select the superior method for GCM downscaling. As far as the authors know, this comparison has not been done in other studies and thus is the contribution made by this research.

CASE STUDY

The Latyan watershed is situated northeast of Tehran between latitudes 35°45′N and 36°15′N and longitudes 51°20′E and 51°55′E. This basin is generally divided into nine sub-basins and Amameh is considered as one of its sub-basins. Amameh River is one of the branches of the Jajrood River that reaches to the dam of Latyan. This dam supplies a large proportion of the water demands of Tehran city. Thus, hydro-climatic studies on this river are of major importance. In Figure 1, the location of the area of interest in Tehran province, and also separately, is shown.

In this study, observed daily precipitation data (1980–2005) are fed to the data-driven models as output (target) data. The output parameters for the same period from HadCM3 and CanESM2 were also downloaded from the Canadian Climate Data and Scenarios website (www.cccsn.ec.gc.ca). These parameters are described in Table 2. These data are fed as input data to the data-driven models. In this study, 85% of the data (1980–2001) was allocated for model calibration (train and validation phases) and 15% (2002–2005) to validate (test) the models.

Long-term time series of standardized daily values of parameters are extracted into a one column text file per grid cell (box). The 128 × x64 grid cells cover a global domain according to T42 Gaussian grid. This grid is uniform along the longitude with horizontal resolution of 2.8125° and nearly uniform along the latitude of roughly 2.8125°.

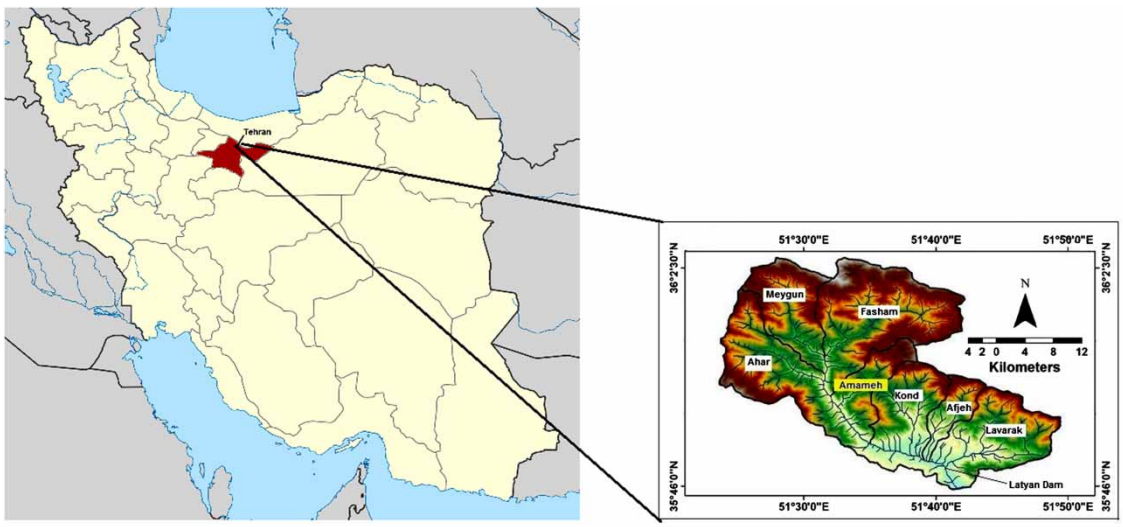

Figure 1 | Location of the Latyan dam watershed and detailed map of the subbasin upstream of Amameh station.

Table 2 | Climate predictor variables for the HadCM3 and CanESM2 models

Row	Evaluator variable	Definition	Description
1	P5_f	Geostrophic airflow velocity at 500 hPa	Geostrophic flow velocity at 500 hectopascal
2	P5_u	Horizontal wind at 500 hPa	Horizontal wind at 500 hPa
3	P5_v	Zonal wind at 500 hPa	Wind area of 500 hPa
4	P5_z	Vorticity at 500 hPa	Vorticity at 500 hPa
5	P5th	Wind direction at 500 hp	Wind at 500 hPa
6	P5zh	Divergence at 500 hPa	Divergence at 500 hPa height
7	P500	Geopotential height at 500 hPa	Geopotential height at 500 hPa
8	R500	Relative humidity at 500 hPa	Relative humidity at 500 hPa
9	P_f	Surface geostrophic airflow	Geostrophic air flow surface
10	P_u	Surface horizontal wind	Surface horizontal wind
11	P_v	Surface zonal wind	Wind surface area
12	P_z	Surface vorticity	A measure of the air vorticity
13	P_th	Surface wind direction	Surface wind direction
14	P_zh	Surface divergence	Surface divergence
15	P8_f	Geostrophic airflow velocity at 850 hPa	Geostrophic flow velocity at 850 hPa
16	P8_u	Horizontal wind at 850 hPa	Horizontal wind at 850 hPa
17	P8_v	Zonal wind at 850 hPa	Wind region at 850 hPa
18	P8_z	Vorticity at 850 hPa	Vorticity at 850 hPa
19	P8th	Wind direction at 850 hp	Wind direction at 850 hPa
20	P8zh	Divergence at 850 hPa	Divergence at 850 hp
21	P850	Geopotential height at 850 hPa	Geopotential height at 850 hPa
22	R850	Relative humidity at 850 hPa	Relative humidity at 850 hPa
23	Mslp	Mean sea level pressure	Medium pressure from sea level
24	Prcp	Total precipitation	Total precipitation
25	Shum	Near surface specific humidity	Humidity near the surface
26	Temp	2 m air temperature	2 m air temperature

METHODOLOGY

The proposed framework for the downscaling process of this study is shown in Figure 2. The whole process is performed for outputs of both HadCM3 and CanESM2 GCM models. The PCA method is also used for selection of the most informative inputs (predictors) for data-driven models. The advantage of PCA is that by using a small number of principal components it is possible to represent the variability of the original multivariate data set. At the same time, the principal components are uncorrelated and therefore there is no redundant information (Shashikanth & Ghosh 2013).

Artificial neural network

The ANN has shown a good performance as a widely used method, in modeling and assessing nonlinear and unstable time series for processes that have no explicit solution and explicit recognition and description of them (Zohdi 2000). The neural network has the ability to recognize the pattern, and establishes a good relationship between input and output data. Compared to other methods, ANN has less sensitivity relative to input errors. ANN after training can evaluate system responses without the need for any explicit mathematical relationship (Bustami *et al.* 2007).

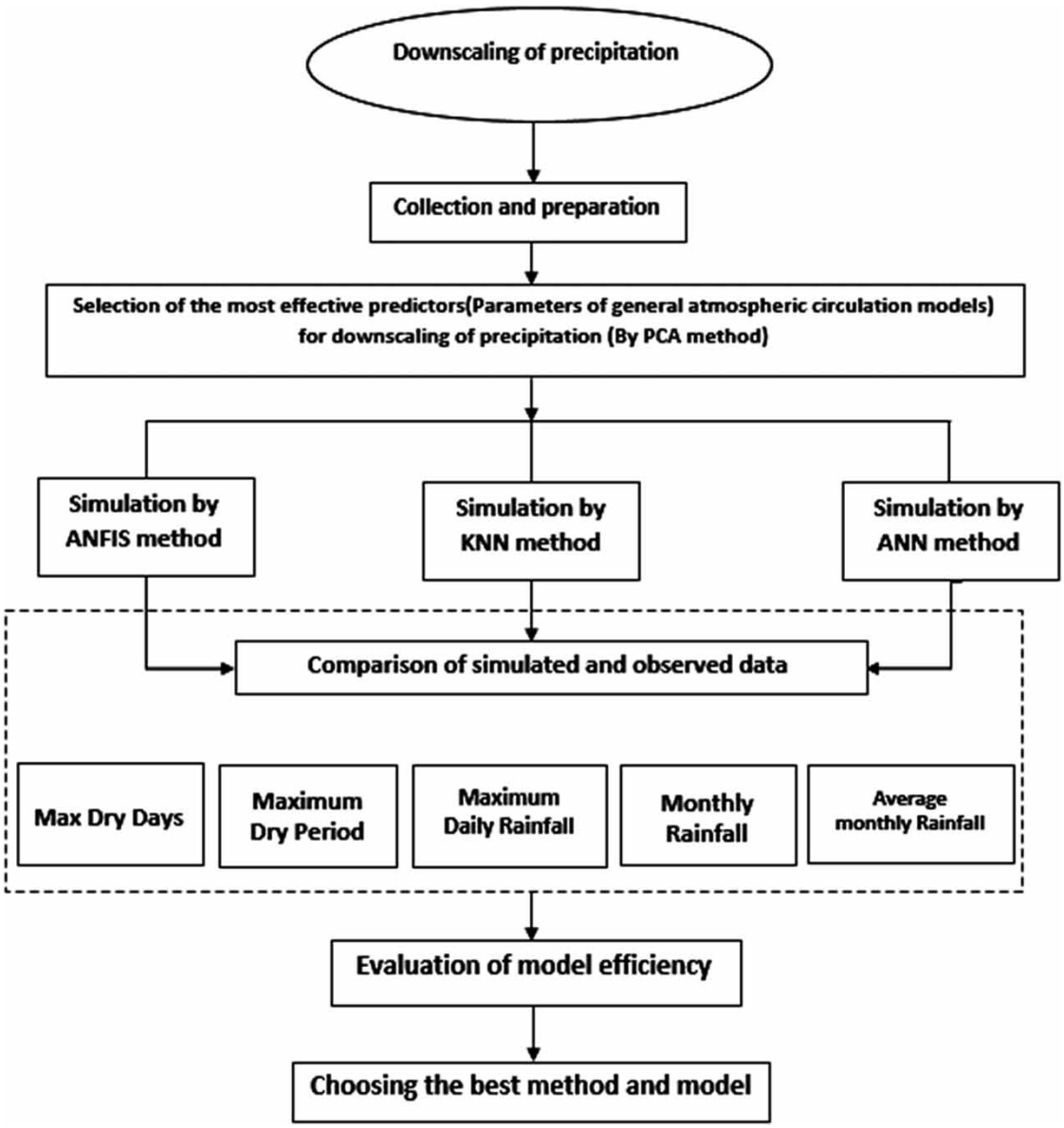

Figure 2 | Suggested algorithm for precipitation downscaling.

The input layer is used to enter the data into the network, the output layer to generate the appropriate responses of the inputs, and one or more intermediate layers composed of processor nodes which in fact are the locations of data processing. The number of neurons in the input and output layers is determined by the nature of the problem under consideration. Likewise, the number of hidden layers and the number of neurons in each hidden layer is usually determined by trial-and-error method in order to reduce the amount of network error (Trafalis *et al.* 2005). However, it is recommended that the number of hidden layers be as low as possible. Therefore, the network is trained by one hidden layer first and in case of inappropriate performance, the number of layers are added. This method is also applied to determine the number of neurons in each hidden layer so that a smaller number of neurons is considered first and if the results are not satisfactory, they will be increased. The nodes of

adjacent layers in the network are fully interconnected (Satish *et al.* 2004). Inputs of each node are the values of input variables or output of other nodes. Each node has an activation function. Figure 3 shows a schematic view of the multi-layered ANN. The most widely used activation functions are: sigmoid tangent function, linear and sigmoid logarithm (demo and bile) functions.

The inputs are in the form of $X\ (X_1, X_2,\ldots,X_n)$ vector and each input is related to a processor node by a weight, and finally, a string of the weights as $W(W_1,W_2,\ldots,W_n)$ is related to the considered node. The output of the node, which is called y, is calculated by the following equation:

$$Y = f(X.W - b) \tag{1}$$

In the above relationship, X is the vector of input variables, W is the weight vector, and b is called bias.

Generally, neural networks are divided into two types of backward and forward. The difference is that in backward networks, there is at least one return signal from a neuron to the same neuron or neurons of the same layer or previous layer. In most cases, backward neural networks can be very useful. However, in 80% of applications, forward neural networks are used. The multi-layer perceptron network is one of the most widely used forward ANNs, especially in modeling climatic elements in which each neuron in each layer is connected to all neurons in the previous layer (Wang & Sheng 2010).

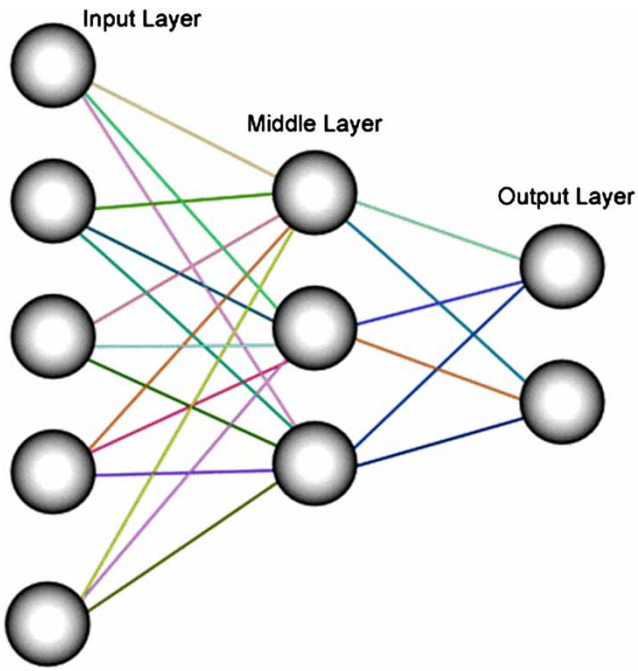

Figure 3 | Multi-layered artificial neural network.

To train neural networks, there are four conventional training algorithms based on the layered perceptron structure. The most widely used methods are Levenberg–Marquardt and conjugate gradient. The Levenberg–Marquardt algorithm has been recognized as the fastest learning method for neural networks since 1993 to date. In this study, we used the daily observation data of the Amameh station as target data and the combination of the daily data of 26 parameters of the HadCM3 and CanESM2 models as input data during the 1980–2005 historical period (according to Table 2). Also, to derive an appropriate architecture for the ANN, the number of neurons in the hidden layer increased from two up to 50 neurons, and the results were evaluated and compared with each other. It was found that the best performance was derived for 30 neurons in the hidden layer with sigmoid and linear activation functions for the hidden and output layers, respectively. We also allocated 85% of the data (1980–2001) for model calibration (train and validation phases) and 15% (2002–2005) to validate (test) the models.

K-nearest neighborhood (KNN)

Nonparametric estimation of probability densities and regression functions is pursued through weighted local averages of the dependent variable. This is the foundation for nearest neighbor methods. KNN methods use the similarity (neighborhood) between observations of predictors and similar sets of historical observations (successors) to obtain the best estimate for a dependent variable (Karlsson & Yakowitz 1987; Lall & Sharma 1996).

Nonparametric regression is a form of regression analysis in which the predictors do not take a predetermined form but are constructed according to information derived from the data. Nonparametric regression requires larger sample sizes than regression based on parametric models because the data must supply the model structure as well as the model estimates. The KNN method imposes a metric on the predictors to find the set of K past nearest neighbors

for the current condition in which the nearest neighbors have the lowest distance. The distance between the current and historical condition can be calculated by the Euclidian (Karlsson & Yakowitz 1987) or Mahalanobis distance (Yates *et al.* 2003) between current and historical predictors.

The algorithmic procedure of a KNN regression is summarized in Figure 4 and is presented as follows.

Determine the vector of current m independent variables also known as predictors, $X_r - \{x_{1r}, x_{2r}, x_{3r} \ldots x_{mr}\}$, associated with the dependent variable, Y_r.

Determine the matrix of $n \times m$ predictors containing n vectors of already observed predictors, $X_t - \{x_{1t}, x_{2t}, x_{3t} \ldots x_{mt}\}$; $t = 1, 2\ldots, n$.

Calculate n distances between current predictors and the observed predictors, Δ_{rt}. Select K sets of predictors/dependent variables (X_k, Y_k), which have the lowest values of Δ_{rt}. Those sets are known as the K-nearest neighbors. Next, a kernel function associated with each K-nearest neighbor is calculated as follows:

$$f_k(\Delta_{rk}) = \frac{1/\Delta_{rk}}{\sum_{k-1}^{k}(1/\Delta_{rk})} \tag{2}$$

Obviously, $\sum_{k-1}^{k} f_k(\Delta_{kr}) = 1$. The unknown Y is finally calculated as:

$$Y_r = \sum_{k-1}^{k} f_k(\Delta_{kr}) \times Y_k \tag{3}$$

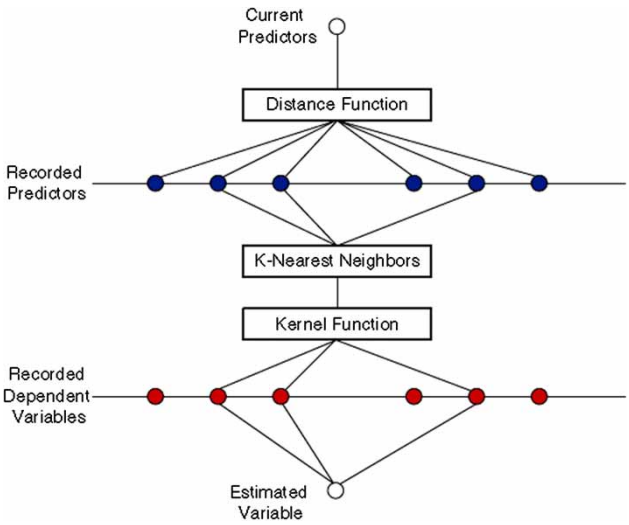

Figure 4 | The schematic of KNN algorithm.

The overall process for the KNN method is shown in Figure 4.

The distance function is usually calculated by a Euclidean distance or a Mahalanobis distance. A Euclidean distance between ith and jth predictors is calculated as:

$$\Delta_{ij} - \sqrt{(X_{1i} - X_{1j)}^2 + (X_{2i} - X_{2j)}^2 + \ldots + (X_{mi} - X_{mj)}^2} \tag{4}$$

where m is the dimension of the predictors. The Mahalanobis distance uses the following equation:

$$\Delta_{ij} - \sqrt{(X_i - X_j)C_t^{-1}(X_i - X_j)^T} \tag{5}$$

where C is the covariance matrix between X and Y.

Mahalanobis distance is a distance measure introduced by Mahalanobis in 1936 (Mahalanobis 1936). It is based on correlations between variables by which different patterns can be identified and analyzed. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale invariant.

Lall & Sharma (1996) suggested that instead of the kernel function of

$$f_k(\Delta_{rk}) - \frac{1/\Delta_{rk}}{\sum_{k-1}^{k}(1/\Delta_{rk})} \tag{6}$$

the following function could be used:

$$\frac{f_k(j) - 1/j}{\sum_{j-1}^{k} 1/j} \tag{7}$$

where j is the order of the neighbors after sorting them in an ascending order. Neighbors with higher distance get higher orders and the lower contribution to the final output.

The KNN classifier is a very simple classifier that works well on basic recognition problems.

After trial and error, the KNN model performs best with $K = 5$. Also, the Euclidean method was considered as the distance function.

Adaptive network-based fuzzy inference system (ANFIS)

One of the methods that has been recently considered in hydrology is modeling based on fuzzy rules. Fuzzy logic

and the theory of fuzzy sets are used to describe human thinking and reasoning in a mathematical framework. Fuzzy modeling is called fuzzy inference system (FIS) and its primary structure consists of three components: (a) the law base contains a set of fuzzy rules, (b) the database that defines the membership functions (MFs) used in fuzzy rules, and (c) the mechanism of the argument, which, according to the rules, relates the input pattern to the corresponding output. Using some if-then rules describes a nonlinear component relationship from the input space to the output space.

The various combinations of membership functions create the input and output variables of the rules and these rules define a fuzzy region from the input space, and finally, the output relationship determines the output of the model. The efficiency of FIS depends on its parameters' estimation which includes the parameters of the membership functions and the output function of each rule. To solve the problem of identifying the parameters in an FIS in neuro-fuzzy models, a comparative network, which is the general state of the multilayer forward neural network, is used.

In this research, ANFIS, which is a fuzzy-neural model, is used. The most common type of FIS that can fit in a matching network is the Sugeno's fuzzy system in which output is a linear relationship and its parameters can be estimated by combining the least error squares methods and the back propagation error based on the gradient reduction. In Figure 5, an example of a first-order Sugeno FIS with two inputs x, y and output z is shown. For this FIS, a sample of the fuzzy rule base containing two rules can be presented as follows:

- First Law: If x equals A1 and y equals to B1, then $f1 = p_1 x + q_1 y + r_1$
- Second Law: If x equals A2 and y equals to B2, then $f2 = p_2 x + q_2 y + r_2$

where *B2*, *B1* and *A2*, *A1*, are the membership functions for input of *y* and *x*, respectively. *r1*, *q1*, *p1*, *r2*, *p2 q2* are also parameters of the output functions for the two defined rules. An example of the usual architecture of the ANFIS model is presented in Figure 6, in which the nodes of each layer have the same function.

Layer 1: Each node in this layer produces the membership classes of an input variable. The output is defined by the following relationships:

$$OP_i^1 = \mu_{Ai}(x) \ \ for\ i = 1,\ 2 \tag{8}$$

$$OP_i^1 = \mu_{Bi}(x) \ \ for\ i = 3,\ 4 \tag{9}$$

where *x* (or *y*) is the input node, *Ai* or (*Bi-2*) is the fuzzy set associated with this node, which is determined by the form of the membership functions of this node, and any suitable function that is continuous and fragmented, such as Gaussian functions, trapezoidal and triangular, can be used as a membership function. Assuming the Gaussian membership function as a membership function, the output of OP_i can be calculated as follows:

$$OP_i^1 = \mu_{Ai}(x) = \exp\left[-0.5\{(x - c_i)/\sigma_i\}^2\right] \tag{10}$$

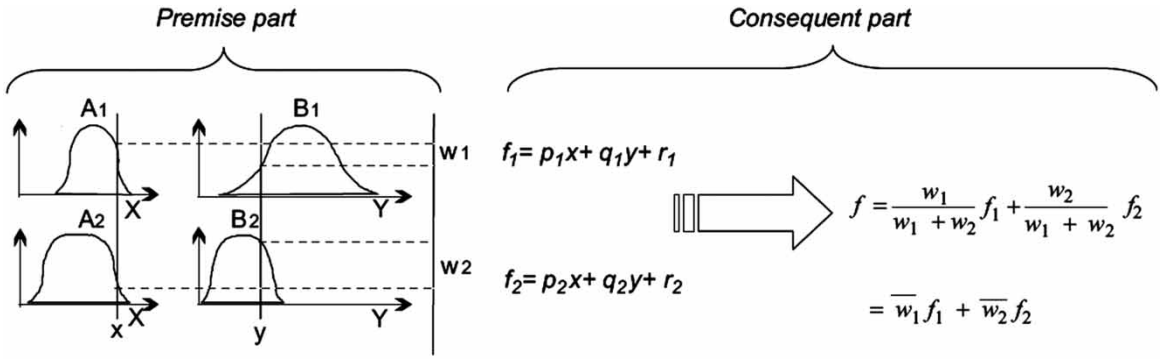

Figure 5 | Sugeno fuzzy inference system (Alemzadeh *et al.* 2004).

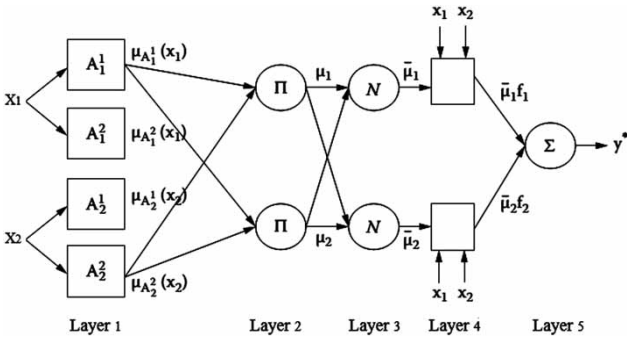

Figure 6 | The architecture of ANFIS model equivalent to the inference system discussed.

where c_i and σ_i are the mean and standard deviation of the i^{th} membership function, respectively.

Layer 2: Each node in this layer is multiplied by the input signal, and the output that represents the power of the excitation of a rule is calculated as follows:

$$OP_i^2 = w_i = \mu_{Ai}(x)\mu_{Bi}(y) \quad i = 1\,,\ 2 \tag{11}$$

Layer 3: The i^{th} node of this layer, which is denoted by N, computes the normalized stimulant power:

$$OP_i^3 = \bar{w}_i = \frac{w_i}{w_1 + w_2} \quad i = 1\,,\ 2 \tag{12}$$

Layer 4: The nodes i in this layer compute the i^{th}-rule to the output of the model using the following function node:

$$OP_i^4 = \bar{w}_i f_i = \bar{w}_i(p_i x + q_i y + r_i) \tag{13}$$

where the output of layer 3 and {pi, qi, ri} are the set of parameters of the linear function of the output of the i-th rule.

Layer 5: The only node in this layer calculates the overall output of ANFIS as follows:

$$OP_i^5 = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{14}$$

The main function of the adaptive system is to optimize the model parameters. Jang *et al.* (1997) devised a hybrid teaching method for the neuro-fuzzy model which is faster and more accurate than the back-propagation method based on gradient reduction in calculating model parameters. Combined training algorithm for ANFIS consists of two alternating phases:

- Reducing the gradient that returns the generated error signals from the output layer to the input layer. This phase corrects the parameters of the front part of the model (membership functions).
- The method of least squares corrects the parameters of the model portion of the model (linear relationship coefficients).

Principal component analysis (PCA)

PCA is a statistical procedure that uses an orthogonal transformation using the Eigen value–Eigen vector decomposition technique to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Noori *et al.* 2011). The number of distinct principal components is equal to the smaller of the number of original variables or the number of observations minus one. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set.

PCA is sensitive to the relative scaling of the original variables. Working with these independent variables might provide better accuracy of evaluation for the predictant variable(s) depending on the problem at hand. When the volume of information and input (predictor) variables are relatively high, this may have negative effects on the accuracy of evaluation because of the noises imported to the data-driven model.

In this study, 26 original variables (predictors) were converted to the independent variables by PCA. Through the PCA approach, independent variables are sorted from the most important to the least in terms of the value of information. A sensitivity analysis has been done and independent variables were omitted from the last variables, one by one, and in each round, the data-driven model was trained and tested. The results showed that using 15 GCM

independent variables, the best performance is achieved. It should be noted that each independent variable (obtained by PCA) is a linear function of all original 26 GCM variables.

Data normalization

Normalization scales all input variables in the same order and often have a positive effect on evaluation accuracy. Working with raw data can reduce network speed and accuracy. Therefore, by using the following equation, all input and output data are initially normalized and then entered into the neural network.

$$X_n = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{15}$$

In the above relation, X_n is normalized data and the indexes i, X_{min}, X_{max} are respectively the rows, minimum and maximum of that data in their set.

Model evaluation

In order to evaluate the performance of ANN, KNN, and ANFIS models, three statistical criteria, namely, Nash–Sutcliffe model efficiency coefficient (*NSE*), relative mean absolute error (*RMAE*), and correlation coefficient (*R*) were used as follows:

$$E = 1 - \frac{\sum_{t=1}^{T} (Q_m^t - Q_o^t)^2}{\sum_{t=1}^{T} (Q_o^t - \bar{Q}_o^t)^2} \tag{16}$$

$$RMAE = \frac{1}{M} \sum_{i=1}^{M} \frac{|X_{i.m} - X_{i.o}|}{X_{i.o}} \tag{17}$$

$$R = \frac{M \sum_{i=1}^{M} X_{i.m}.X_{i.o} - \sum_{i=1}^{M} X_{i.m}.\sum_{i=1}^{M} X_{i.o}}{\sqrt{\left[M \sum_{i=1}^{M} X_{i.m}^2 - (\sum_{i=1}^{M} X_{i.m})^2\right].\left[M \sum_{i=1}^{M} X_{i.0}^2 - (\sum_{i=1}^{M} X_{i.0})^2\right]}} \tag{18}$$

where M is the total number of input data, X_{im} represents the i^{th} estimated data using one of the four above models, and X_{io} represents the i^{th} data.

To evaluate the performance of rainfall downscaling from the CanESM2 and HadCM3 models, daily data at Amameh station located above the Latyan dam were used. In this research, five important meteorological variables were studied to evaluate the performance of ANN, KNN, and ANFIS in downscaling of the GCM model data. These five parameters are: length of the longest sequence of dry days in the month (days), maximum daily precipitation in the month (mm), total precipitation in the month (mm), number of dry days in the month, and average monthly precipitation (mm). For instance, Karamouz *et al.* (2013) used these parameters to evaluate ANN and SDSM methods for downscaling GCM outputs. In other studies, such as those of Fu *et al.* (2012), Verbist *et al.* (2010), Jones *et al.* (2009), and Schoof & Pryor (2001), similar parameters were used for the assessment of downscaling results. In this study, the downscaling methods are compared with the same parameters.

RESULTS

In this paper, we examine the climate change and downscaling of GCMs for the upper basin of Latyan dam and subbasin of Emamah in the period between 1980 and 2005. For this purpose, daily data were collected from 1980 to 2005. These data were fed into the device as output (target). Furthermore, the data of two models of global climate, named HadCM3 and CanEM2, were downloaded for the period between 1980 and 2005. These models contain 26 parameters that are shown in Table 2. These data are fed to the model as inputs.

Given the high number of input parameters, these 26 parameters are prioritized using the PCA method and simulation was performed according to the priorities that derived from the PCA method. After several experiments on the data, the first 15 PCs eventually offer the best answers for simulating climatic data. After preparing the input–output data sets for simulation, the model's training and validation begin. As indicated before, in this study, 85% of the data (1980–2001) was allocated for model calibration (training and validation phases) and 15% (2002–2005) to validate (test) the models.

In order to evaluate the accuracy of the downscaling performance of three downscaling methods, namely, ANN, KNN, and ANFIS based on outputs of two GCM models,

i.e., HadCM3 and CanESM2, monthly datasets were calculated from daily data. Next, these data sets were analyzed in the form of five important and fundamental parameters, including length of the longest sequence of dry days in the month (days), maximum daily precipitation in the month (mm), total precipitation in the month (mm), number of dry days in the month, and average monthly precipitation (mm). The reason for selecting these parameters is that they can present both the average and extreme hydroclimatological state of a region relating to precipitation.

In this research, in order to use the neural network and to derive an appropriate architecture for the ANN, the number of neurons in the hidden layer was increased from 2 up to 50 neurons, and the results were evaluated and compared with each other. It was found that the best performance was derived for 30 neurons in the hidden layer with sigmoid and linear activation functions for the hidden and output layers, respectively.

In the KNN method also, after trial and error, the KNN model performs best with $K = 5$. Also, the Euclidean method was considered as the distance function.

The results of three downscaling methods, i.e., ANN, KNN, and ANFIS, based on outputs of the HadCM3 and CanESM2 models, are presented in Figures 7 and 8, respectively.

As is shown in Figure 7, all three simulation methods provide acceptable results for each of the five parameters. As expected, the models' performance is better in the calibration phase compared to the test period. Also, it can be seen that for the parameters which are directly related to rainfall, i.e., maximum daily precipitation and total precipitation in a month, all downscaling models reveal weaker performance for the first four months of a year in addition to November and December, i.e., winter and spring months, and then the simulation graphs for all simulation methods reveal similar behavior and very close to observation.

Also, in Figure 8, downscaling outputs of CanESM2 model data yielded satisfactory results compared to observation data. Again, for the two parameters, i.e., maximum daily precipitation and total precipitation in the month, the results from all downscaling methods deviated from observation for winter and spring months.

In order to evaluate the performance of the ANN, KNN, and ANFIS models, three statistical criteria, namely, Nash–Sutcliffe model efficiency coefficient (NSE), relative mean absolute error (RMAE), and correlation coefficient (R) were used, and their results for the ANN, KNN, and ANFIS methods are presented in Tables 3–5, respectively.

It can be seen that for both HadCM3 and CanESM2 models, the performance of all downscaling methods, i.e., ANN, KNN, and ANFIS, is acceptable for our study region. However, with a slight difference, the KNN method has more accurate results. For example, for total precipitation in a month, the correlation values for the CanESM2 model are 0.805, 0.874, and 0.789 and for the HadCM3 model are 0.824, 0.835, and 0.784 for the ANN, KNN, and ANFIS methods, respectively. The values of RMAE are 0.36, 0.18, and 0.2 for the CanESM2 model and 0.24, 0.17, and 0.3 for the HadCM3 model, for the ANN, KNN, and ANFIS methods, respectively.

Based on Table 3, it has been seen that for all parameters, the correlation values (R) resulting from downscaling of HadCM3 model have better results compared to CanESM2 model. Also, by examining RMAE values, the error rate for HadCM3 in all parameters is less than the CanESM2 model. Finally, with regard to the NSE index, again, HadCM3 performed better compared to the CanESM2 model, i.e., had higher NSE values.

Table 4 contains the results of KNN downscaling method. It has been seen that for all parameters, R values for CanESM2 are higher than the HadCM3 model. Also, with regard to RMAE and NSE performance criteria, the CanESM2 model exhibits better results compared to the HadCM3 model, i.e., lower RMAE and higher NSE values.

Finally, from Table 5, one can see the downscaling performance of the ANFIS method. Again, based on all performance criteria, i.e., R, RMAE, and NSE indices, downscaling the outputs of the CanESM2 model caused better results compared to the HadCM3 model, i.e., higher values of R and NSE and lower values for RMAE.

As one can see, there is a small difference between the simulated and observed mean daily precipitation in all three downscaling methods. It is also evident that the simulation results of the ANN and especially the ANFIS model have a good performance when there is no noise in input data. Due to the considerable uncertainties in GCM outputs, the performance of the ANFIS model is

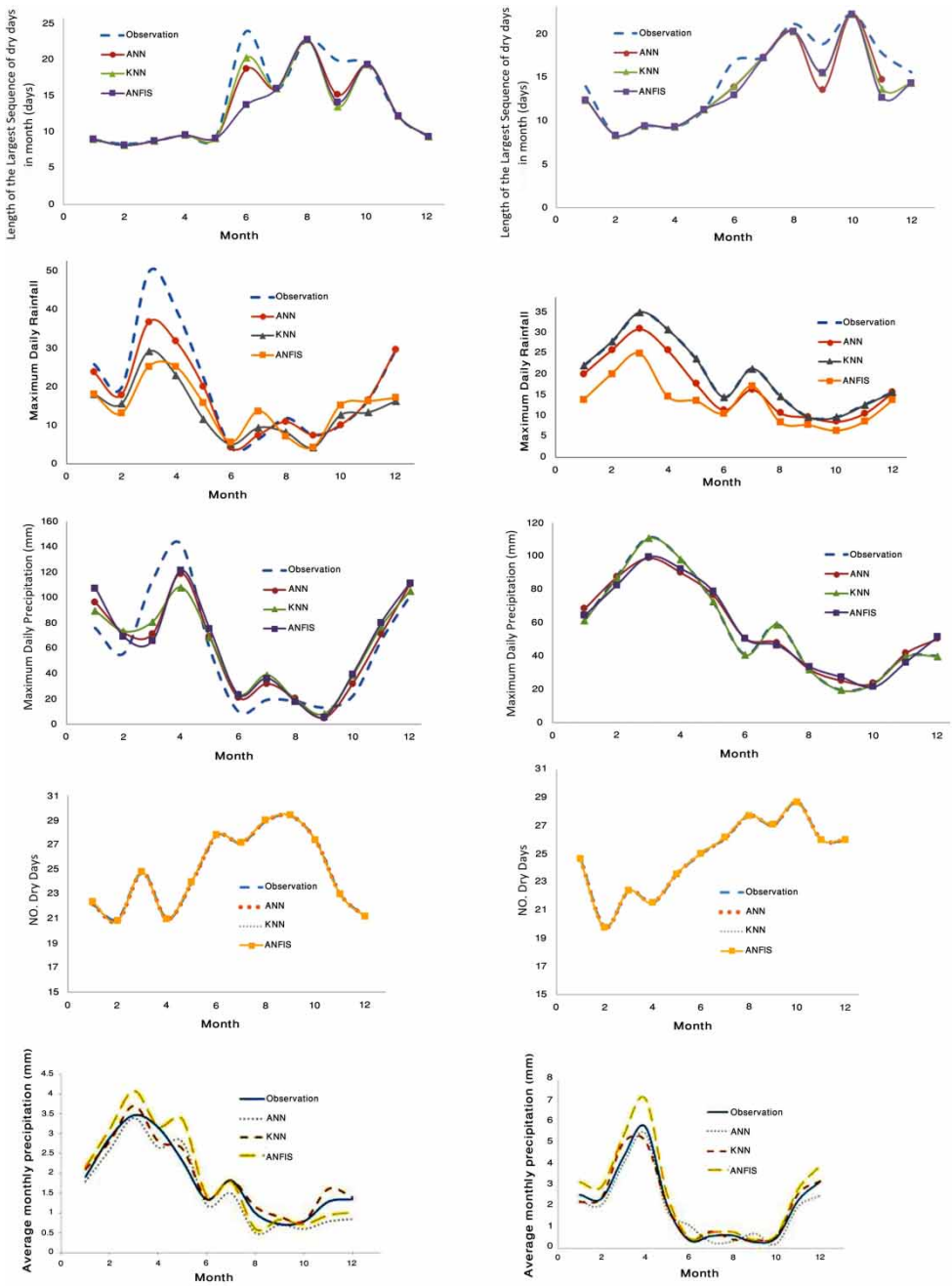

Figure 7 | Observed vs. simulated values for HadCM3 model calculated from ANN, KNN, and ANFIS models. The graphs on the right are related to the calibration and the left graphs, are related to the testing phases.

not as satisfactory as that of the KNN and ANN models. The credibility of the KNN model in noisy spaces has already been confirmed by other researchers (e.g., Eum *et al.* 2010). This is the case particularly for the maximum dry period in Figure 7 which is an extreme variable compared to other predictions.

CONCLUSION

The major goal of this research was to evaluate the performance of three data-driven models, namely, ANN, KNN, and ANFIS in downscaling the outputs of the CanESM2 model from the Fifth Assessment Report (AR5) and those of the

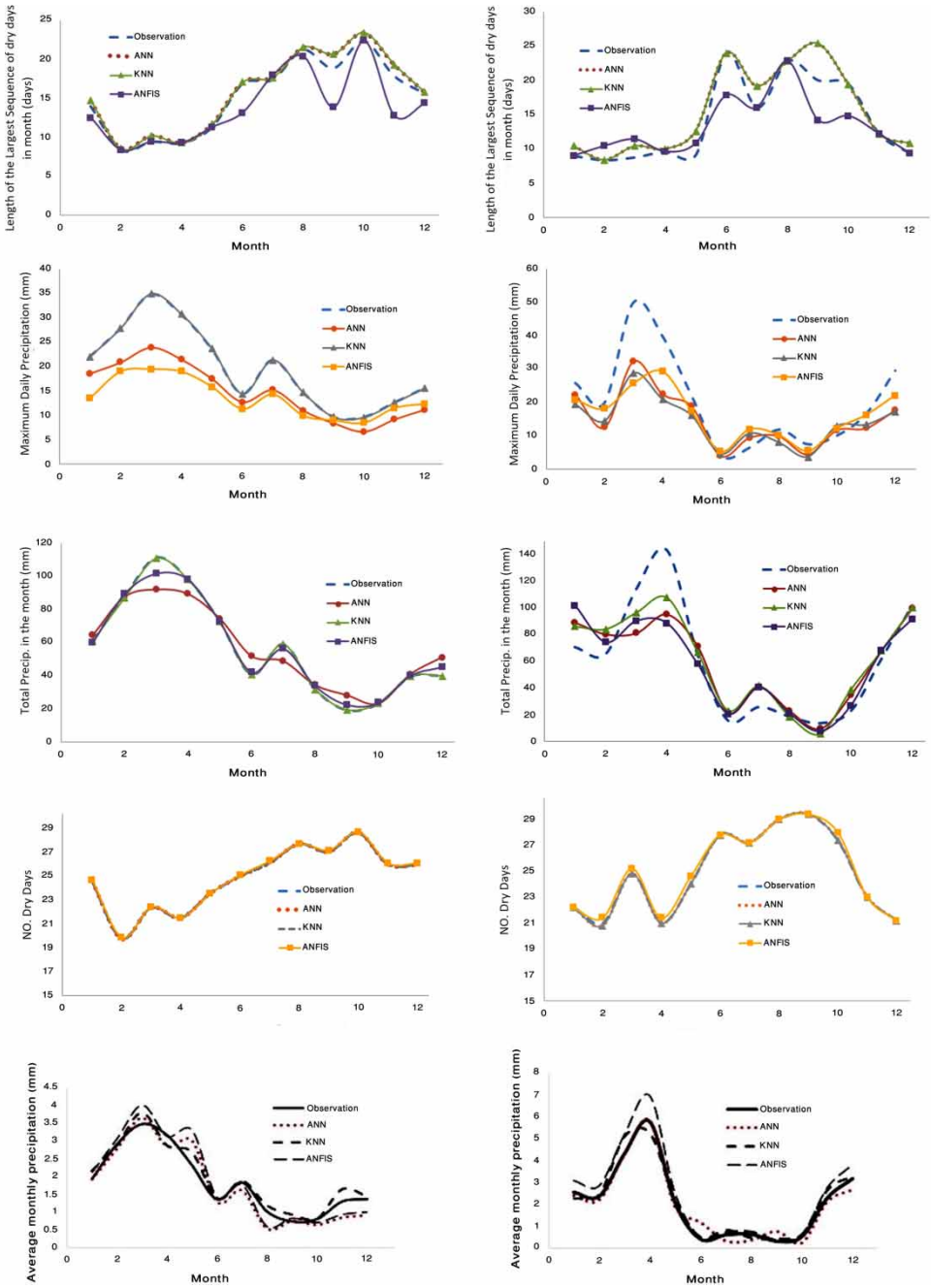

Figure 8 | Observed vs. simulated values for CanESM2 model calculated from ANN, KNN, and ANFIS models. The graphs on the left are related to the calibration and the right graphs are related to the testing phases.

HadCM3 model from the Fourth Assessment Report (AR4). The Amameh Basin located at the upstream of the Latyan Dam in northern Tehran was selected as the case study. With regard to the high number of input parameters for data-driven downscaling methods, the 26 outputs of the GCM models were prioritized using the PCA method.

With regard to the results, it is shown that except for the maximum daily precipitation in the month (mm) and total

Table 3 | Results of ANN method for the two models of HadCM3 and CanESM2

GCM model	CanESM2						HadCM3					
Evaluation criteria	R		RMAE		NSE		R		RMAE		NSE	
Calibration/Test	Test	Calibration	Test	Calibration	Test	Calibration	Test	Calibration	Test	Calibration	Test	Calibration
Length of the longest sequence of dry days in the month (day)	0.9082	0.989	0.314	0.152	0.832	0.968	0.920	0.869	0.02	0.036	0.854	0.791
Maximum daily precipitation in the month (mm)	0.898	0.963	0.56	0.073	0.966	0.9528	0.966	0.9515	0.04	0.06	0.966	0.941
Total precipitation in the month (mm)	0.804	0.935	0.36	0.111	0.831	0.9390	0.824	0.946	0.24	0.059	0.871	0.934
Number of dry days in the month (#)	0.98	0.98	0.084	0.015	0.97	0.98	0.98	0.98	0.0003	0.0002	0.98	0.97
Average monthly rainfall (mm)	0.91	0.959	0.1	0.08	0.954	0.919	0.9658	0.959	0.001	0.001	0.991	0.969

Table 4 | Results of KNN method for the two models of HadCM3 and CanESM2

GCM model	CanESM2						HadCM3					
Evaluation criteria	R		RMAE		NSE		R		RMAE		NSE	
Calibration/Test	Test	Calibration	Test	Calibration	Test	Calibration	Test	Calibration	Test	Calibration	Test	Calibration
Length of the longest sequence of dry days in the month (day)	0.920	0.9897	0.01	0.01	0.865	0.968	0.8919	0.9013	0.05	0.03	0.85	0.834
Maximum daily precipitation in the month (mm)	0.901	0.98	0.17	0.003	0.96	0.93	0.8776	0.97	0.21	0.003	0.921	0.945
Total precipitation in the month (mm)	0.874	0.98	0.18	0.003	0.97	0.95	0.8353	0.97	0.17	0.003	0.841	0.93
Number of dry days in the month (#)	0.98	0.98	0.0003	0.0002	0.98	0.97	0.98	0.98	0.0003	0.0002	0.98	0.97
Average monthly rainfall (mm)	0.9623	0.96	0.04	0.06	0.975	0.952	0.9686	0.976	0.0008	0.0005	0.966	0.97

Table 5 | Results of ANFIS method for the two models of HadCM3 and CanESM2

GCM model	CanESM2						HadCM3					
Evaluation criteria	R		RMAE		NSE		R		RMAE		NSE	
Calibration/Test	Test	Calibration	Test	Calibration	Test	Calibration	Test	Calibration	Test	Calibration	Test	Calibration
Length of the longest sequence of dry days in the month (day)	0.796	0.812	0.032	0.043	0.86	0.968	0.712	0.8569	0.06	0.03	0.84	0.82
Maximum daily precipitation in the month (mm)	0.863	0.945	0.117	0.113	0.94	0.89	0.765	0.8005	0.2	0.15	0.92	0.93
Total precipitation in the month (mm)	0.789	0.989	0.2	0.025	0.9	0.93	0.784	0.9375	0.3	0.07	0.84	0.92
Number of dry days in the month (#)	0.98	0.98	0.0042	0.0013	0.98	0.97	0.98	0.98	0.007	0.0004	0.98	0.97
Average monthly rainfall (mm)	0.954	0.963	0.01	0.04	0.981	0.966	0.941	0.9128	0.07	0.09	0.92	0.973

precipitation in the month (mm) parameters, other parameters, i.e., length of the longest sequence of dry days in the month (days), number of dry days in the month, and average monthly rainfall (mm) presented higher correlation for both GCM models. With regard to the results obtained in this research, i.e., the high accuracy of data-driven models in calibration phase and also the low rate of errors for the test period, it can be inferred that the application of all these data-driven models for downscaling the outputs of GCM models can be advised for our study area and also for other case studies. It was shown that among three data-driven methods, the KNN and ANN approaches presented better results compared with ANFIS, while KNN had the best performance.

For both the CanESM2 and HadCM3 GCM models, it was shown that the performance of all downscaling methods was satisfactory based on statistical indices, but it was shown that the CanESM2 model exhibited better performance compared to HadCM3. This could be related to the fact that CanESM2 is a newer GCM model with more updated data and modeling approaches. Finally, it can be concluded that downscaling the outputs of the CanESM2 general circulation model by the KNN approach provided the best results compared to all other combinations of GCM and downscaling methods.

According to the conducted literature survey, the ANFIS method has not been used so far for precipitation/temperature downscaling, although it has been one of the widely used data-driven models for simulation of hydrological variables. The accuracy of data-driven methods employed in this research can be compared with other statistical downscaling methods, e.g., LARS-WG, SDSM, SOGDS and also dynamic models over this study area and other regions.

Statistical downscaling of GCM data on climate change is built on the implicit assumption that the statistical relationships between the large-scale predictors and the local predictants would not be affected by climate change. On relatively short time scales (up to a few decades), which was the case for our study, this problem should not be too grave, as the anticipated (and GCM-simulated) scale of change is still of the order of the natural interannual and interdecadal variability. It should be noted that strong

nonlinearity in climate change, on the other hand, could crash any downscaling approach.

REFERENCES

Alemzadeh, M., Halavati, R., Bagheri Shouraki, S., Eshraghi, M. & Ziaie, P. 2004 A Novel Fuzzy Approach to Speech Recognition, Hybrid Intelligent Systems. In: *Fourth International Conference on Hybrid Intelligent Systems*, Kitakyushu, Japan.

Bring, A., Asokan, S., Jaramillo, F., Jarsjö, J. & Prieto, C. 2015 Implications of freshwater flux data from the CMIP5 multimodel output across a set of Northern Hemisphere drainage basins. *Earth's Future* **3** (6), 206–217.

Bustami, R., Bessaih, N., Bong, C. & Suhaili, S. 2007 Artificial Neural Network for precipitation and water level predictions of Bedup River. *IAENG International Journal of Computer Science* **34** (2), 228–233.

Campozano, L., Tenelanda, T., Sanchez, E., Samaniego, E. & Feyen, J. 2016 Comparison of statistical downscaling methods for monthly total precipitation: case study for the Paute River basin in southern Ecuador. *Advances in Meteorology* **2016**, 6526341, 13.

Djamil, Y. S. & Aldrian, E. 2008 Application of multivariate anfis for daily rainfall prediction: influences of training data size. *Makara Seri Sains* **12** (1), 7–14.

Duhan, D. & Pandey, A. 2015 Statistical downscaling of temperature using three techniques in the Tons River basin in Central India. *Theoretical and Applied Climatology* **121**, 605.

Emamgholizadeh, S., Moslemi, K. & Karami, G. 2014 Prediction the groundwater level of bastam plain (Iran) by artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). *Water Resour. Manage.* **28**, 5433. https://doi.org/10.1007/s11269-014-0810-0.

Eum, H. I., Simonovic, S. P. & Kim, Y. O. 2010 Climate change impact assessment using K-nearest neighbor weather generator: case study of the Nakdong River basin in Korea. *Journal of Hydrologic Engineering* **15** (10), 772–785. DOI: 10.1061/(ASCE)HE.1943-5584.0000251.

Farzaneh, M. R., Eslamian, S., Samadi, S. Z. & Akbarpour, A. 2012 An appropriate general circulation model (GCM) to investigate climate change impact. *International Journal of Hydrology Science and Technology* **2** (1), 34–47.

Fu, G., Charles, S. & Kirshner, S. 2012 Daily rainfall projections from general circulation models with a downscaling nonhomogeneous hidden Markov model (NHMM) for south-eastern Australia. *Hydrological Processes* **27** (25), 3663–3673.

Hesami, M. & Zeynolabedini, M. 2016 Forecasting of precipitation using statistical downscaling from outputs of CanESM2 model. In: *Second International Conference in New Research on Civil*, Architectural & Urban Management, Bangkok.

IPCC 2007 *Synthesis Report 2007: AR4*. Cambridge University Press, Cambridge, UK & New York, USA.

IPCC 2008 Towards new scenarios for analysis of emissions, climate change, impacts, and response strategies. In: *IPCC Expert Meeting Report on New Scenarios, Noordwijkerhout, Intergovernmental Panel on Climate Change*.

Jang, J.-S. R., Sun, C.-T. E. & Mizutani, E. 1997 *Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Pearson Education, London, UK.

Jones, P., Thornton, P. & Heinke, J. 2009 *Generating Characteristic Daily Weather Data Using Downscaled Climate Model Data From the IPCC's Fourth Assessment*. Project report. ILRI, Nairobi, Kenya.

Karamouz, M., Nazif, S. & Zahmatkesh, Z. 2013 Self-organizing Gaussian-based downscaling of climate data for simulation of urban drainage systems. *Journal of Irrigation and Drainage Engineering* **139** (2), 98–112.

Karlsson, M. & Yakowitz, S. 1987 Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. *Water Resources Research* **23** (7), 1300–1308.

Lall, U. & Sharma, A. 1996 A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research* **32** (3), 679–694.

Mahalanobis, P. C. 1936 On the generalised distance in statistics. *Proceedings of the National Institute of Science India* **2** (1), 49–55.

Mahani, M. 2015 Assessment of the effects of climate change on the discharge of the Ploud river using artificial neural network. In: *International Conference on Engineering*, Nagoya, Japan. Science and Technology.

Noori, R., Karbassi, A. R., Moghddamnia, A., Han, D., Ashtiani, H., Farokhnia, A. & Ghafari Gousheh, M. 2011 Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream flow prediction. *Journal of Hydrology* **401** (3–4), 177–189.

Rouhani, H., Bahleke, M. & Fathabadi, A. 2016 Studying of temperature changes using the application of statistical downscaling models and outputs of CanESM2 model. In: *National Conference of Knowledge and Technology of Agricultural Science*, Natural Resources and Environment of Iran.

Samadi, S. Z., Sagareswar, G. & Tajiki, M. 2010 Comparison of general circulation models: methodology for selecting the best GCM in Kermanshah Synoptic Station, Iran. *International Journal of Global Warming* **2** (4), 347–365.

Satish, B., Swarup, K. S., Srinivas, S. & Hanumantha Rao, A. 2004 Effect of temperature on short term load forecasting using an integrated ANN. *Electric Power Systems Research* **72**, 95–101.

Schoof, J. & Pryor, S. 2001 Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks. *International Journal of Climatology* **21** (7), 773–790.

Shahverdi, F., Ahmadi, M., Avazmoghadam, S. & Shahverdi, M. 2017 Modeling of Ni(II) Separation from Aqueous Solutions onto

Aspergillus awamori using Artificial Neural Technique. In: *Fourth International Conference on Planning and Management*, Faculty of Environment, University of Tehran, Tehran, Iran.

Shashikanth, K. & Ghosh, S. 2013 Fine resolution Indian summer monsoon rainfall projection with statistical downscaling. *International Journal of Chemical, Environmental & Biological Sciences* **1** (4), 615–618.

Syed, C., Soltani, A. & Gholipour, M. 2012 Simulation of the effect of climate change on growth, yield and water consumption of chickpea. *Journal of Agricultural Sciences and Natural Resources* **2**, 131–140.

Talei, A., Lloud, H. C. C. & Wong, T. S. W. 2010 Evaluation of rainfall and discharge inputs used by Adaptive Network-based Fuzzy Inference Systems (ANFIS) in rainfall–runoff modeling. *Journal of Hydrology* **391** (3–4), 248–262.

Trafalis, T. B., Santosa, B. & Richman, M. B. 2005 Learning networks in rainfall estimation. *CMS* **2**, 229–251.

Tue Vu, M., Aribarg, T., Supratid, S., Raghavan, S. & Liong, S. 2016 Statistical downscaling rainfall using artificial neural network: significantly wetter Bangkok. *Theoretical and Applied Climatology* **126** (3–4), 453–467.

Verbist, K., Robertson, A. W., Cornelis, W. M. & Gabriels, D. 2010 Seasonal predictability of daily rainfall characteristics in Central Northern Chile for dry-land management. *Journal of Applied Meteorology and Climatology* **49**, 1938–1955.

Wang, Z. L. & Sheng, H. H. 2010 Rainfall Prediction Using Generalized Regression Neural Network: Case study Zhengzhou. In: *International Conference on Computational and Information Sciences*, 17–19 December, pp. 1265–1268.

Wu, W., Liu, Y., Descomes, G., Ge, M., Warner, T., Swerdlin, S., Rostkier-Edelstein, D., Kunin, P. & Givati, A. 2010 *Application of A K-Nearest Neighbor Simulator for Seasonal Precipitation Prediction in A Semiarid Region with Complex Terrain*. EGU General Assembly 2–7 May, Vienna, Austria, p. 5237.

Xu, C.-H. & Xu, Y. 2012 The projection of temperature and precipitation over China under RCP scenarios using a CMIP5 multi-model ensemble. *Atmospheric and Oceanic Science Letters* **5** (6), 527–533.

Yates, D., Gangopadhyay, S., Rajagopalan, B. & Strzepek, K. 2003 A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resources Research* **39** (7), 1114–1121.

Zohdi, R. 2000 Industrial applications of logic and fuzzy neural networks. *Isiran Institute Publications* **2000**, 426 (in Persian).

First received 28 November 2017; accepted in revised form 6 August 2018. Available online 3 September 2018