Conjunction of wavelet-entropy and SOM clustering for multi-GCM statistical downscaling

Aida Hosseini Baghanam, Vahid Nourani, Mohammad-Ali Keynejad, Hassan Taghipour and Mohammad-Taghi Alami

ABSTRACT

Important issues in statistical downscaling of general circulation models (GCMs) is to select dominant large-scale climate data (predictors). This study developed a predictor screening framework, which integrates wavelet-entropy (WE) and self-organizing map (SOM) to downscale station rainfall. WEs were computed as the representatives of predictors and fed into the SOM to cluster the predictors. SOM-based clustering of predictors according to WEs could lead to physically meaningful selection of the dominant predictors. Then, artificial neural network (ANN) as the statistical downscaling method was developed. To assess the advantages of different GCMs, multi-GCM ensemble approach was used by Can-ESM2, BNU-ESM, and INM-CM4 GCMs. Moreover, NCEP reanalysis data were used to calibrate downscaling model as well for comparison purposes. The calibration, validation, and projection of the proposed model were performed during January 1951 to December 1991, January 1992 to December 2005 and January 2017 to December 2100, respectively. The proposed data screening model could reduce the dimensionality of data and select appropriate predictors for generalizing future rainfall. Results showed better performance of ANN than multiple linear regression (MLR) model. The projection results yielded 29% and 21% decrease of rainfall at the study area for 2017–2050 under RCPs 4.5 and 8.5, respectively.

Key words | artificial neural networks (ANN), general circulation models, mutual information (MI), self-organizing map (SOM), Tabriz city rainfall, wavelet-entropy (WE)

Aida Hosseini Baghanam (corresponding author)
Vahid Nourani
Mohammad-Taghi Alami
Department of Water Resources Engineering,
Faculty of Civil Engineering,
University of Tabriz,
Tabriz,
Iran
E-mail: *hosseinibaghanam@gmail.com*

Vahid Nourani
Department of Civil Engineering,
Near East University,
P.O. Box: 99138, Nicosia, North Cyprus, Mersin 10,
Turkey

Mohammad-Ali Keynejad
Faculty of Civil Engineering,
Sahand University of Technology,
Sahand,
Iran

Hassan Taghipour
Department of Environmental Health Engineering,
Faculty of Public Health,
Tabriz Medical Science University,
Tabriz,
Iran

INTRODUCTION

Rainfall as the major component of the hydrologic cycle deposits most of the fresh water on the earth and can lead to disasters such as floods, erosion, sedimentation, and surface and groundwater contamination. The role of precipitation in water resource management cannot be disregarded, therefore, an extensive range of studies on the issue is being carried out, of which, the basic requirement of such studies is rainfall data. Access to high-resolution rainfall data can lead to more reliable studies; however, availability of high-resolution rainfall data is limited due to large spatiotemporal variation of rainfall as well as restrictions of cost, technical capability, and scarcity of hydrological stations. In this way, general circulation models (GCMs) can be considered as reliable hydro-climatologic data sources. GCMs use physical-based equations on various processes of the atmosphere and ocean for simulating the response of global climate system against increasing greenhouse gas concentrations (IPCC 2013). The skill of GCMs in reproducing hydrologic parameters such as precipitation is of critical importance. Although GCMs are tools that provide reliable atmospheric data, coarse spatial resolution of GCMs may lead to their poor applicability as

doi: 10.2166/nh.2018.169

the input to local-scale hydrologic models. Downscaling is an approach to obtain local-scale weather data from large-scale GCMs. The two methods of downscaling are dynamical and statistical downscaling. Dynamical downscaling is a method to derive smaller-scale climatic information over a bounded area via a high-resolution regional model driven by boundary conditions from GCMs. Statistical downscaling encompasses statistically relating large-scale climate features (predictors), to local climate (predictand) data (Wilby & Wigley 1997); in this way, physical characteristics of the study area can be involved in downscaling by extracting the statistical pattern between predictors and predictand, which indeed demonstrates a black box modeling.

In spite of high-resolution outputs of dynamical downscaling, intensive computational procedures due to the numerical solution of motion and thermodynamic equations, use of physical principles to reproduce local climates, the need for large volumes of data, and a high level of expertise to implement and interpret results, limit the application of dynamical downscaling (Trzaska & Schnarr 2014; Danandeh Mehr & Kahya 2016). Instead, in statistical downscaling techniques, the purpose is to find statistical relations between GCM and local weather data without needing any physical knowledge of the region. Therefore due to the convenience in implementing and interpreting results, statistical downscaling techniques have been used widely in several studies (e.g., Sailor & Li 1999; Olsson *et al.* 2001, 2004; Harpham & Wilby 2005; Chen & Adams 2006; Sousa *et al.* 2007; Beecham *et al.* 2014). The main concept of statistical downscaling is to make a relationship between predictors and predictand by means of a statistical method such as: (i) linear regression models, e.g., statistical downscaling model (SDSM) (Wilby *et al.* 2002); (ii) nonlinear regression models, e.g., artificial neural network (ANN) (Zorita & Von Storch 1999), support vector machine (SVM) (Tripathi *et al.* 2006), relevance vector machines (RVM) (Ghosh & Mujumdar 2008) gene expression programming (GEP) (Sachindra & Perera 2016); (iii) weather generators, e.g., Long Ashton Research Station-Weather Generator (LARS-WG) (Racsko *et al.* 1991).

Trzaska & Schnarr (2014) prepared a thorough review of downscaling methods for climate change projections, which can be referred to in order to study more about downscaling methods.

Among nonlinear regression techniques, ANN has high potential to extract complex patterns and relations between predictors and predictand. The capability of ANN in simulating the nonlinear, and time-varying characteristics of atmospheric variables at different scales, leads to several successful applications of ANNs in downscaling issues within the literature (e.g., Wilby & Wigley 1997; Dibike & Coulibaly 2006; Chadwick *et al.* 2011; Okkan & Fistikoglu 2014; Okkan & Kirdemir 2016).

Focus on the consequences of studies with ANN-based downscaling methods shows contradictory results, with some studies stating the superiority, while others denote the drawbacks and inefficiency of ANN in downscaling GCM data (e.g., Dibike & Coulibaly 2006; Khan *et al.* 2006; Tisseuil *et al.* 2010; Abdellatif *et al.* 2013). These inconsistent results can depend on the quality and quantity of the applied data. Huge data sets available by GCMs are one of the issues that has become a major challenge in ANN-based modeling. Redundant information and the involved noise in data are the other challenging issues concerning ANN, because the noises might be magnified while using nonlinear models such as ANN. In this case, application of input screening as a pre-processing scheme can largely enhance the efficiency of the ANN-based downscaling model.

Input data screening is an important step in any data-driven modeling and which usually improves the modeling performance. Since different researchers have already proved that the application of input data screening approaches can enhance the efficiency of data-based models such as ANN (Nourani *et al.* 2017), it is expected that application of a robust input data screening technique can increase the efficiency of ANN in downscaling problems as well. In statistical downscaling problems, the main concern is encountering huge data sets relevant to several large-scale climate variables with long historical time series (i.e., temporal variation) at diverse grid points (i.e., spatial variation). Generally, such large input data sets may lead to decreasing ANN performance, thus implementing input data screening on GCM data by reduction of the input data size, or in other words, dominant input selection among GCM variables might be useful.

In downscaling subjects, apart from different climate variables at multiple grid points over long time intervals,

selection of effective GCMs among several GCMs is the other factor to manifold the data dimensionality. Although various research centers around the globe prepare GCM-based climate data, generally they do not offer the same values for a specific variable in a particular region; the reason is due to different parameterization schemes, variation in boundary layers and different resolutions, relying on the output from a single model, and finally, abrupt spatial variability in climate (Khan & Pilz 2017). Thus, proper selection of the GCM is of prime importance in downscaling approach (Lee & Kim 2017). In this way, the question is which variables and GCMs should be included in a SDSM.

In spite of developing new GCM parameter selecting techniques (Khan & Pilz 2017) and extensive research projects which aim to standardize and coordinate climate models (i.e., the Coupled Modelling Inter comparison Project (CMIP5: http://cmip-pcmdi.llnl.gov/cmip5/) and the Paleoclimate Modelling Inter comparison Project (PMIP3: https://pmip3.lsce.ipsl.fr/)) (Varela *et al.* 2015), general approaches such as correlation analysis (Devak & Dhanya 2014), principal component analysis (PCA) (Ahmadi & Han 2013; Chuang *et al.* 2016), fuzzy and Gamma test (Ahmadi *et al.* 2015) have been employed for input selection of statistical downscaling models. Hence, a robust method of dominant input selection among various GCMs and several climate variables is needed, which the current study tries to address.

One of the effective methods to decrease the dimensionality of input space is clustering and selecting a representative member from each cluster (Bowden *et al.* 2005). Although several studies have already used different clustering based methods in dominant input selection of ANN models in hydrological applications (e.g., May *et al.* 2008; Markus *et al.* 2010; Nourani & Parhizkar 2013; Li *et al.* 2015; Nourani *et al.* 2017), the implementation of such a clustering based input screening is indeed scarce in statistical downscaling of GCMs. The advantage of the proposed input screening model is to select a comprehensive group of variables in generalizing future rainfall while the whole data domain contributes in selecting the dominant predictor, since an appropriate clustering method categorizes data space into homogenous clusters with similar characteristics and extracts feature from the whole domain of observed data even with moderate or low relevancy criteria.

In the current study, the self-organizing map (SOM) as a robust clustering method which is fed by wavelet-entropies (WEs), classifies the features of large-scale climate variables to select dominant inputs. SOMs are often described as a type of neural network; they are more easily understood as a type of constrained k-means. Various studies indicated the superiority of SOM to classical method such as k-means (Openshaw & Openshaw 1997; Bação *et al.* 2005; Hsu & Li 2010; Nourani & Parhizkar 2013). Some examples of SOM advantages to k-means are: the k-means method is sensitive to initialization which SOM is not; the k-means gradient orientation forces a premature convergence which, depending on the initialization, may frequently yield local optimum solutions while SOM is less prone to local optima than k-means and can deal effectively with problems having multiple optima; the performance of the K-mean algorithm largely depends on not only the number but also the placement of the initial codebooks. This has been a limiting factor in the application of the k-mean algorithm, while the SOM mapping tries to preserve topological relations, i.e., patterns that are close in the input space will be mapped to units that are close in the output space, and vice versa. To allow an easy visualization, SOM offers the opportunity for an early exploration of the search space, and as the process continues it gradually narrows the search. By the end of the search process (providing the neighborhood radius decreases to zero) the SOM is exactly the same as k-means, which allows for a minimization of the distances between the observations and the cluster centers.

Since climatologic data sets involve non-stationary time series, entropy as a measure of information content (Shannon 1948) could be preferred to statistical moments (i.e., mean, variance, skewness, etc.) in order to represent whole time series in the clustering procedure. On the other hand, underlying multi-resolution seasonality of the process is drawn out by wavelet transform via extracting various features of time series at different time scales (Johnson *et al.* 2011; Nourani *et al.* 2014; Rashid *et al.* 2015).

The proposed novel input data screening method is incorporated into ANN-based statistical downscaling of GCMs in order to reduce redundant information of input data set for projection of monthly rainfall of Tabriz station located in Iran. Furthermore, to enjoy effective

characteristics of each GCM in future rainfall simulation, an ensemble of three GCMs (i.e., Can-ESM2, BNU-ESM, and INM-CM4 GCMs) is used in the proposed downscaling. The ensemble of GCMs has been recommended by some other studies in order to cover the uncertainties involved in each GCM (Li *et al.* 2012; Miao *et al.* 2014; Rashid *et al.* 2015). Since various GCMs are developed according to different structures, the climate downscaling and projection using various GCMs can lead to diverse results (Sachindra *et al.* 2014). Application of a multi-GCM approach can lead to efficient results since there is the opportunity to select dominant variables among various GCMs with different structural inherence than being forced to select from one specific GCM. In this way, ANN downscaling was performed by using ensemble of GCMs and reanalysis data (i.e., NCEP).

This article is organized in four sections as follows: the section below describes the study area as well as the applied data, including the station and GCM data; furthermore, the applied methodology is included at the last part of this section. This is followed by a results and discussion section, and the final section provides concluding remarks.

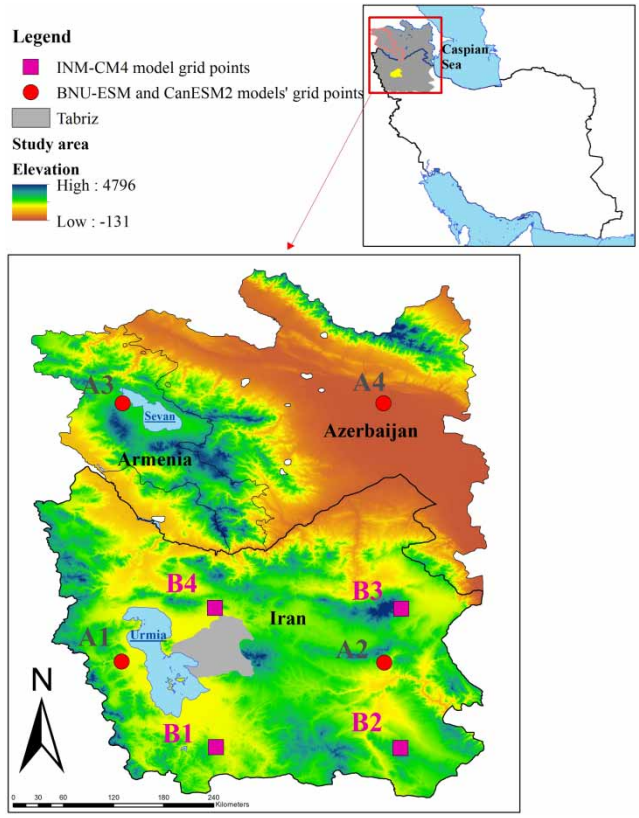

Figure 1 | Study area with the grid points of BNU-ESM, Can-ESM (i.e., A1, A2, A3, A4) and INM-CM4 (i.e., B1, B2, B3, B4) GCMs.

MATERIAL AND METHODS

Study area and data set

Tabriz City (latitude: 38°08′*N*, longitude: 46°29′*E*) is the capital city of East Azerbaijan province in the north-west of Iran and is located in the valley of a seasonal river (Figure 1). The city lies on the Tabriz plain which has a moderate slope and at 60 km west ends on the east bank of Urmia Lake. The elevation of the city varies between 1,350 and 1,600 meters above sea level. The city's weather is semi-arid with regular seasons. The annual precipitation of Tabriz is around 280 millimeters, mostly falling in wet seasons (i.e., October to April). During the winter, precipitation is snow and during spring and fall it rains. Overall, the city's weather is mild and fine in spring, dry and semi-hot in summer, humid and rainy during fall, and cold with snowfall in winter. The average annual temperature is 12.6 °C. Cool winds blow from east to west, mostly in summer.

In order to project large-scale GCM data to local-scale station data, Tabriz hydrologic observational station was considered. The monthly rainfall data of the station during January 1951 to December 2016 prepared by the Meteorological Organization of East Azerbaijan were used in the current study. In order to develop the proposed downscaling model, large-scale historical climate data were extracted during the period January 1951–December 2005 from Can-ESM2, BNU-ESM, and INM-CM4 GCMs developed respectively in research centers of Canada, China, and Russia with mean monthly atmospheric variables (Table 1). Since several studies described the beneficial effects in applying several grid points around the study location (Frost *et al.* 2011; Guo *et al.* 2012; Beecham *et al.* 2014), predictors on four grid points around the study station were adopted in this study as well (Figure 1). The four closest grid points from each GCM around the Tabriz station were considered as the potential grid points based on the grid size of each GCM. But according to the proposed

Table 1 | Characteristics of applied GCMs

Centre	Centre acronym	Model	RCP (W/m^2)	Grid size (approximately)	Predictor number	Applied climate variables in GCMs	Pressure levels (Pa)
Beijing Normal University, China	BNU	BNU-ESM	RCP4.5; RCP8.5	$2.81^\circ \times 2.81^\circ$	380	[a]ua: eastward wind; [a]va: northward wind; [a]zg: geo-potential height; [a]hur: relative humidity; [a]hus: specific humidity; tas: air temperature; uas: eastward near-surface wind; vas: northward near-surface wind; psl: air pressure at sea level; hfls: surface upward latent heat flux; prc: convective precipitation flux; pr: precipitation flux; hurs: near-surface relative humidity; huss: near-surface specific humidity; evspsbl: water evaporation flux	[b]100; [b]200; [b]300; [b]500; [b]700; 1,000; 2,000; 3,000; 5,000; 7,000; 10,000; 15,000; 20,000; 25,000; 30,000; 40,000; 50,000; 60,000; 70,000; 85,000; 92,500; 100,000
Canadian Centre for Climate Modelling and Analysis, Canada	CCCma	Can-ESM2	RCP4.5; RCP8.5	$2.81^\circ \times 2.81^\circ$	480		
Russian Academy of Sciences, Institute of Numerical Mathematics, Russia	INM	INM-CM4	RCP4.5 RCP8.5	$1.5^\circ \times 2^\circ$	380		

[a]Variables which vary at pressure levels.
[b]Pressure levels of Can-ESM2 model in addition to others.

methodology, effective climate variables and consequently effective grid points were selected as the dominant grid points. As Figure 1 shows, there are grid points A1, A2, A3, A4 relevant to BNU and CAN GCMs which coincide spatially and B1, B2, B3, B4 relevant to INM GCM. In order to validate GCM-based downscaling, monthly reanalysis data sets of the National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) (termed NCEP reanalysis afterwards in this study) were collected from the National Oceanic and Atmospheric Administration/Earth System Research Laboratory (NOAA/ESRL). The resolution of NCEP data is $2.5^\circ \times 2.5^\circ$, which, in order to match GCMs, the GCM outputs were linearly interpolated over the NCEP grid points.

It is noted that the Fifth Assessment Report (AR5) of the United Nations Intergovernmental Panel on Climate Change (IPCC) under the representative concentration pathways (RCPs), i.e., RCP4.5 and RCP8.5 scenarios for future simulation, was considered in this study. The applied data sets in the current study were collected from the IPCC data distribution center (http://cera-www.dkrz.de/) for the period 1951–2100.

Table 1 shows the list of available variables for each considered GCM. Ninety-five large-scale atmospheric variables at each grid point of BNU-ESM and INM-CM4 models and 120 variables at each grid point of the Can-ESM2 model were considered, so in total 1,240 variables in three GCMs are the potential input candidates of the downscaling model. In the case of trial and error procedure, $2^{1240} - 1$ cases should be examined which is an exhausting process and not a feasible technique to select the dominant inputs. Although linear correlation coefficient may be used as a criterion to select dominant inputs, nonlinear input selection method is essential for a nonlinear process as judged by Nourani *et al.* (2015). In this regard, nonlinear input screening method was addressed in this study to find the significant inputs of ANN-based downscaling model.

Proposed methodology

In order to downscale GCM data, an ANN-based statistical method was applied, and to improve the efficiency of the ANN model, pre-processing on variables of GCMs over

the four nearest grid points around the studied station was performed. The most effective GCMs in this study were selected according to the results of Cai *et al.* (2009). Selection of appropriate predictors was the next step to statistically downscale rainfall. Since involving a full set of potential variables simultaneously in a downscaling model can negatively impact the outcomes due to redundant information, pervasive assessments on selecting particular predictors becomes necessary due to the lack of general guidelines. Important predictor selection among three GCMs was implemented in this study using a newly proposed method involving WE and SOM tools. In this way, the ensemble of three different GCMs was performed in the screening step before downscaling. The ensemble idea in the current study was ensemble in input not outputs, therefore all the considered predictors in Table 1 from four grid points of three GCMs (i.e., in total 12 sources of predictors) were integrated into the screening methodology in order to select the important variables which impact on Tabriz rainfall generation. Since the selected variables belonged to different GCMs at diverse grid points, it was named multi-GCM ensemble procedure.

The proposed methodology considers three stages to achieve statistical downscaling of predictand. According to the presented schematic diagram in Figure 2, the first step is dominant input selection procedure, the second step is ANN-based downscaling model, and finally, the third step is projection of future rainfall at Tabriz hydrological station according to a multi-input ensemble ANN model under RCPs 4.5 and 8.5.

First step

Climatic variables due to non-stationary fluctuations inherent in climate phenomena encompass temporal features as well as seasonal attributes, and wavelet transform can extract such features. Therefore, discrete wavelet transform (DWT) was applied to decompose time series of large-scale climate variables into multi-resolution sub-series. In this way, large- and small-scale hidden features of time series were broken up into approximation and detailed sub-series to represent general trend and different levels of periodicity involved in the time series. Dominant feature extraction among such huge data sets including numerous sub-series would lead to poor outcomes due to decrease of accuracy caused by the imposed noises. Therefore, it is suggested to compute and use entropy of each sub-series as the representatives. Entropy shows information content or the degree of order/disorder of WT decomposed time series and is designated as WE here.

In this way, each sub-series is substituted by only one number (i.e., WE) so that the dimensionality of data domain is reduced greatly. In more detail, monthly time series with many components first is decomposed to $l+1$ sub-series via wavelet transform at level l and then $l+1$ WEs are computed to represent whole time series with only $l+1$ numbers, each representing a specific feature of original time series. In order to use WEs in the clustering tool, vectors of $l+1$ components including WEs relevant to approximation and decomposed time series were established as the feature of each predictor variable. Generally, all of the features are not informative equally, some may include noise, some others may have correlation, and some have no remarkable relevancy to the output variable. Therefore, screening features are necessary to select the impressive features. SOM as an ANN-based clustering method was used here to screen the inputs. Basically, SOM is an unsupervised ANN tool that can capture linear and nonlinear statistical relations in complex high dimensional data sets and illustrate them in a comprehensible geometric relation map. SOM-based clustering was implemented over computed WEs of three GCM variables. Application of WEs instead of whole decomposed sub-series in the clustering procedure can lead to optimizing the SOM input layer and improve clustering performance by reducing the dimensionality of the data set.

Due to the complexity of climate variables, poor linear correlation between predictors and predictand is probable in spite of strong nonlinear relation. In this regard, the mutual information (MI) concept was used as a measure to detect nonlinear relations between predictors at clusters and predictand. Finally, the variables which passed the screening procedure were considered to be used in the rainfall downscaling model via ANN.

Second step

In the second step, the ANN-based downscaling model was developed. In this way a multi-GCM ensemble model was

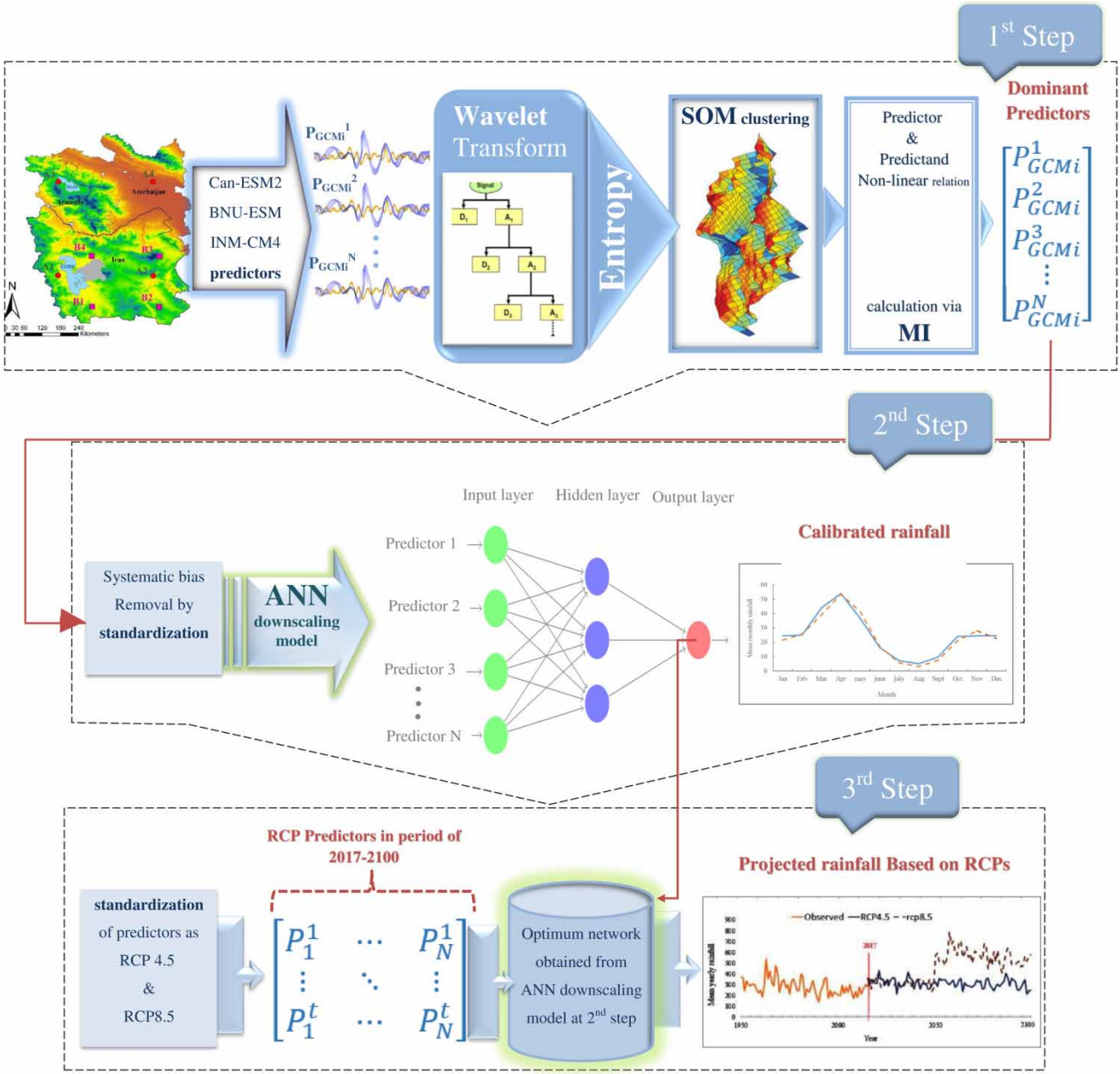

Figure 2 | Schematic of proposed methodology to screen predictor (first step), downscale GCM data (second step) and project future rainfall (third step). At the first step, P^N_{GCMi} denotes the *N*th variable belong to GCMi, *i* shows three applied GCMs. At the second step, predictor *N* denotes *N* dominant variables which were selected at the first step. At the third step, the matrix denotes the input data to the simulation model, where, P^t_N denotes the monthly value of dominant variable at RCP 4.5 or 8.5, *t* shows date which starts from 2017 to 2100, and *N* displays the dominant variables selected at the first step.

trained according to screened variables of three GCMs at four surrounding grid points of the study area to calculate Tabriz station rainfall values. NCEP-derived reanalysis data were also used to calibrate the ANN downscaling model. Since application of GCM-derived data in downscaling has shown systematic biases between the downscaled results and observations, statistical amendment of GCM data is often essential (Wilby *et al.* 2004). One of the common procedures to remove the systematic biases in mean and variance is standardization (Tripathi *et al.* 2006;

Chen *et al.* 2010; Acharya *et al.* 2013). According to Sachindra *et al.* (2014), standardization of predictors (i.e., subtracting the mean from data and dividing by the standard deviation) scales down the predictor data to a single uniform scale and removes the units of the variables.

Third step

Finally, stage three was performed under RCPs 4.5 and 8.5 to simulate future monthly rainfall of Tabriz hydrological station for near and distant future during the periods 2017–2050 and 2051–2100. It is noted that according to various plausible viewpoints about future anthropogenic actions due to different rates of growth in population, economic, energy, and socioeconomic development and their impact on greenhouse gas emissions, each RCP demonstrates distinct radiative forcing pathways by 2100. In this way, RCPs 4.5 and 8.5 are related to intermediate and high emission scenarios, respectively.

The sections below briefly explain the required tools for the proposed methodology.

Wavelet-entropy

The **seasonal** pattern encompassed in climate time series are captured with WT by decomposing the time series into subseries with different periods. The wavelet makes localization of a signal in time and scale domain by comparing the relationship between wavelet function and signal. Here, in this study, discrete form of WT is used and screened as Equation (1) (Mallat 1998):

$$g_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} g^*\left(\frac{t - nb_0a_0^m}{a_0^m}\right) \quad (1)$$

where * is the complex conjugate and $g(t)$ named wavelet function or mother wavelet. Different types of mother wavelets are *ciof2*, Daubechies (db) family, sym3 Meyer, etc. (Mallat 1998); m and n denote the wavelet dilation and translation, respectively; a_0 is a specified fined dilation step greater than 1; and b_0 is the location parameter which must be greater than zero. The most common and simplest choice for parameters are $a_0 = 2$ and $b_0 = 1$. This power-of-two logarithmic scaling of the dilation and translation is known as the dyadic grid arrangement. The dyadic wavelet can be written in more compact notation as (Mallat 1998):

$$g_{m,n}(t) = 2^{-m/2} g(2^{-m}t - n) \quad (2)$$

For a discrete time series, x_i, the dyadic WT becomes (Mallat 1998):

$$T_{m,n} = 2^{-m/2} \sum_{i=0}^{N-1} g(2^{-m}i - n)x_i \quad (3)$$

where $T_{m,n}$ corresponds to wavelet coefficient of the discrete wavelet at scale $a = 2^m$ and location $b = 2^m n$. Equation (3) considers a finite time series, x_i, $i = 0, 1, 2, \ldots, N = 1$, and $N = 2^q$. This gives the ranges of m and n as, respectively, $0 < n < 2^{q-m} - 1$ and $1 < m < q$.

The inverse discrete transform is given by (Mallat 1998):

$$x_i = \bar{T} + \sum_{m=1}^{q} \sum_{n=0}^{2^{q-m}-1} T_{m,n} 2^{-m/2} g(2^{-m}i - n) \quad (4)$$

or in a simple format as (Mallat 1998):

$$x_i = \bar{T}(t) + \sum_{m=1}^{q} W_m(t) \quad (5)$$

In which $\bar{T}(t)$ is called approximation sub-signal at level q and $W_m(t)$ are detail sub-signals at levels $m = 1, 2, \ldots, q$.

The wavelet coefficients, $W_m(t)$ ($m = 1, 2, \ldots, q$), provide the detail signals, which can capture small features of interpretational value in the data; the residual term, $\bar{T}(t)$, represents the background information of data which is named approximation signal.

Entropy is a statistical measure of the randomness or uncertainty in terms of probability distribution, presented by Shannon (1948). So, entropy of each wavelet decomposed sub-series $T_{m,n}$, which takes values $T^1_{m,n}$, $T^2_{m,n}$, …, $T^N_{m,n}$ with probabilities $P(T^1_{m,n})$, $P(T^2_{m,n})$, …, $P(T^N_{m,n})$, respectively, is defined as (Shannon 1948):

$$E(T_{m.n}) = -\sum_{i=1}^{N_s} P(T^i_{m.n}) \log\left(P(T^i_{m.n})\right) \quad (6)$$

where $E(T_{m.n})$ is entropy of $T_{m,n}$ (also referred to as entropy function) and N_s is the number of intervals or bins to form the histogram and thereinafter *PDF*.

Entropy can statistically represent the rate of uniformity for a single variable but is not able to discover the uniformity and relations between two variables, thus, in order to overcome this shortcoming, joint entropy between two variables is defined as (Gao *et al.* 2008):

$$E(X.Y) = -\sum_{I=1}^{N_s}\sum_{j=1}^{M_s} P(x_i\ .y_i)\log(P(x_i\ .y_i)) \tag{7}$$

Here the examples of two variables X and Y in this study are predictors (X) and predictand (Y), where $p(x_i, y_i)$ is the joint probability of x_i and y_i with number of bins N_s and M_s, respectively. In order to find the relations of two variable MI concepts, an entropy-based criterion can be used as well. In this way, the mutual dependence between the two variables is calculated by quantifying the information content of random variables relative to each other (Nourani *et al.* 2015). While correlation coefficient detects the linear relation between two variables, MI has the capability to detect the nonlinear dependency of variables (Cover & Thomas 1991). MI is computed by the following equation for a predictor x_i and the predictand y_i in this study (Yang *et al.* 2000):

$$MI(X.Y) = E(X) + E(Y) - E(X.Y) \tag{8}$$

For numerical calculation of MI and E using Equations (6) and (8), PDF for all of the variables should be specified. Histogram method is the most common approach of calculating PDF (Yang *et al.* 2000).

Self-organizing map (SOM)

SOMs are unsupervised ANNs which reduce dimensionality of data by generating a low-dimensional, discretized representation of the input space of the training samples, called a map. Neighborhood function is used to keep the topological attributes of the input space. Therefore, SOMs are effective in visualizing low-dimensional illustration of high-dimensional data.

The SOM structure contains components named nodes which are in various arrangements. The common arrangement of nodes is a two-dimensional hexagonal grid. A weight vector of the same dimension as the input data vectors is assigned to each node. In order to detect the node with the closest (smallest distance metric) weight vector (w) to the n-dimensional input vector x Euclidean distance concept is used (Kohonen 1997):

$$\|\boldsymbol{x} - \boldsymbol{w}\| = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{w}_i)^2} \tag{9}$$

The weight with the closest match to the input data is the winner node named best matching unit (*BMU*). In order to further decrease the distance between the weights and *BMU* learning continues by changing the weights at each training iteration t (Kohonen 1997):

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \alpha(t) h_{lm}(\boldsymbol{x} - \boldsymbol{w}(t)) \tag{10}$$

where α corresponds to the learning rate ranging in [0 1], h_{lm} denotes the neighborhood function. The most commonly used neighborhood function is the Gaussian function (Kohonen 1997):

$$h_{lm} = \exp\left(-\frac{l - m^2}{2\sigma(t)^2}\right) \tag{11}$$

where l and m correspond to *BMU* and its neighboring output nodes' position, and σ is the width of the topological neighborhood at iteration t.

Finally, SOM clusters homogeneous data with a similar pattern in a cluster, and reduces the dimensionality of data.

Artificial neural network (ANN)

ANN learns according to input and output data, while information flows through the network. The underlying pattern in the data affects the structure of ANN and leads to generation of the appropriate network according to data. ANN comprises neurons in different layers which interconnect through the network. The simplest ANN includes three layers: input, hidden, and output layers. The input layer includes input neurons that send information to the hidden layer of neurons, and finally processed information

goes to the third layer of output neurons. The greater the number of layers, the more complex is the network.

ANN is extensively applied as a prediction tool and it is recommended to use the back-propagation (BP) algorithm with three-layer network to predict and simulate climatological problems. The structure of ANN is demonstrated by the following equation (Haykin 1994):

$$\hat{y}_k = f_0\left[\sum_{j=1}^{M_N} w_{kj}.f_h\left(\sum_{i=1}^{N_N} w_{ji}x_i + w_{j0}\right) + w_{k0}\right] \quad (12)$$

where i, j, and k show the input, hidden, and output layer neurons, respectively. w_{ji} is a weight in the hidden layer, which connects the ith neuron in the input layer to the jth neuron in the hidden layer; w_{jo} is the bias for the jth hidden neuron, f_h is the activation function of the hidden neuron; w_{kj} is a weight in the output layer connecting the jth neuron in the hidden layer to the kth neuron in the output layer; w_{ko} is the bias for the kth output neuron, f_o is the activation function for the output neuron, x_i is ith input variable for the input layer and $\hat{y}_k$ is computed output. N_N and M_N are the neuron numbers of input and hidden layers, respectively. For more information on the ANN, readers are referred to study relevant books in this field such as Haykin (1994).

Evaluation criteria

Clustering evaluation criterion

In order to measure validity of SOM-based clustering, silhouette coefficient (SC) is applied. SC refers to a technique of verification of consistency within clusters of data. It is obvious that increasing the number of clusters leads to more homogeneous clusters and, generally, the consistency of clusters increases by growing the number of clusters, but such increment may not be wise in high dimensional data such as in the current study where the dominant feature from each cluster will be used by ANN. Since numerous inputs to ANN can decrease the efficiency of the model, a logical number of inputs obtained from clusters should be considered, and to do so, the mean value of SC for each number of cluster was drawn up. In the following, SC calculation is explained.

It is assumed that the data have been clustered via SOM (any technique of clustering can be used) into k clusters. For each datum i, two kinds of dissimilarities are defined; first the average dissimilarity of i with all other data in the same cluster is defined by a(i), which denotes how well i is assigned to its cluster (the smaller the value, the better the assignment). Second, the average dissimilarity of point i to each of the other clusters other than own cluster of i is defined, which is the average distance from i to all points in an individual cluster. The lowest average dissimilarity of i to any other clusters, of which i is not a member is named neighboring cluster and shown by b(i). Indeed, the neighboring cluster is the subsequent best fit cluster for point i. It is noted that any distance metric (e.g., Euclidean distance) can be used to compute the average dissimilarity between two data. The following equation corresponds to silhouette calculation (Amorim & Hennig 2015):

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (13)$$

The S(i) varies between −1, 1 and as a result, a component of a cluster that is closer to 1 indicates that the clustering is done correctly; low or negative results of S(i) show the incorrect placement of a component in the respective cluster; and near zero S(i) means that the datum is located on the border of two clusters.

In order to assess the overall clustering configuration, the average S(i) over all data of all clusters is calculated, which shows how appropriately the data have been clustered.

If there are too many or too few clusters, as may occur when a poor choice of k is used in the clustering algorithm, some of the clusters will typically display much narrower silhouettes than the rest. Thus silhouette averages may be used to determine the natural number of clusters within a data set while too many or too few clusters can occur when a poor choice of cluster numbers is used. The average S(i) for the entire data set, is defined as (Hsu & Li 2010):

$$SC = \frac{1}{n}\sum_{i=1}^{n} S(i) \quad (14)$$

where n is the total number of arrays in the data space.

Simulation evaluation criteria

The comparison between calibrated and observed values is the simplest form of model validation. Several statistical measures are available to evaluate the association between calibrated and observed data; correlation coefficient (CC) is the most common method and denotes the linear relevancy between two variables and can range from $+1$ to -1. In this way, values greater than zero show a positive association and values less than zero indicate a negative relation, while the value of zero states no association between the two variables:

$$CC = \frac{\sum OC - \frac{(\sum O)(\sum C)}{N}}{\sqrt{\left(\sum O^2 - \frac{(\sum O)^2}{N}\right)\left(\sum C^2 - \frac{(\sum C)^2}{N}\right)}} \quad (15)$$

Moreover, in order to evaluate the precision of prediction in a downscaling model, determination coefficient (DC) is used. It measures how well observed results based on their proportion of total variation are replicated by the proposed model (Draper & Smith 1998) and ranges from 0 to 1. The greater DC denotes the agreement of observed and modeled outcomes.

Root mean square error (RMSE) is a measure of accuracy and assesses how effectively downscaling model predicts rainfall:

$$DC = 1 - \frac{\sum_{i=1}^{T} (O_i - C_i)^2}{\sum_{i=1}^{T} (O_i - \bar{O})^2} \quad (16)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{T} (O_i - C_i)^2}{N}} \quad (17)$$

where N, O_i, C_i, and $\bar{O}$ are number of observation data, observed data, calculated values, and mean of observed data, respectively. It is necessary to mention that Legates & McCabe (1999) proved the adequacy of DC and RMSE in hydro-climatology prediction processes as evaluation criteria of models.

Also, in order to test downscaling accuracy, bias measurement index (BI) was used as a criterion to show how well model calibration magnitudes approach the values of the observed data; in other words, BI locates and quantifies errors between calibrated and observed values. BI is defined as a logarithm of the fraction of calibrated and observed rainfall and can take the range of $-\infty . + \infty$ where zero, negative, and positive values, respectively, indicate exact matching, under-estimation, and over-estimation between calibrated and observed rainfall. Here, in this study, the acceptable BI is in the range $[-1.1]$ in which calibration magnitudes fall inside 10% deviation from the observed values:

$$BI = log\frac{C}{O} \quad (18)$$

RESULTS AND DISCUSSION

Among several GCMs, the three potential GCMs (i.e., CanESM2, BNU-ESM, and INM-CM4) were assigned to select dominant predictors. In this regard, important climate variables of all the GCMs were identified according to the WT-EN-SOM methodology, which was proposed as the first step in the current study.

First step-input screening

Through the first step, DWT was applied to decompose time series to the approximation and detail sub-series. The proper selection of the mother wavelet and decomposition level is the important issue while using DWT. Based on Nourani *et al.* (2014), the structure of the Daubechies mother wavelet with four vanishing moments (db4) is similar to the time series of hydro-climatologic processes; thus, it can accurately capture the hidden features of the time series through wavelet analysis. Hence, decomposition procedure was established with the db4 mother wavelet. In regard to selection of the decomposing level, owing to the fact that the aim of the study is to downscale rainfall for the future and the need to have knowledge about different rainfall frequencies for the distant horizon, the level 7 was selected as

the appropriate decomposition level to capture a large range of frequencies from approximately monthly to seasonal up to five-yearly frequencies. Therefore, each predictor was decomposed into one approximation and seven detail sub-series, then, WE of each sub-series was calculated. In this way, a vector of eight members was replaced by the eight long sub-series of a predictor. The procedure was conducted on all predictors of each GCM.

Afterwards, SOM clustering approach was used to cluster the similar features of the predictors in several groups. In order to reduce dimensionality of input space, dominant features from each cluster were selected according to MI.

For each of the three GCMs (i.e., Can-ESM2, BNU-ESM, and INM-CM4), 15 atmospheric predictors at four grid points around the hydrological station were considered to screen dominant variables, as shown in Table 1. The advantage of imposing four grid points around the downscaling station is the ability to take into account the physical characteristics of locations, which affect behavior of climate predictors due to spatial variation on an area covering nearly 103,850 square kilometers (area is calculated based on $2.81^\circ \times 2.81^\circ$ grid size).

After handling the non-stationary property of predictors by wavelet, it is time to select the dominant features among large volumes of data (e.g., for Can-ESM2 GCM, it is 3,840 sub-series in the length of 50 years of monthly data; the number came from 120 large-scale climate variables considering 22 pressure levels multiplied by eight decomposed sub-series at four grid points, as seen in Table 1). In this regard, entropy of an approximation and seven detail sub-series were calculated and a matrix with eight rows and N columns (N depends on the number of predictors at each GCM, (e.g., N = 120 in Can-ESM2 GCM)) was constructed for each GCM. Since the wavelet representation of the predictors differs according to the frequency state of the process, WE values of them also vary. In fact, high values of WEs refer to high stochastic fluctuations obtained from contributions of all frequency bands. In contrast, low values for WEs are expected in the case of ordered sub-series with deterministic and smooth signals.

Henceforth, SOM classified approximately 1,200 feature vectors of predictors obtained from the three GCMs in several clusters. The dominant features of clusters were selected based on maximum MI computed between the predictors in a cluster and the predictand, thus, a predictor with the most nonlinear relevancy to observed rainfall data was designated as the agent for that cluster.

In order to determine the optimum cluster number, the SC measure was calculated and plotted against the number of clusters, which varies between 4 and 100 (Figure 3) due to two-dimensional SOM structures, i.e., 2×2. 3×3. ... 10×10 cluster numbers. According to Figure 3, the SC values increase very smoothly by growing the number of clusters after a considerable drop around ten clusters. Although increase of cluster number raises the SC value, too many clusters should be avoided in order to maintain the efficacy of ANN (numerous inputs can lead to poor performance of ANN in the verification step). Therefore, the optimum cluster number for approximately 1,200 climate features considering ANN efficiency and SC acceptable value can be around 9 to 25. The growth of SC around 9 to 25 clusters is smooth in comparison to the gradient of SC after and before the specified range, as depicted by vertical arrows in Figure 3.

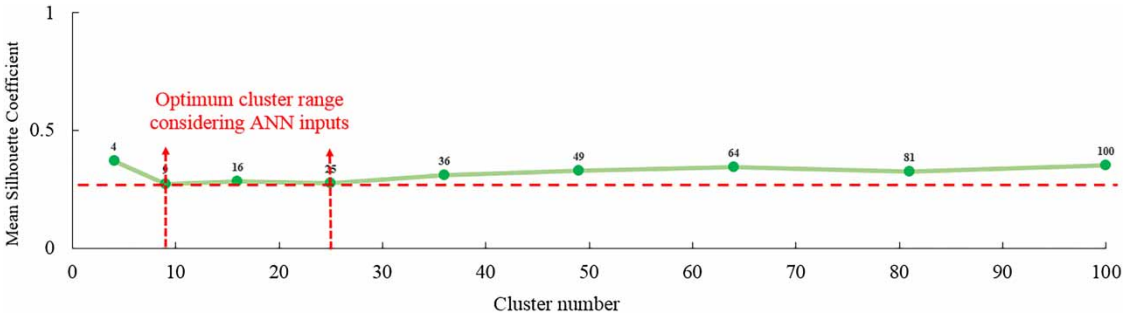

Figure 3 | Mean SC measure against the number of clusters.

After grouping approximately 1,200 climate features of the three GCMs in nine clusters, the dominant agents of clusters, which can operate as the representative of a cluster, were determined by MI. Figure 4(a) illustrates the number of members in each of the nine clusters at SOM hit plot and Figure 4(b) shows the SOM neighbor weight distances plot. The hit plot shows the neuron locations in the topology, and indicates how many of the training data are associated with each of the neurons (cluster centers). The topology is a 3-by-3 grid, so there are nine neurons. The maximum number of hits associated with any neuron is the maximum hit number (here 273), thus, there are 273 input vectors in the relevant cluster. There are eight elements in each input vector, so the input space is eight-dimensional. The weight vectors (cluster centers) fall within this space. The weight distance matrix (also called the U-matrix) is a visualization tool for the SOM, in which, the small hexagons represent the neurons and the lines connect neighboring neurons. The regions containing the lines indicate the distances between neurons, in which darker and lighter regions represent larger and smaller distances, repectively.

According to the SOM hits plot in Figure 4(a), among the nine groups of clustered data, the first cluster (i.e., at left bottom corner) with just two components is considerable. These two members refer to east- and northward winds (85,000 pressure level) at the grid point B4 relevant to the INM-CM4 model with a majority of zero values. The sensitivity analysis proved that these components have no effect on the rainfall of Tabriz; therefore, they were eliminated and not considered in the modeling procedure. The SOM neighbor weight distances plot (Figure 4(b)) shows the discrepancy of these two variables by the darkest shade. It is worth noting that east- and northward winds at 92,500 and 100,000 pressure level of the INM-CM4 model at all the four grid points were totally zero, which was not considered in the clustering process. Dominant features of the clusters, selected in the first step of the proposed methodology, are tabulated in Table 2.

Although various predictors were selected from the three GCMs through the clusters, representative predictors were related to humidity, wind velocity, geopotential height, and heat flux of the atmosphere; accordingly, some previous studies more or less pointed to such atmospheric predictors in rainfall downscaling (Wilby *et al.* 2004; Frost *et al.* 2011; Rashid *et al.* 2015).

In the absence of clustering and selection of important variables based on MI, only variables associated with high values in terms of the MI criterion were selected, which also do not cover the entire variable domain, while other

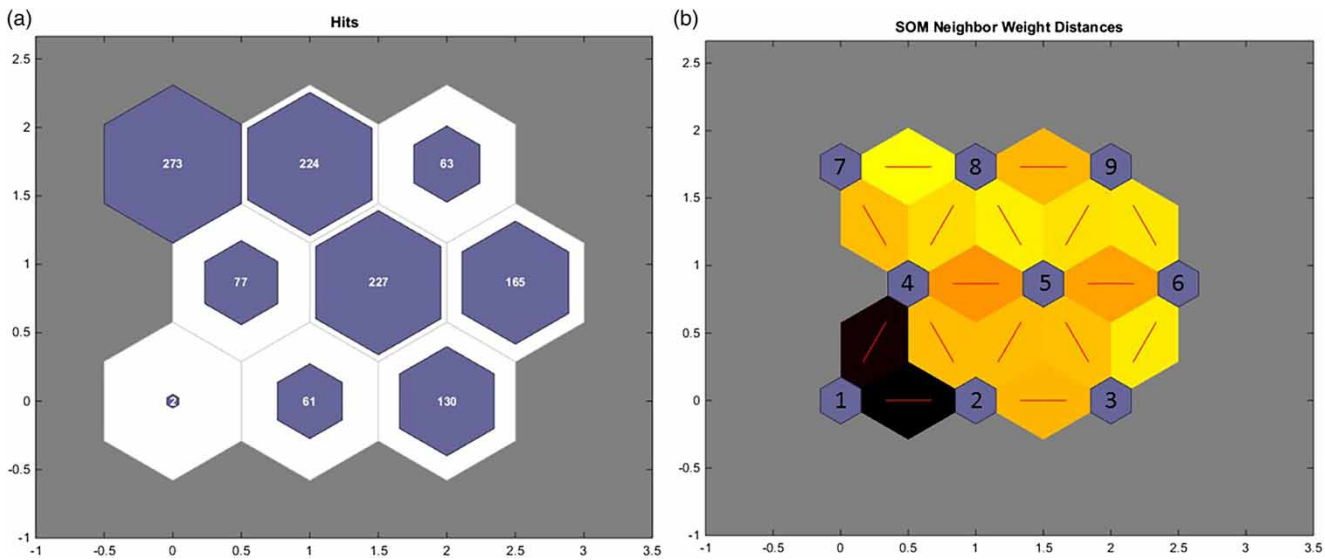

Figure 4 | Clustering presentation by (a) hit plot, the numbers in hexagons show the number of similar climate features and (b) neighbor weight distances plot, the numbers in hexagons show the cluster name.

Table 2 | Dominant features of the clusters, selected in the first step of the proposed methodology

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Variable	va, ua 85000	hfls	ua 100000	hus7000	hfls	zg3000	hus7000	hfls	hfls
Model	INMCM4	BNU-ESM	BNU-ESM	INMCM4	Can-ESM2	INMCM4	Can-ESM2	BNU-ESM	Can-ESM2
Grid No.	B4	A1	A2	B4	A3	B1	A2	A4	A4

variables with moderate or low relation values from the input domain may be associated in generalizing future rainfall and its fluctuation. However, application of clustering makes it happen. The schematic diagram in Figure 5 clarifies the advantage of SOM-MI with regard to the simple MI feature extraction method. In the first approach of Figure 5, which illustrates the simple MI method, ranked values of MI were arranged in a descending order. The issue in this case is how to separate dominant inputs from the MI ranked inputs. In other words, how many of the maximum ranked MIs should be selected? The second approach of Figure 5, which shows application of the clustering method like SOM before MI-based ranking, solves the problem of the simple MI ranked method. In this way, by clustering potential inputs into specified groups, the number of dominant inputs can be defined as the number of clusters, and not necessarily inputs with the largest MI values are selected; instead, different variables with various MIs are selected. Dominant input selection according to the largest MI values leads to selection of only one kind of input; however, in the case of using the clustering (SOM), the input variables are clustered into groups of similar inputs with specific patterns, various ranges of MI. Accordingly, various kinds of inputs can be selected, which causes inputs with different patterns to be imposed on the AI models and, subsequently, elevates the accuracy of prediction of unseen data in the verification step. Therefore, by application of a clustering approach in this study, dominant variables from whole domain of variables, even with low or moderate scores, were selected according to the MI measure.

Figure 6 displays dominant predictors obtained in the first step from different grid points. It is noted that the area between the grid points of Can-ESM2 and BNU-ESM

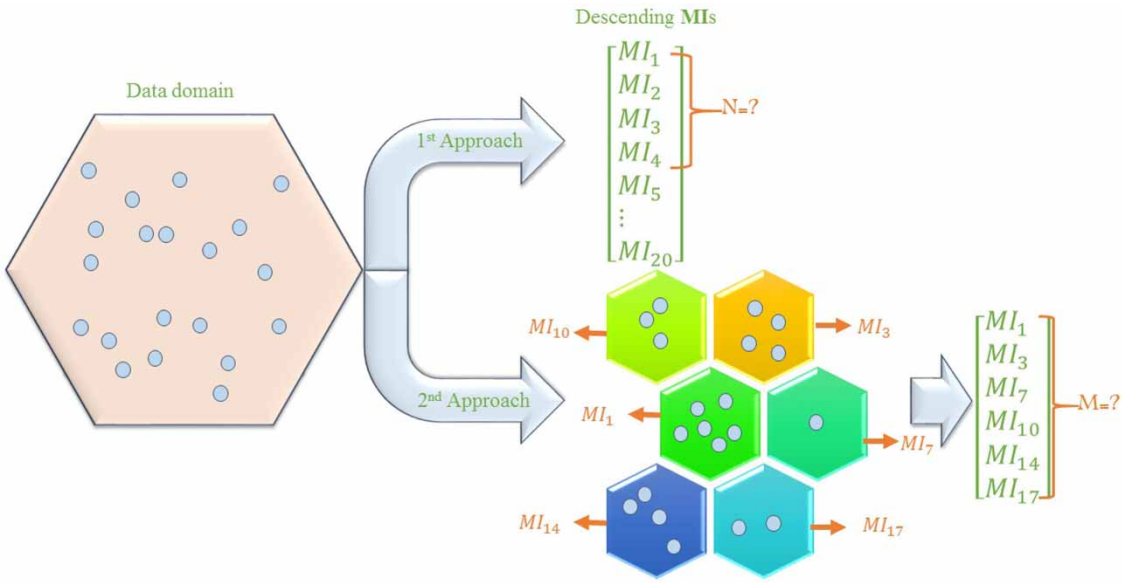

Figure 5 | Schematic of feature extraction methods based on 1st approach: simple MI, in which N is the maximum numbers of MI that should be selected from descending list of MIs; 2nd approach: SOM-MI, in which M is the selected MIs from each cluster involving whole data domain.

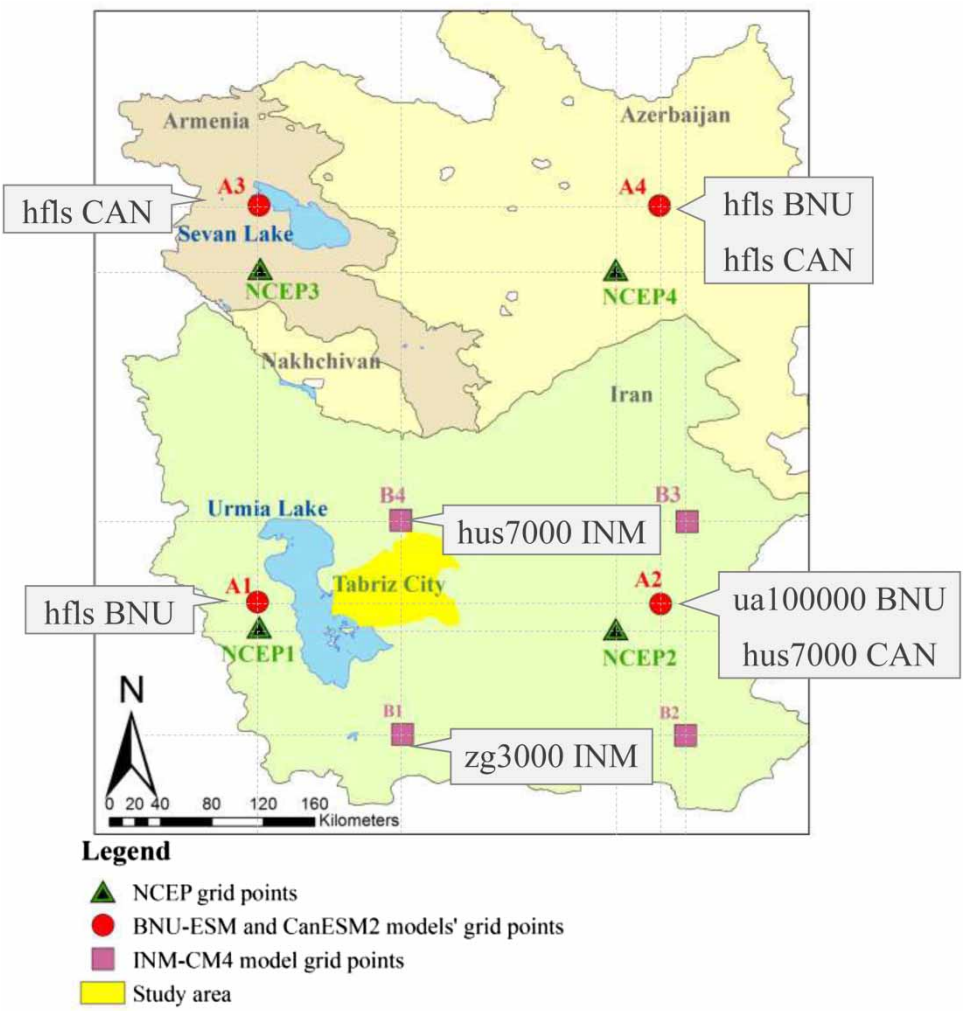

Figure 6 | Important predictors of each GCM depicted on the map of study area.

GCMs, which overlapped, is larger than the area made by the INM-CM4 grid points around the study station. Important predictors on each grid can denote the morphologic impact of that region on atmospheric variables, which consequently affects the Tabriz station rainfall. Concerning the issue, humidity-relevant variables became important on grids, which are influenced by close water bodies. In this way, the specific humidity variable hus7000 showed a dominant effect on Tabriz rainfall at the nearest grids to the study area (i.e., the grids A2 and B4) around the study location. The Caspian Sea and Urmia Lake, as the sources of moisture, lead to dominancy of the humidity variable (i.e., hus7000) on the grid A2 and B4, respectively.

Eastward wind (i.e., ua100000) at the nearest grid (i.e., the grid A2) was the other dominant variable in downscaling rainfall; the wind variable was in line with the prevailing wind of the region.

Hfls was the most important climate variable, which affected Tabriz rainfall from distant grid points and became significant at the grids A1, A3, and A4. The latent heat flux emerges as winds carry moisture away from the surface; thus, it can be involved in regional rainfall as an important factor. In this way, faraway water bodies can also influence Tabriz rainfall through the hfls climate variable. The source of such moisture can be Urmia Lake adjacent to grid A1, Sevan Lake close to grid A3, and the Caspian Sea near grid A4 (see Figure 6).

Near surface geopotential height (i.e., zg 3000) was selected as the dominant climate variable at the southern grid point B1 of the Tabriz station. The way air layers are situated and the amount of atmospheric pressure at one point varies with temperature variations. The hot air has low pressure, whereas the cold air is high in pressure. Thus, alteration of the seasons leads to temperature changes, which in turn, causes change in atmospheric pressure. The change in the atmospheric pressure is the source of wind generation at the earth's surface. On the other hand, winds are the principal cause of cloud displacement, which can be effective in producing rainfall. In this way, the input screen methodology of the current study with the ability to detect nonlinear relationships extracted the effect of the geopotential height on precipitation production.

From Figure 6, it is noticed that the variables of the eastern grid points of INM-CM4 model (i.e., B2 and B3) did not contribute in SOM-MI input screening.

Second step ANN-based downscaling model

Eight element sets of climatic variables were selected as inputs of the ANN model in the first step. The ANN inputs included two types of data: (i) ensemble of GCMs and similar (ii) NCEP reanalysis data. The large-scale climate variables that passed through the SOM screening procedure were standardized over the period 1951–2005. The first 75% of the dominant predictors and observed rainfall data was used for the training of downscaling model and the remaining 25% used for the validating purpose. Hence, the calibration period of the ANN model was from 1951 to 1991, and the verification period was from 1992 to 2006.

The three-layer feed-forward neural network with back propagation algorithm and the Levenberg–Marquardt scheme, with tangent sigmoid (Tansig) as the activation function was used to downscale Tabriz station rainfall. Different hidden neurons (which were selected via trial and error method) through 1,000 epochs were examined. Evaluating criteria of ANN model demonstrated that the maximum efficiency occurred at 480 epochs with four hidden neurons. Results of multi-GCM ensemble downscaling model is tabulated in Table 3.

In order to determine the efficiency of the proposed screening methodology, standard correlation analysis also was used to select dominant predictors (the acceptable range set to 0.3 and up). The obtained predictors from CC-based method were fed into the ANN and results also tabulated in Table 3. The results showed that the proposed screening method was 46% more successful than CC-based downscaling due to recognizing the nonlinear relations and filtering redundant information, and thus could detect a more robust pattern between predictor and predictand.

Table 3 | ANN and MLR downscaling results based on various inputs determined by WE-SOM-MI and CC predictor screening methodology

Model	GCM-based inputs selection	DC train	DC verify	N[a]-RMSE train	N[a]-RMSE verify	CC train	CC verify
ANN	CC	0.23	0.18	0.91	0.99	0.30	0.28
ANN	WE-SOM-MI	0.50	0.41	0.74	0.69	0.70	0.67
MLR	WE-SOM-MI	0.30	0.40	0.86	0.70	0.55	0.59

[a]N denotes normalized RMSE values.

To investigate the relationship between the grid point distance from the Tabriz station and the impact of variables at the grid points over the predictand forecasting, ANN-based sensitivity analysis was performed on the variables of each grid point and the results are shown in Table 4.

According to Table 4, none of the two grid points (i.e., A1 and B4) with the least distance from the study area could catch the maximum effect on production of the predictand; however, the middle distance grid point (i.e., A2) had the maximum effect on calibration of the predictand. Moreover, the two distant grid points (i.e., A3 and A4) with little difference with point A2 were more effective for downscaling the Tabriz station rainfall.

Therefore, the significance of large-scale climate variables in the rainfall production of the region was not related to the distance of the grid points from the studied

Table 4 | The sensitivity analysis over grid point distance from Tabriz station and their impact on station rainfall, according to DC verify of ANN downscaling model

Grid points	Point A1	Point A2	Point A3	Point A4	Point B1	Point B4
DC verify of ANN downscaling	0.19	0.30	0.28	0.27	0.08	0.22

area, so that the distant points could be more effective. This sensitivity analysis surely confirmed that the simultaneous effect of climate variables from different locations could lead to a more robust downscaling model.

Since some of the commonly used statistical downscaling models such as SDSM use the multiple linear regression (MLR) method to establish a statistical relationship between predictors and the predictand, in order to assess the performance of the ANN model, the MLR model was also developed in this study (Table 3) and compared with the results of the proposed ANN-based methodology. Results of Table 3 demonstrate that the ANN-based downscaling model outperforms the MLR model in terms of evaluating criteria, because MLR was not capable of distinguishing nonlinear relations among dominant predictors and the predictand. It should be mentioned that the MLR model was also built based on the dominant climate variables, which were used in the ANN downscaling model.

In general, the ANN model showed better performance in reproducing the rainfall statistics of the Tabriz station than the MLR model in reproducing a reasonable trend and variation of the observed data. Figure 7 depicts the mean monthly rainfall for calibration and validation steps according to the observed data of the Tabriz station as well as the ANN (with two types of inputs: multi-GCM and NCEP) and MLR downscaling models. Moreover, it was observed that the ANN-based downscaling model calibrated by NCEP data performed better than ANN model calibrated by multi-GCM historical data.

Moreover, Figure 7 demonstrates that both the ANN downscaling models showed high efficiency in forecasting peak points of rainfall so that, as an example, the April mean rainfall was forecasted as accurately as the observed data, while the MLR model could not perform as efficiently as the ANN model.

Furthermore, in order to test the ANN downscaling accuracy, the BI criterion was calculated for the multi-GCM ensemble model at the verification step. Figure 8 shows that, during the dry months of summer (i.e., June, July, August, and September), the proposed multi-GCM ANN model overestimates Tabriz rainfall, while in other months, the model performs more reliably with acceptable BI values falling inside 10% deviation from the observed values (i.e., the two horizontal dashed lines specify BI in the range of $[-1.1]$, which is acceptable BI range).

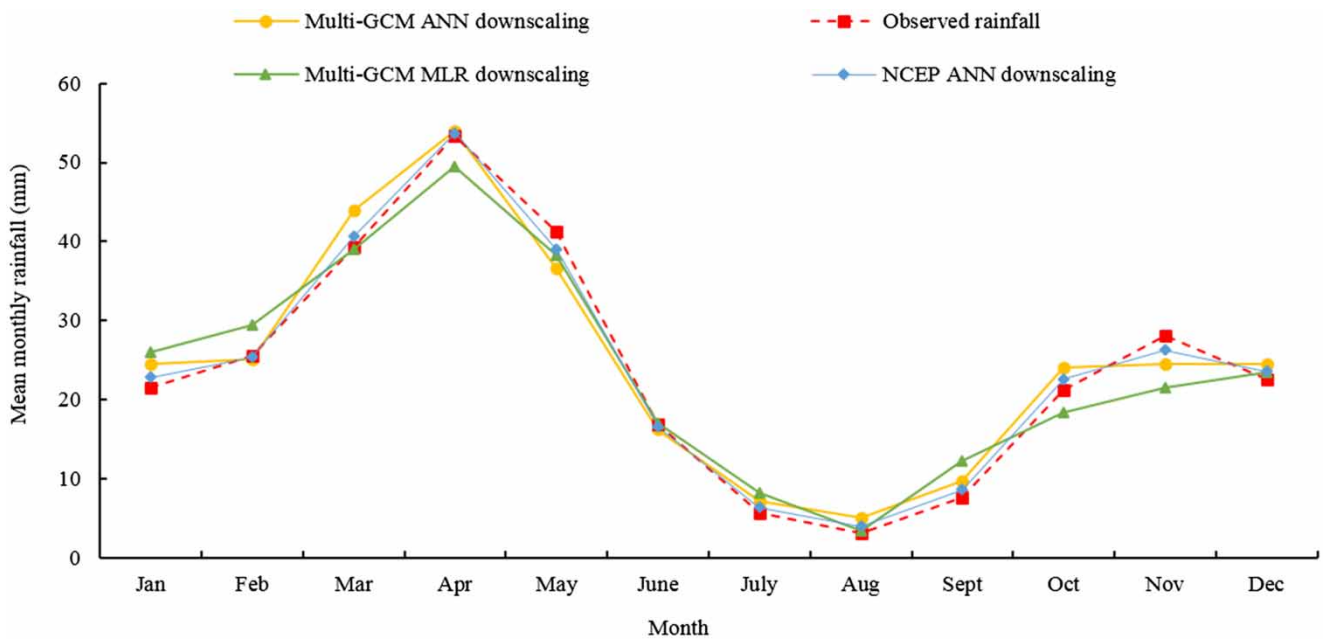

Figure 7 | Mean monthly rainfall of the observed and calibrated rainfall for the ANN and MLR downscaling models.

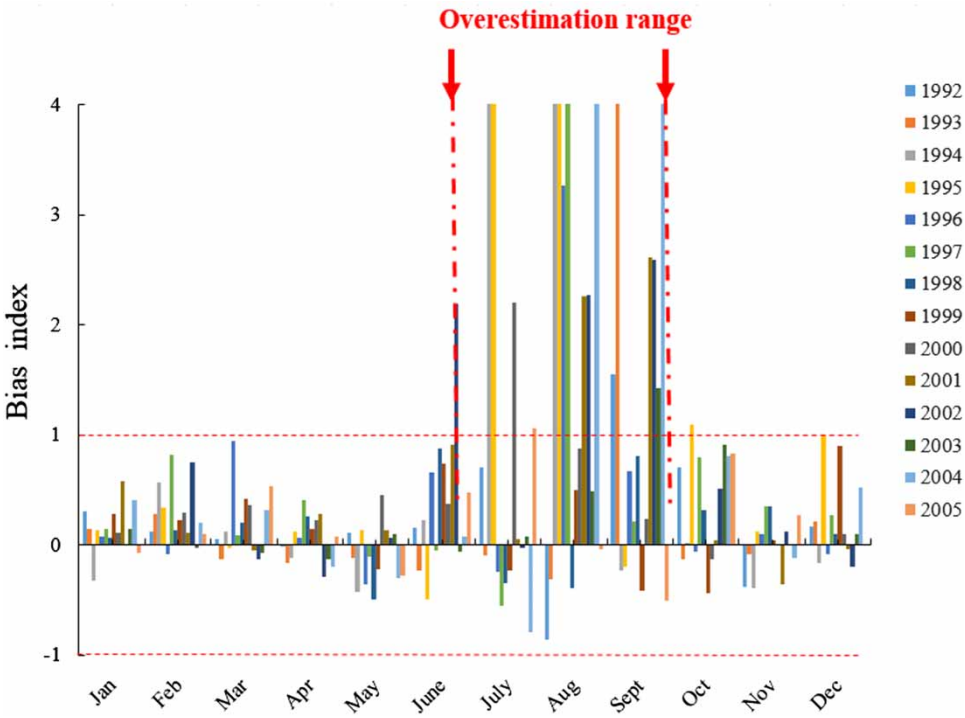

Figure 8 | BI results based on observed and ANN downscaling of the multi-GCMs ensemble model during the validation period 1992–2005.

Third step-rainfall projection for future

Since the projection of faraway future emissions and other human factors that impact atmosphere is troublesome, researchers utilize a variety of situations utilizing different suppositions about future economic, social, technological, and environmental conditions, named projection scenarios (here RCPs). RCPs indicate four greenhouse gas concentration (not emissions) paths adopted by the IPCC-AR5. Among the four RCPs, including RCP2.6, RCP4.5, RCP6, and RCP8.5, RCPs 4.5 and 8.5 were used in the current study.

Future rainfall projection under RCPs 4.5 and 8.5 was conducted for near and distant future over the 2017–2050 and 2051–2100 periods after calibrating the ANN downscaling model with NCEP data. The boxplots in Figure 9 compare the distribution of rainfall data based on the minimum, first quartile, median, third quartile, and maximum for observation, calibration and projection values in near and distant future under the two RCPs.

According to Figure 9(a), the mean values of the ANN calibrated rainfall at baseline (i.e., 1951–2005) were almost equal to the observed rainfall values; however, the data dispersion was reduced in calibrated models. The comparison of the annual observed and simulated rainfall values demonstrated that over both RCPs and periods (i.e., near and distant future), the mean rainfall values decrease. The highest decrement in annual rainfall will be shown in RCP4.5 at the end of the century. The decrease in rainfall under RCP4.5 and RCP8.5 revealed decrease of 0.29%–36% and 21%–30% for near and distant future compared to baseline, respectively. Figure 9(b) demonstrates the seasonal projections of observed and simulated rainfall during baseline and future (2017–2100), in which mean rainfall during all seasons under both RCPs will show decrease, while during autumn and winter decreasing is more notable than decline during spring and summer. Especially during spring, the rainfall under both RCPs tends to be almost constant comparing observation at baseline. Projections show that although mean rainfall will decrease, the extreme events will be more probable. This outcome is also in line with the IPCC 2013 report which stated that 'the amount of rain falling in heavy precipitation events is likely to increase in most regions, while storm tracks are projected to shift poleward.'

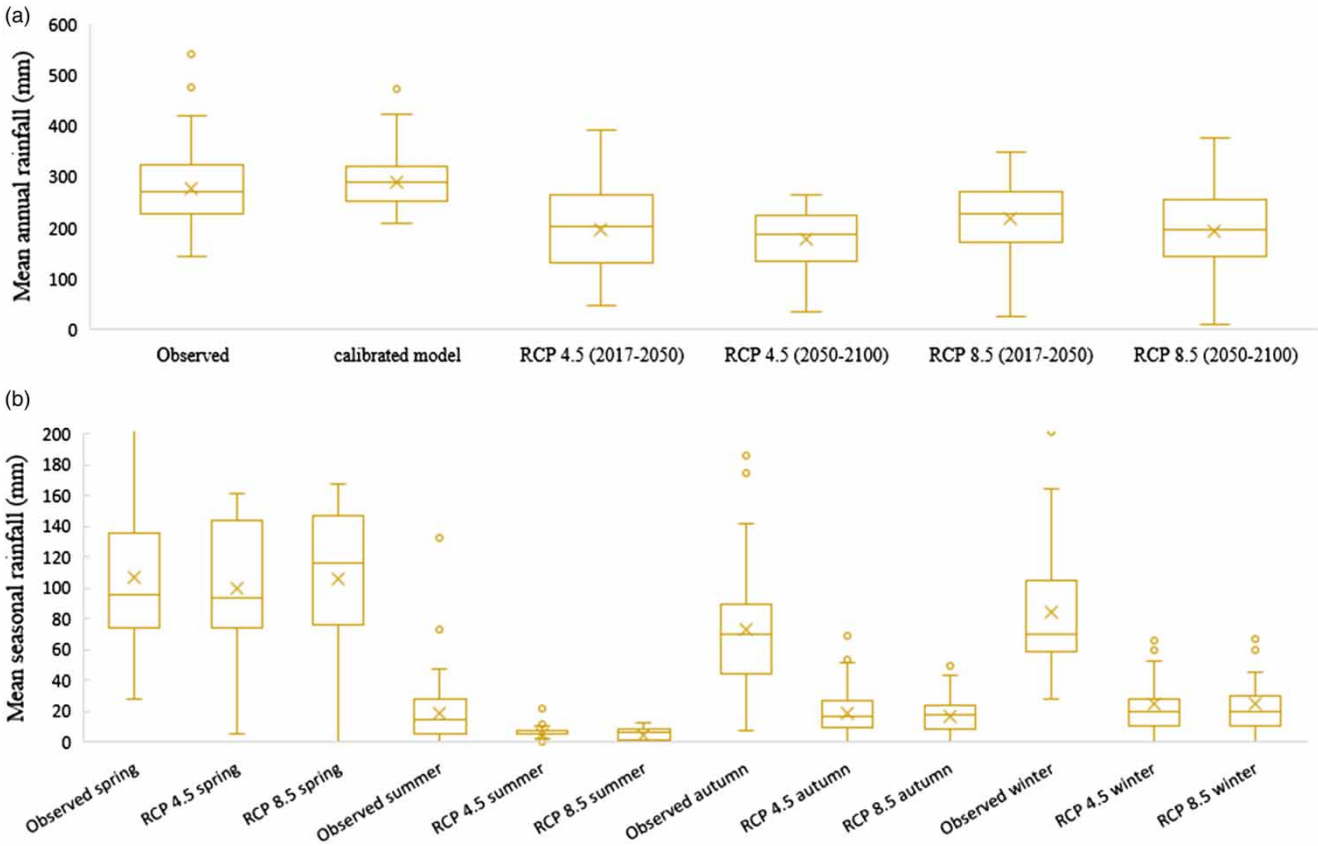

Figure 9 | (a) Mean annual boxplots of the observed, calibrated, and simulated rainfall in the near and distant future. (b) Mean seasonal boxplots of the observed and simulated rainfall for 2017–2100.

Figure 10 depicts annual variations of rainfall for the observed and simulated values under both RCPs 4.5 and 8.5. Figure 10 indicates the decreasing trend line and the changes in annual rainfall encompass considerable year-to-year variations.

CONCLUDING REMARKS

The Tabriz hydrological station rainfall variation is the important parameter in developing hydrologic impact studies of the city. In this regard, the future rainfall of the

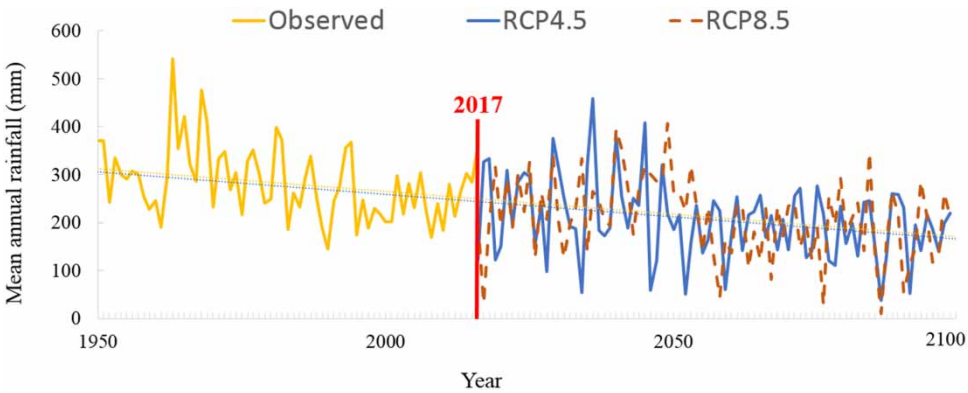

Figure 10 | Mean annual rainfall variation during 1950–2100.

city was assessed by an ANN-based SDSM. The dominant climate predictors were determined via the novel WE-SOM predictor screening methodology among three various GCMs as the multi-GCM ensemble model and reanalysis NCEP data sets. According to proper calibration results of the NCEP-based ANN downscaling, the later model was used in the simulation step as well.

The advantages of using the proposed predictor screening methodology was that the wavelet transform handled non-stationary effects of climate variables by depicting various periods involved in the time series and WE diminished the dimensionality of data by replacing the decomposed sub-series, while preserving the information content underlaid in the data. Moreover, SOM clustered the features of the predictors in homogenous clusters, so one agent from each cluster could represent that cluster instead of applying several resembling data. The agents were selected by MI with the capability of detecting nonlinear relation to find the most correlated predictors to the predictand in each cluster. Although the proposed screening methodology with different mathematical tools seems to be complicated (which may be considered as a limitation of modeling), the selection of optimum variables (i.e., 8 variables) from 1,200 potential variables resulted in the establishment of a robust and suitable downscaling model.

Finally, the selected dominant predictors were applied in ANN to downscale rainfall of the location. ANN captured a nonlinear statistical relationship between the large-scale circulation variables and rainfall of the Tabriz station. The comparison of the linear (i.e., MLR) and nonlinear (i.e., ANN) regression models indicated that ANN can be more reliable than the MLR model in downscaling rainfall at the study station due to 40% better performance in the model training step.

In contrast to research, in which the variables were selected at important grid points according to inverse distance weighting (IDW) based on the distance from the study station (Zhou & Zhang 2014), the present study showed that sometimes the distance cannot be a suitable criterion for choosing an effective variable to simulate predictand. Therefore, in this study, the contribution of the variables in distant points was even more than in close points in downscaling the Tabriz station rainfall.

In the last step, future rainfall of the Tabriz station under RCPs 4.5 and 8.5 was projected. Results of the simulations indicated that, according to intermediate and high emission scenarios, the station rainfall will decrease in average (i.e., 0.29%–36%) and high states (i.e., 21%–30%), respectively.

Overall, the results of this study provide promising evidence for statistical downscaling and, more specifically, for the proposed WE-SOM input screening method, to select climate variables. In order to build on the current study, it is recommended in the future that the proposed WE–SOM screening method is used to downscale other climatologic parameters of the station (e.g., temperature). Moreover, the proposed downscaling method can be compared with statistical and dynamical downscaling models (other than the MLR, which was used in this study). In this way, due to the different abilities of artificial intelligence and machine learning methods, it may be suggested to use other versions and kinds of AI models (e.g., SVM, ANFIS, GEP) and clustering approaches.

REFERENCES

Abdellatif, M., Atherton, W. & Alkhaddar, R. 2013 A hybrid generalised linear and Levenberg–Marquardt artificial neural network approach for downscaling future rainfall in North Western England. *Hydrology Research* **44** (6), 1084–1101. doi:10.2166/nh.2013.045.

Acharya, N., Chattopadhyay, S., Mohanty, U. C., Dash, S. K. & Sahoo, L. N. 2013 On the bias correction of general circulation model output for Indian summer monsoon. *Meteorological Applications* **20**, 349–356. doi: 10.1002/met.1294.

Ahmadi, A. & Han, D. 2013 Identification of dominant sources of sea level pressure for precipitation forecasting over Wales. *Journal of Hydroinformatics* **15** (3), 1002–1021. doi:10.2166/hydro.2012.110.

Ahmadi, A., Han, D., Kakaei Lafdani, E. & Moridi, A. 2015 Input selection for long-lead precipitation prediction using large-scale climate variables: a case study. *Journal of Hydroinformatics* **17** (1), 114–129. doi:10.2166/hydro.2014.138.

Amorim, R. C. & Hennig, C. 2015 Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences* **324**, 126–145. doi:10.1016/j.ins.2015.06.039.

Bação, F., Lobo, V. & Painho, M. 2005 Self-organizing maps as substitutes for k-means clustering. *Computational Science–ICCS* **2005**, 9–28.

Beecham, S., Rashid, M. & Chowdhury, R. K. 2014 Statistical downscaling of multi-site daily rainfall in a South Australian catchment using a Generalized Linear Model. *International Journal of Climatology* **34** (14), 3654–3670. doi:10.1002/joc.3933.

Bowden, G. J., Dandy, G. C. & Maier, H. R. 2005 Input determination for neural network models in water resources applications. Part 1 – Background and methodology. *Journal of Hydrology* **301** (1–4), 75–92. doi:10.1016/j.jhydrol.2004.06.021.

Cai, X., Wang, D., Zhu, T. & Ringler, C. 2009 Assessing the regional variability of GCM simulations. *Geophysical Research Letters* **36** (2). doi:10.1029/2008GL036443.

Chadwick, R., Coppola, E. & Giorgi, F. 2011 An artificial neural network technique for downscaling GCM outputs to RCM spatial scale. *Nonlinear Processes in Geophysics* **18** (6), 1013–1028. doi:10.5194/npg-18-1013-2011.

Chen, J. & Adams, B. J. 2006 Integration of artificial neural networks with conceptual models in rainfall-runoff modeling. *Journal of Hydrology* **318** (1), 232–249. doi:http://dx.doi.org/10.1016/j.jhydrol.2005.06.017.

Chen, S. T., Yu, P. S. & Tang, Y. H. 2010 Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology* **385**(1–4), 13–22. doi:10.1016/j.jhydrol.2010.01.021.

Chuang, J. M., Lin, S. S., Kan, P. H., Li, C. Y. & Hu, Y. L. 2016 Applying bootstrap and radial basis function neural networks developing a climate change statistical downscaling model. *Taiwan Water Conservancy* **64** (4), 48–58.

Cover, J. & Thomas, A. 1991 *Elements of Information Theory*. John Wiley and Sons, Hoboken, NJ. doi:10.1002/0471200611.

Danandeh Mehr, A. & Kahya, E. 2016 Grid-based performance evaluation of GCM-RCM combinations for rainfall reproduction. *Theoretical and Applied Climatology* **129**(1–2), 47–57.

Devak, M. & Dhanya, C. T. 2014 Downscaling of precipitation in Mahanadi Basin, India using Support Vector Machine, K-Nearest Neighbor and Hybrid of Support Vector Machine with K-Nearest Neighbor. In: *The 16th International Association for Mathematical Geosciences – Geostatistical and Geospatial Approaches for the Characterization of Natural Resources in the Environment: Challenges, Processes and Strategies Conference*, New Delhi, India.

Dibike, Y. B. & Coulibaly, P. 2006 Temporal neural networks for downscaling climate variability and extremes. *Neural Networks* **19** (2), 135–144. doi:10.1016/j.neunet.2006.01.003.

Draper, N. R. & Smith, H. 1998 *Applied Regression Analysis*, 3rd ed. John Wiley and Sons, Hoboken, NJ.

Frost, A. J., Charles, S. P., Timbal, B., Chiew, F. H. S., Mehrotra, R., Nguyen, K. C., Chandler, R. E., McGregor, J. L., Fu, G., Kirono, D. G. C., Fernandez, E. & Kent, D. M. 2011 A comparison of multi-site daily rainfall downscaling techniques under Australian conditions. *Journal of Hydrology* **408**(1–2), 1–18. doi:10.1016/j.jhydrol.2011.06.021.

Gao, Z., Gu, B. & Lin, J. 2008 Monomodal image registration using mutual information based methods. *Image and Vision Computing* **26** (2), 164–173. doi:http://dx.doi.org/10.1016/j.imavis.2006.08.002.

Ghosh, S. & Mujumdar, P. P. 2008 Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Advances in Water Resources* **31** (1), 132–146. doi:10.1016/j.advwatres.2007.07.005.

Guo, J., Chen, H., Xu, C. Y., Guo, S. & Guo, J. 2012 Prediction of variability of precipitation in the Yangtze River Basin under the climate change conditions based on automated statistical downscaling. *Stochastic Environmental Research and Risk Assessment* **26** (2), 157–176. doi:10.1007/s00477-011-0464-x.

Harpham, C. & Wilby, R. L. 2005 Multi-site downscaling of heavy daily precipitation occurrence and amounts. *Journal of Hydrology* **312** (1–4), 235–255. doi:10.1016/j.jhydrol.2005.02.020.

Haykin, S. 1994 *Neural Networks (Computer Science)*. MacMillan College Publishing Co., New York.

Hsu, K. C. & Li, S. T. 2010 Clustering spatial-temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources* **33** (2), 190–200. doi:10.1016/j.advwatres.2009.11.005.

IPCC 2013 *Climate Change 2013: The Physical Science Basis*. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex & P. M. Midgley, eds) Cambridge University Press, Cambridge and New York, 1535 pp, doi:10.1017/CBO9781107415324.

Johnson, F., Westra, S., Sharma, A. & Pitman, A. J. 2011 An assessment of GCM skill in simulating persistence across multiple time scales. *Journal of Climate* **24** (14), 3609–3623. doi:10.1175/2011JCLI3732.1.

Khan, F. & Pilz, J. 2017 A Bayesian Approach for GCMs Selection and Ensemble Projections under the latest Emission Scenarios. *Proceedings of the 19th EGU General Assembly*, Vienna, Austria.

Khan, M. S., Coulibaly, P. & Dibike, Y. 2006 Uncertainty analysis of statistical downscaling methods. *Journal of Hydrology* **319** (1–4), 357–382. doi:10.1016/j.jhydrol.2005.06.035.

Kohonen, T. 1997 Exploration of very large databases by self-organizing maps. In: *Paper Presented at the Proceedings of the 1997 IEEE International Conference on Neural Networks*. Part 4 (of 4), Piscataway, NJ, USA

Lee, J. K. & Kim, Y. O. (2017). Selection of representative GCM scenarios preserving uncertainties. *Journal of Water and Climate Change*. In Press. doi: 10.2166/wcc.2017.101.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35** (1), 233–241. doi:10.1029/1998WR900018.

Li, G., Zhang, X., Zwiers, F. & Wen, Q. H. 2012 Quantification of uncertainty in high-resolution temperature scenarios for North America. *Journal of Climate* **25** (9), 3373–3389. doi:10.1175/JCLI-D-11-00217.1.

Li, X., Maier, H. R. & Zecchin, A. C. 2015 Improved PMI-based input variable selection approach for artificial neural

network and other data driven environmental and water resource models. *Environmental Modelling & Software* **65**, 15–29. http://dx.doi.org/10.1016/j.envsoft.2014.11.028.

Mallat, S. G. 1998 *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA.

Markus, M., Hejazi, M. I., Bajcsy, P., Giustolisi, O. & Savic, D. A. 2010 Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois. *Journal of Hydroinformatics* **12** (3), 251–261. doi:10.2166/hydro.2010.064.

May, R. J., Maier, H. R., Dandy, G. C. & Fernando, T. M. K. G. 2008 Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling and Software* **23**(10–11), 1312–1326. doi:10.1016/j.envsoft.2008.03.007.

Miao, C., Duan, Q., Sun, Q., Huang, Y., Kong, D., Yang, T., Ye, A., Di, Z. & Gong, W. 2014 Assessment of CMIP5 climate models and projected temperature changes over Northern Eurasia. *Environmental Research Letters* **9** (5). doi:10.1088/1748-9326/9/5/055007.

Nourani, V. & Parhizkar, M. 2013 Conjunction of SOM-based feature extraction method and hybrid wavelet-ANN approach for rainfall–runoff modeling. *Journal of Hydroinformatics* **15** (3), 829–848. doi:10.2166/hydro.2013.141.

Nourani, V., Baghanam, A. H., Adamowski, J. & Kisi, O. 2014 Applications of hybrid wavelet–artificial intelligence models in hydrology: a review. *Journal of Hydrology* **514**, 358–377. http://dx.doi.org/10.1016/j.jhydrol.2014.03.057

Nourani, V., Khanghah, T. R. & Baghanam, A. H. 2015 Application of entropy concept for input selection of wavelet-ANN based rainfall-runoff modeling. *Journal of Environmental Informatics* **26** (1), 52–70. doi:10.3808/jei.201500309.

Nourani, V., Andalib, G. & Dąbrowska, D. 2017 Conjunction of wavelet transform and SOM-mutual information data pre-processing approach for AI-based Multi-Station nitrate modeling of watersheds. *Journal of Hydrology* **548**, 170–183. https://doi.org/10.1016/j.jhydrol.2017.03.002

Okkan, U. & Fistikoglu, O. 2014 Evaluating climate change effects on runoff by statistical downscaling and hydrological model GR2M. *Theoretical and Applied Climatology* **117** (1), 343–361. doi:10.1007/s00704-013-1005-y.

Okkan, U. & Kirdemir, U. 2016 Downscaling of monthly precipitation using CMIP5 climate models operated under RCPs. *Meteorological Applications* **23** (3), 514–528. doi:10.1002/met.1575.

Olsson, J., Uvo, C. B. & Jinno, K. 2001 Statistical atmospheric downscaling of short-term extreme rainfall by neural networks. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere* **26** (9), 695–700. http://dx.doi.org/10.1016/S1464-1909(01)00071-5

Olsson, J., Uvo, C. B., Jinno, K., Kawamura, A., Nishiyama, K., Koreeda, N., Nakashima, T. & Morita, O. 2004 Neural networks for rainfall forecasting by atmospheric downscaling. *Journal of Hydrologic Engineering* **9** (1), 1–12. doi:10.1061/(ASCE)1084-0699(2004)9:1(1).

Openshaw, S. & Openshaw, C. 1997 *Artificial Intelligence in Geography*. John Wiley & Sons, Chichester, UK.

Racsko, P., Szeidl, L. & Semenov, M. 1991 A serial approach to local stochastic weather models. *Ecological Modelling* **57** (1), 27–41. http://dx.doi.org/10.1016/0304-3800(91)90053-4

Rashid, M. M., Beecham, S. & Chowdhury, R. K. 2015 Statistical downscaling of CMIP5 outputs for projecting future changes in rainfall in the Onkaparinga catchment. *Science of the Total Environment* **530–531**, 171–182. doi:10.1016/j.scitotenv.2015.05.024.

Sachindra, D. A. & Perera, B. J. C. 2016 Statistical downscaling of general circulation model outputs to precipitation accounting for non-stationarities in predictor-predictand relationships. *PLoS ONE* **11** (12). doi:10.1371/journal.pone.0168701.

Sachindra, D. A., Huang, F., Barton, A. F. & Perera, B. J. C. 2014 Multi-model ensemble approach for statistically downscaling general circulation model outputs to precipitation. *Quarterly Journal of the Royal Meteorological Society* **140** (681), 1161–1178. doi:10.1002/qj.2205.

Sailor, D. J. & Li, X. 1999 A semiempirical downscaling approach for predicting regional temperature impacts associated with climatic change. *Journal of Climate* **12** (1), 103–114.

Shannon, C. E. 1948 A mathematical theory of communication. *Bell System Technical Journal* **27** (3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

Sousa, S. I. V., Martins, F. G., Alvim-Ferraz, M. C. M. & Pereira, M. C. 2007 Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software* **22** (1), 97–103. http://dx.doi.org/10.1016/j.envsoft.2005.12.002

Tisseuil, C., Vrac, M., Lek, S. & Wade, A. J. 2010 Statistical downscaling of river flows. *Journal of Hydrology* **385**(1–4), 279–291. doi:10.1016/j.jhydrol.2010.02.030.

Tripathi, S., Srinivas, V. V. & Nanjundiah, R. S. 2006 Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of Hydrology* **330**(3–4), 621–640. doi:10.1016/j.jhydrol.2006.04.030.

Trzaska, S. & Schnarr, E. 2014 *A Review of Downscaling Methods for Climate Change Projections*. Report for the United States Agency for International Development by Tetra Tech ARD.

Varela, S., Lima-Ribeiro, M. S. & Terribile, L. C. 2015 A short guide to the climatic variables of the last glacial maximum for biogeographers. *PLoS ONE* **10** (6). doi:10.1371/journal.pone.0129037.

Wilby, R. L. & Wigley, T. M. L. 1997 Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography* **21** (4), 530–548.

Wilby, R. L., Dawson, C. W. & Barrow, E. M. 2002 Sdsm – a decision support tool for the assessment of regional climate change impacts. *Environmental Modelling & Software* **17** (2), 145–157. https://doi.org/10.1016/S1364-8152(01)00060-3

Wilby, R. L., Charles, S. P., Zorita, E., Timbal, B., Whetton, P. & Mearns, L. O. 2004 *Guidelines for use of Climate Scenarios Developed From Statistical Downscaling Methods*. Supporting material of the Intergovernmental Panel on

Climate Change (IPCC), prepared on behalf of Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA). doi:citeulike-article-id:8861447.

Yang, H. H., Vuuren, S. V., Sharma, S. & Hermansky, H. 2000 Relevance of time–frequency features for phonetic and speaker-channel classification. *Speech Communication* **31** (1), 35–50. http://dx.doi.org/10.1016/S0167-6393(00)00007-8

Zhou, Y. & Zhang, J. 2014 Application of GIS in downscaling regional climate model results over the province of Ontario. *Environmental Systems Research* **3** (1), 8. doi:10.1186/2193-2697-3-8.

Zorita, E. & Von Storch, H. 1999 The analog method as a simple statistical downscaling technique: comparison with more complicated methods. *Journal of Climate* **12** (8), 2474–2489. https://doi.org/10.1175/1520-0442(1999)012

First received 24 October 2017; accepted in revised form 1 February 2018. Available online 20 March 2018