Amirkabir Journal of Civil Engineering

Amirkabir J. Civil Eng., 52(4) (2020) 219-222
DOI: 10.22060/ceej.2018.14986.5806

The conjunction of the feature extraction method with AI-based ensemble statistical downscaling models

Z. Razzaghzadeh, V. Nourani* , N. Behfar

Department of Water Resources Engineering, Faculty of Civil Engineering, University of Tabriz, Tabriz, Iran.

ABSTRACT: In this study, two general circulation models (GCMs) (Can-ESM2, BNU-ESM) were used to simulate the future precipitation of Tabriz city. The weakness of GCMs is the coarse resolution of climate variables in which the different methods of downscaling is about to solve this deficiency. In this study, the Artificial Intelligence (AI) models, i.e., Artificial Neural Network (ANN) and Adaptive neuro-fuzzy inference system (ANFIS), were used to statistically downscale the climate variables of GCMs. Without any doubt, the most important step during the use of these models is selecting the dominant inputs among huge large-scale GCM data. So in this study for the selection of dominant inputs, decision tree, and mutual information (MI) feature extraction methods were used. Also, the ensemble techniques were used to evaluate the efficiency of downscaling models and to decrease the uncertainties. A comparison of the result of downscaling models indicated that the ensemble technique (i.e., hybrid of ANN and ANFIS) with dominant inputs based on decision tree feature extraction methods presents better performance. In both GCMs, the application of the downscaling ensemble couple with dominant predictors based on a decision tree model in precipitation downscaling showed 10%-38% increase in DC in versus the individual ANN and ANFIS downscaling models. The projection precipitation of Tabriz synoptic station for future (2020-2060) by proposed ensemble AI-based model indicated 30%-40% precipitation decreases under RCP4.5 and RCP8.5 scenarios.

Review History:
Received: 9/15/2018
Revised: 10/1/2018
Accepted: 10/29/2018
Available Online: 12/15/2018

Keywords:
General Circulation Models (GCMs)
Adaptive neuro-fuzzy inference system (ANFIS)
Artificial Neural Network (ANN)
Mutual Information (MI)
Statistical Downscaling

1- INTRODUCTION

Water resources are extremely influenced by climate change. The negative effects of climate change on a different component of water resources such as; agriculture, industry, and so on have caused concern in human societies. Therefore, the survey of precipitation changes as the major component of the hydrologic cycle seems necessary. General circulation models (GCMs) can be considered as reliable tools in the prediction of precipitation. GCMs use physical-based equations on various processes of the atmosphere so, have widely used in various sciences [1]. The outputs of GCMs developed in coarse spatial resolutions may lead to their poor applicability as the input to local-scale hydrologic models. Downscaling is a technique to catch local-scale climate data from large scale GCMs [2]. Therefore, the current study proposes a novel statistical downscaling methodology by ensembling AI-based models (i.e., ANN and ANFIS). Plenty of data sets available by GCMs are the most challenging issue in AI-based modeling. In this case, the application of feature extraction methods as a pre-processing technique can largely increase the efficiency of the AI-based downscaling model. Multilinear and non-linear feature extraction methods i.e., decision tree and mutual information, respectively, are used in the current study. Finally, the proposed novel predictor screening method is incorporated into AI-based statistical downscaling methods to achieve the optimal model for projection of monthly precipitation of Tabriz station.

*Corresponding author's email: vahid.nourani96@gmail.com

2- MATERIAL AND METHODS

2-1- Study Area and Data

Tabriz City (latitude: 38˚08'N, longitude: 46˚ 29'E38), the capital city of East Azerbaijan province, located in the northwest of Iran (Fig. 1). The monthly precipitation data of the Tabriz station during the base period (1951-2017) prepared by the Meteorological Organization of East Azerbaijan were utilized in the current study. To develop the proposed downscaling model, large-scale GCMs data in the base period (1951-200) and forecasting period (2020-2060) were from Can-ESM2 and BNU-ESM GCMs developed respectively in research centers of Canada and China. Future climate variables were extracted under RCP4.5 and RCP8.5 scenarios. Predictors on four grid points around the study area (i.e., 1, 2, 3, 4) were adopted in the current study (Fig. 1).

2-2- Proposed Methodology

In order to downscale GCM data, the ensemble AI-based models (i.e., ANN and ANFIS) was used, and to increase the efficiency of the AI-based downscaling model, pre-processing on variables of GCMs over the four nearest grid points to

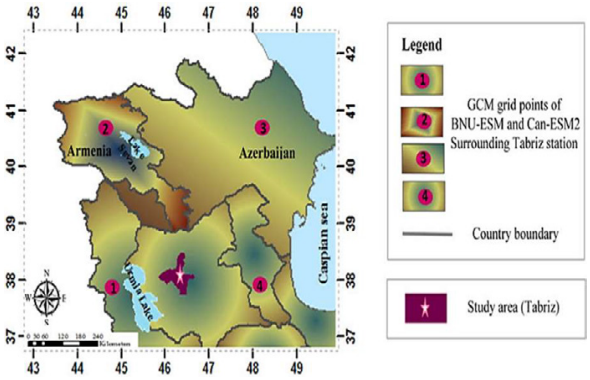

Fig. 1. Study area with the grid points of BNU-ESM and Can-ESM2

study area was performed. The proposed methodology includes three steps. The first step is a data pre-processing step, where the important GCMs for study area are determined and then, dominant predictors of selected GCMs are identified by two feature extraction methods (i.e., MI, decision tree). The second step is training an AI-based statistical downscaling model with the inputs obtained from the first step, and finally, the third step is the projection of future precipitation of Tabriz station according to ensemble AI-based model under RCPs 4.5 and 8.5.

The following subsections briefly describe the utilized mathematical tools in the proposed methodology.

2-3-Mutual Information (MI)

The entropy and information content has been mathematically formulated and measured using the distribution of data probability. It has been introduced as a measurement criterion of disorder, turbulence, and uncertainty [3]. In general, the theory of Shannon's information content is used in the discrete or continuous form based on the data and problem nature. MI between two random variables of X and Y is computed by the following equation [4]:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \quad (1)$$

Where $H(X)$ and $H(Y)$ are the entropies of X and Y respectively and $H(X,Y)$ is the common entropy of X and Y which is calculated by equation (2 and 3) [5, 6]:

$$H(X) = -\sum_{i=1}^{N} P(X_i)\log[P(X_i)] \quad (2)$$

$$H(X,Y) = -\sum_{i=1}^{N}\sum_{i=1}^{N} P(X_i,Y_i)\log[P(X_i,Y_i)] \quad (3)$$

2-4-Decision Tree

The decision tree method with a supervised learning approach is a powerful tool to perform classification and prediction in the data mining field. Generally, the structure of a tree is composed of four parts, including root, branch, node, and leaf. The root (i.e., the first node) locates at the upper level while the end chain of branches and nodes are leaves (i.e., end node). The root of the tree is the best node for the classification, and other variables in the lower nodes of the decision tree catch less importance. The M5 model tree as a decision tree method is maximizing the reduction of the standard deviation of each class of data that has been acquired in each node. The reduction of standard deviation is computed as [7]:

$$sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \quad (4)$$

2-5-Artificial Neural Network (ANN)

For statistical downscaling of GCM outputs in the current study, the three-layer Feed-Forward Neural Network (FFNN) structure is utilized. The previous studies indicated that FFNN with backpropagation algorithm could lead to reliable outcomes in predicting and simulation of hydro-climatologic variables [8]. For more information about the mathematical basis of neural networks, the readers are referred to [9].

2-6- Adaptive Neuro-Fuzzy Inference System (ANFIS)

ANFIS, as a neuro-fuzzy model, combines the neural network and fuzzy logic concepts to enjoy the benefits of both within a unique framework. Any ANFIS is consists of a five-layer neural network. The first layer is used for the input fuzzification. In the second layer, the fuzzy rule performance weight is calculated. The third layer is the normalization layer. In the fourth layer, the consequent rule values are calculated and multiplied by the respective rule performance weight and the fifth layer does the defuzzification (for more details see [10]).

2-7- Model Ensembling

It is well known that a combination of different predictors as a post-processing approach can improve overall predictions for a time series. In this sense, using combined predicts is safer and less risky than relying on a single method [11, 12].

2-8-Evaluation Criteria

Due to the adequacy of DC and RMSE in hydro-climatology prediction processes as evaluation criteria of models, in this study, to evaluate the performance of the downscaling model, in the training and validation process, two DC and RMSE criteria were used.

$$DC = 1 - \frac{\sum_{i=1}^{n}(O_i - Y_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O}_i)^2} \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - Y_i)^2}{N}} \quad (6)$$

Where O_i and Y_i are predictand and predictors, respectively, $\bar{O}$ is mean of predictand and N is the total number of observations.

3- RESULTS AND DISCUSSION

In the current study, to accurately assess the pattern of precipitation of Tabriz station in the future (2020-2060), GCM-based downscaling was applied. In this way, a robust

Table 1. dominant predictors based on MI and decision tree

Model	Feature extraction method	(Dominant predictors)[a]
BNU-ESM	MI	ua(850)$^{(2)}$, ta(850)$^{(1)}$, ua(200)$^{(1)}$,ta(850)$^{(3)}$
Can-ESM2		ta(500)$^{(3)}$, ta(500)$^{(1)}$, tas$^{(3)}$, ua(200)$^{(3)}$
BNU-ESM	Decision tree	ta(200)$^{(3)}$,ts$^{(3)}$,ta(1000)$^{(2)}$,pr$^{(1)}$
Can-ESM2		Prc$^{(3)}$,hus(850)$^{(3)}$,tas$^{(2)}$,ta(500)$^{(2)}$
Grid numbers based on Figure1 [a] (a=1, 2, 3, 4)		

Table 2. downscaling results of precipitation

GCMs	Feature extraction methods of Inputs	Downscaling model	training		testing	
			DC	N[1]-RMSE	DC	N[1]-RMSE
BNU-ESM	MI	ANN	0.65	0.11	0.61	0.10
		ANFIS	0.50	0.13	0.46	0.12
		Ensemble model	0.79	0.09	0.76	0.08
	Decision tree	ANN	0.79	0.09	0.77	0.08
		ANFIS	0.66	0.11	0.64	0.10
		Ensemble	0.88	0.06	0.84	0.07
Can-ESM2	MI	ANN	0.69	0.11	0.67	0.09
		ANFIS	0.54	0.13	0.51	0.11
		Ensemble model	0.85	0.07	0.83	0.07
	Decision tree	ANN	0.87	0.07	0.83	0.07
		ANFIS	0.72	0.10	0.66	0.10
		Ensemble model	0.94	0.05	0.91	0.05
N[1]-RMSE denotes normalized RMSE values.						

pre-processing method coupled with downscaling ensemble model resulted in predictand pattern recognition. Since the proposed methodology includes three steps, the results are presented and discussed in three steps as follow:

3-1-Results of First Step (Input screening)

The selection of the appropriate GCMs for the case study was performed based on MI calculation between monthly observed and GCM data from 1951-2005. Two GCMs i.e., Can-ESM2 and BNU-ESM with higher nonlinear relations based on MI were selected and utilized for the modeling. After assigning the important GCMs, dominant predictors of selected GCMs were determined by MI and decision tree feature extraction methods (see Table 1).

The dominant predictors obtained based on the MI method were zonal wind (a) and temperature-related variables (i.e., ta, tas and ts), which gained based on the nonlinear relation between observed precipitation with temperature and ua variables. The dominant predictors based on the decision were humidity and temperature-related predictors (i.e., hus, pr, prc, ts, ta, tas). Evaluation of dominant predictors selected by the decision tree method indicated that the decision tree model acted as an aggregation of CC and MI methods, due to selection of humidity and temperature-related variables like CC and MI, respectively.

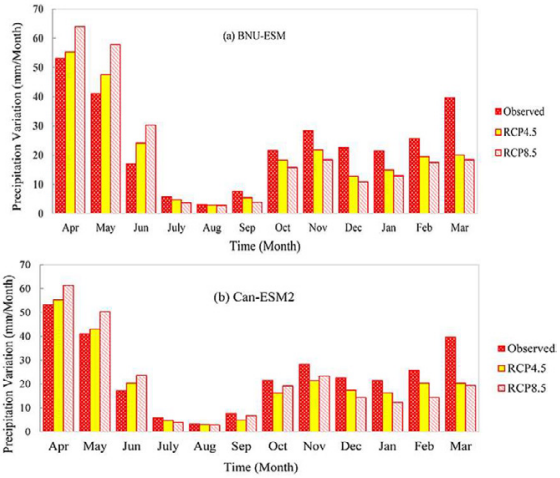

Fig. 2. Mean monthly observations and predicted precipitation for the future (2020-2060) under RCP4.5 and RCP8.5 scenarios.

3-2-Results of Second Step (Downscaling)

In both GCMs based on MI and decision tree, four dominant climatic variables were selected as inputs of the AI-based downscaling model in the first step. 75% of the dominant predictors and observed precipitation data were utilized for the training and the remaining 25% utilized for the validating purpose. The proper outcomes of precipitation downscaling according to DC and RMSE evaluation criteria, using ANN, ANFIS and ensemble models, are tabulated in Tables 2. The results of the ensemble methods indicated that ensemble models produced better outcomes concerning the single models in precipitation downscaling. So that, ensemble AI-based downscaling models methods increased the performance of AI-based modeling (according to the obtained DC values) up to 10%-65% in the verification step. The proper performance of the ensemble model compared to the single downscaling models indicated that the ensemble model covers the uncertainties in each model and lead to take advantage of both models for the AI-based downscaling models. According to Table 2, the decision tree as a multi-linear method, performs better than nonlinear MI method, because decision tree can accumulate and handle both linear and non-linear properties together in selecting the dominant predictors.

3-3-Results of Third Step (precipitation projection for future)

In the third step, monthly precipitation of Tabriz station was projected for the future (2020-2060) under RCP4.5 and RCP8.5 scenarios for Can-ESM2, BNU-ESM GCMs. According to the results of the second step, the downscaling ensemble model with inputs obtained via the decision tree method was selected as the benchmark model, and utilized in the projection step. To assess the projection results, the

predicted precipitation results for the period 2020-2060 under RCP4.5 and RCP8.5 scenarios are indicated versus the observed precipitation (1951-2005) (Fig. 2). As it is shown in Fig. 2, the study area in the future will experience both decrease and increase in precipitation values. It seems the precipitation decrease in the cold months of the year is due to the increase of pollutant density. The spread of particles and pollutions don't let the condensation of many minute vapor particles. So, it prevents the formation of raindrops.

4- CONCLUSIONS

To assess the future precipitation of the Tabriz city AI-based ensemble model (i.e., ANN and ANFIS) was used for the downscaling of GCMs outputs. The dominant predictors were determined via the decision tree and MI feature extraction methods. The advantage of using the proposed predictor screening methodology was selecting the dominant predictors among huge data sets without considering the non-stationary effects of climate variables. The comparison of the ensemble AI-based downscaling models and single models indicated that ensemble models produced better approximation than the single downscaling models and model combination improved the modeling performance by up to 24%-65%. Also, feature extraction methods indicated that the decision tree could be more reliable than the MI in downscaling precipitation at the study station. Finally, the selected feature extraction methods incorporate with downscaling ensemble model indicated reliable results.

REFERENCES

[1] Rezaei, M., Nahtani, M., Abkar, A., Rezaei, M. and Mirkazehi Rigi, M., 2013. "The survey of the efficiency of SDSM for predicting temperature parameters in two dry and superhero climates (Case study: Kerman and Bam)". *Watershed Management Research*, pp. 117-131. (In Persian).

[2] Le Roux, R., Katurji, M., Zawar-Reza, P., Quénol, H., and Sturman, A., 2018. "Comparison of statistical and dynamical downscaling results from the WRF model" *Environmental Modelling & Software*, 100, pp. 67-73.

[3] Shannon, C.E., 1948. "A mathematical theory of communications I and II" *Bell Labs Technical Journal*, 27, pp. 379-423.

[4] Yang, H.H., Van Vuuren, S., Sharma, S., and Hermansky, H., 2000. "Relevance of time-frequency features for phonetic and speaker-channel classification" *Speech Communication*, 31(1), pp. 35-50.

[5] Singh, V.P., 2011. "Hydrologic synthesis using entropy theory" *Journal of Hydrologic Engineering*, 16(5), pp. 421-433.

[6] Gao, Z., Gu, B., and Lin, J., 2008. "Monomodal image registration using mutual information-based methods" *Image and Vision Computing*, 26(2), pp. 164-173.

[7] Pal, M., and Deswal, S., 2009. "M5 model tree-based modeling of reference evapotranspiration" *Hydrological Processes: An International Journal*, 23(10), pp. 1437-1443.

[8] Maier, H.R., and Dandy, G.C., 2000. "Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications" *Environmental Modelling & Software*, 15(1), pp. 101-124.

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

[10] Jang, J.S.R., Sun, C.T., and Mizutani, E., 1997. *Neuro-Fuzzy and Soft Computing; a Computational Approach to Learning and Machine Intelligence*. Prentice-Hall.

[11] Sharghi, E., Nourani, V., and Behfar, N., 2018. "Earthfall dam seepage analysis using ensemble artificial intelligence-based modeling" *Journal of Hydroinformatics*, 20(5), pp. 1071-1084.

[12] Jang, J.S., 1993. "ANFIS: adaptive-network-based fuzzy inference system" *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3), pp. 665-685.

HOW TO CITE THIS ARTICLE

Z. Razzaghzadeh, V. Nourani, N. Behfar, The conjunction of the feature extraction method with AI-based ensemble statistical downscaling models, Amirkabir J. Civil Eng., 52(4) (2020) 219-222.

DOI: 10.22060/ceej.2018.14986.5806