Journal of Hydrology 492 (2013) 1–14

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Performance assessment of different data mining methods in statistical downscaling of daily precipitation

M. Nasseri [a,*], H. Tavakol-Davani [a], B. Zahraie [b]

[a] School of Civil Engineering, University of Tehran, Tehran, Iran
[b] Center of Excellence for Engineering and Management of Civil Infrastructures, School of Civil Engineering, University of Tehran, P.O. Box 11155-4563, Tehran, Iran

ARTICLE INFO

Article history:
Received 16 April 2012
Received in revised form 7 April 2013
Accepted 9 April 2013
Available online 18 April 2013
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Sheng Yue, Associate Editor

Keywords:
Statistical downscaling
Nonlinear data-mining method
Climate change

SUMMARY

In this paper, nonlinear Data-Mining (DM) methods have been used to extend the most cited statistical downscaling model, SDSM, for downscaling of daily precipitation. The proposed model is Nonlinear Data-Mining Downscaling Model (NDMDM). The four nonlinear and semi-nonlinear DM methods which are included in NDMDM model are cubic-order Multivariate Adaptive Regression Splines (MARS), Model Tree (MT), *k*-Nearest Neighbor (*k*NN) and Genetic Algorithm-optimized Support Vector Machine (GA-SVM). The daily records of 12 rain gauge stations scattered in basins with various climates in Iran are used to compare the performance of NDMDM model with statistical downscaling method. Comparison between statistical downscaling and NDMDM results in the selected stations indicates that combination of MT and MARS methods can provide daily rain estimations with less mean absolute error and closer monthly standard deviation and skewness values to the historical records for both calibration and validation periods. The results of the future projections of precipitation in the selected rain gauge stations using A2 and B2 SRES scenarios show significant uncertainty of the NDMDM and statistical downscaling models.

1. Introduction

Outputs of Global Circulation Models (GCMs) are the base of climate change studies. Spatial resolution of these data is not enough to determine local climate change effects and they must be recalculated to a suitable resolution to be valid for local meteorological analysis. The methods of extracting regional scale meteorological variables from GCM outputs have been known as downscaling approaches. Four general categories of downscaling approaches include regression (empirical) methods (Enke and Spekat, 1997; Faucher et al., 1999; Li and Sailor, 2000; Wilby et al., 2002; Hessami et al., 2008; Raje and Mujumdar, 2011), weather pattern approaches (Bárdossy and Plate, 1992; Yarnal et al., 2001; Bárdossy et al., 2002; Wetterhall et al., 2009; Anandhi et al., 2011), stochastic weather generators (Semenov and Barrow, 1997; Bates et al., 1998) and regional climate models (Mearns et al., 1995).

Regression or empirical methods are the most cited approaches in downscaling simulation. Simplicity in use, relatively lower costs of pre-processing and straightforwardness of computational procedure are the main reasons of the popularity of these downscaling techniques.

Finding the empirical relationships between global and local scales of climate circulation is the basic statement of any statistical downscaling method. According to this assumption, correlation of global GCM meteorological variables (predictors) and local meteorological variables such as observed precipitation and temperature (predictands) is the key point of this type of downscaling procedure. The most well-known regression based downscaling methods are structured for separate estimation of occurrence and amount of meteorological variables. Advantages and disadvantages of statistical regression based downscaling methods have been comprehensively discussed by Hessami et al. (2008).

Different nonlinear Data Mining (DM) methods such as Artificial Neural Networks (ANNs) (Tomassetti et al., 2009; Pasini, 2009; Mendes and Marengo, 2010; Fistikoglu and Okkan, 2011), *k*-Nearest Neighbor (*k*NN) (Yates et al., 2003; Gangopadhyay et al., 2005; Raje and Mujumdar, 2011), Support Vector Machine (SVM) (Tripathi et al., 2006; Chen et al., 2010), Model Tree (MT) (Li and Sailor, 2000), Multivariate Adaptive Regression Splines (MARS) (Corte-Real et al., 1995) beside linear regression methods (Wilby et al., 2002; Hessami et al., 2008) have been used in the previous studies for climatological research.

Statistical downscaling model (SDSM) is the most cited concepts and packages among regression based statistical downscaling methods. This computer package benefits from Multiple Linear Regression (MLR) method to estimate the amount and/or the occurrence of local meteorological predictands.

In this paper, efficiency of four nonlinear and semi-nonlinear DM methods and their previous applications in climatological research, namely MARS, MT, *k*NN and Genetic Algorithm-optimized SVM (GA-SVM) have been evaluated versus application of the standard MLR in estimating both occurrence and amount of precipitation.

* Corresponding author. Tel.: +98 912 209 4881.
E-mail addresses: mnasseri@ut.ac.ir (M. Nasseri), h_tavakol@ut.ac.ir (H. Tavakol-Davani), bzahraie@ut.ac.ir (B. Zahraie).

0022-1694/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2013.04.017

In this article, the structure of SDSM has been used as the main platform to develop Nonlinear Data-Mining Downscaling Model (NDMDM) model by replacing their MLR kernels with the selected DM methods. In the next sections, local scale (areas of interest and predictands) and large scale datasets (predictors) which are used in this study are described. Then, SDSM and the utilized DM methods are briefly described. The next sections of the paper present the results of the case study and concluding remarks and recommendations for further studies.

2. Datasets

2.1. Local dataset

To assess the efficiency of the proposed downscaling method, twelve rain gauge stations scattered in five different climatological basins in Iran, namely Hamoon-Jazmoorian, Sefidrood, Mordab-Anzali, Shapoor-dalky and Mond are used. These basins are located in an arid region in southeast of Iran near Iran-Pakistan border, a wet region in north of Iran near Caspian Sea and a semi-arid region in southwest of Iran in Persian Gulf. Some statistical characteristics such as average, maximum and standard deviation of observed daily precipitation of the selected stations have been presented in Table 1. The locations of these rain gauge stations are also shown in Fig. 1. As presented in Table 1, 26–35 years of daily precipitation records up to the year 2000 (the start year of simulations of the climate change scenarios) are available for the selected rain gauge stations. For each station, the first 75% of the available record has been used for calibration of the downscaling model and the rest of the recorded data has been used for validation of the model. The daily precipitation records have been gathered from the Iran Water Resources Management Company.

2.2. Large scale datasets

The data bank of Hadley Center GCM, namely HadCM3, for A2 and B2 SRES (Special Report on Emission Scenarios) scenarios has been used in this study to project the future climate behavior. The coarse resolution ($2.5^\circ \times 2.5^\circ$) reanalysis of atmospheric data from the U.S. National Center for Environmental Prediction (NCEP) (Table 2) have been used as the downscaling model predictors.

Because of inconsistency of spatial resolution of HadCM3 outputs (3.75° (long.) $\times$ 2.5° (lat.)) and NCEP dataset, projection of large-scale predictors of NCEP on HadCM3 computational grid box has been used in this study. The daily projected data and HadCM3 outputs are available from the Canadian Climate Impacts Scenarios (CCIS) website (www.cics.uvic.ca/scenarios/sdsm/select.cgi).

Twenty-six different atmospheric variables are available for each grid box in this database. For each rain gauge station, nine boxes covering and around the study areas have been selected. Fig. 1 depicts center of each meteorological grid box and location of the selected rain gauge stations. As it is illustrated in this figure, the grid boxes cover a large area over the selected basins and around them. In addition, one to three-day lags of predictors have been considered as candidate model inputs to incorporate cross correlation and auto-correlation in the modeling process. For each station, 936 (9 (grid boxes) $\times$ 26 (meteorological predictors) $\times$ 4 (0 to 3-day time lag)) predictors have been analyzed.

3. Methodology

In the current section at the first, platform of SDSM has been described. Then different data-mining methods which are used in NDMDM are described and at the end, structure of NDMDM is explained.

3.1. Statistical downscaling model (SDSM)

SDSM software is developed based on Multiple Linear Regression Downscaling Model (MLRDM) (Wilby et al., 2002). SDSM outputs are the average of several weather ensembles which are the results of using linear regression models with stochastic terms of bias correction. Because of the linear structure of SDSM, selection of predictors is based on the correlation and partial correlation analysis between the predictand and predictors and weights of the predictors which are estimated via simple least square method. Dual simplex method has been also provided in SDSM because of instability of regression coefficients for non-orthogonal predictor vectors. Hessami et al. (2008) added a new option of using ridge regression (Hoerl and Kennard, 1970) in their downscaling model, namely ASD as a remedy of the non-orthogonality impact of the predictor vectors as well (Hessami et al., 2008).

SDSM model contains of two separate sub-models to determine occurrence and amount of conditional meteorological variables (or discrete variables) such as precipitation and amount model for unconditional variables (or continues variables) such as temperature or evaporation. Statistical downscaling using SDSM consists of the following steps:

1. In first step, suitable predictors should be selected. SDSM provides the ability of some statistical analysis for users to select the best predictors. In SDSM, predictors should have acceptable unconditional and conditional correlations with the predictand. Also, partial correlation, *P*-value and explained variance of the

Table 1
Basic information about 12 rain gauge stations (Max.=Maximum and Std.=Standard deviation).

No.	Station code	Station name	Abbr.	Basin	Length of dataset (year)	Longitude (°E)	Latitude (°N)	Statistical characteristics of observed daily rainfall (mm)		
								Mean	Max.	Std.
1	44-014	Delfard	Del.	Jazmoorian	1975–2000	57.60	29.00	1.20	150	6.31
2	44-009	Dehrood	Deh.	Jazmoorian		57.73	28.87	0.76	132	4.71
3	44-016	Khoramshahi	Kho.	Jazmoorian		57.75	29.00	1.32	194	6.95
4	44-024	Kharposht	Khar.	Jazmoorian		57.83	28.48	0.46	80	3.20
5	17-082	Rasht	Ras.	Sefidrood	1966–2000	49.60	37.25	3.58	188	10.34
6	17-075	Farshekan	Far.	Sefidrood		49.58	37.40	3.30	168	9.93
7	18-007	Kasma	Kas.	Mordab-anzali		49.30	37.31	3.01	317	9.59
8	18-017	Shanderman	Shan.	Mordab-anzali		49.11	37.41	2.67	177	8.23
9	24-033	Khanzanian	Khan.	Mond	1972–2000	52.15	29.67	1.27	92	5.26
10	23-011	Shapoor	Shap.	Shapoor-dalky		51.11	29.58	0.85	75	4.45
11	23-019	Shoorjareh	Shoo.	Shapoor-dalky		51.98	29.25	0.99	120	4.92
12	43-034	Arsanjan	Ars.	Shapoor-dalky		51.30	29.92	0.87	111	4.56