Theoretical and Applied Climatology (2020) 142:1255–1269
https://doi.org/10.1007/s00704-020-03372-4

ORIGINAL PAPER

Identification of potential causal variables for statistical downscaling models: effectiveness of graphical modeling approach

Riya Dutta[1] · Rajib Maity[1]

Received: 3 August 2018 / Accepted: 3 September 2020 / Published online: 12 September 2020

Abstract

Selection of potential causal variables (PCVs) from a pool of many possibly associated variables is a critical issue since it can significantly affect the performance of any statistical downscaling model. Generally, the variable to be downscaled is associated with many other hydrologic and climatic (aka hydroclimatic) variables. Most of the existing approaches, such as correlation analysis (CA), partial correlation analysis (PaCA), and stepwise regression analysis (SRA), rely mostly on the mutual association for the selection of PCVs. However, none of these approaches investigate the detailed dependence structure that may be helpful in eliminating the unwanted information and efficiently selecting the PCVs for downscaling the target variable. In this study, the effectiveness of graphical modeling (GM) approach is explored for the selection of the PCVs as GM can effectively identify the detailed conditional independence structure among all the associated variables. For demonstration, downscaling of monthly precipitation is undertaken using the PCVs, identified by CA, PaCA, SRA, and the proposed GM approach. Two different downscaling models, namely statistical downscaling model (SDSM) and support vector regression (SVR)–based downscaling model, are utilized. The results show that the PCVs identified through the proposed GM approach provides consistent as well as robust performance, across different regions and seasons, due to its ability to capture the complete conditional indepedence structure among the variables. The downscaled monthly precipitation obtained using the proposed approach is better matching with the observed data in terms of the mean, variance as well as the probability distribution. Overall, this study recommends the GM approach for the identification of the PCVs for the downscaling models.

Keywords Statistical downscaling · Potential causal variable selection · Graphical modeling · Correlation analysis · Partial correlation analysis · Stepwise regression analysis

1 Introduction

Downscaling is a general procedure to assess the information of any hydroclimatic variable at a finer scale using the information of the same and other variables at a coarser scale. The downscaling methods are broadly categorized into statistical and dynamical approaches (Wilby et al. 1999; Bergströms et al. 2001; Fowler et al. 2007; Schoof et al. 2009; Pinto et al. 2010). Statistical downscaling methods are less computationally intensive as compared with the dynamical methods, often the reason for its popularity (Wilby et al. 2002; Fowler et al. 2007; Chen et al. 2012; Meenu et al. 2013; Gutmann et al. 2014; Tatsumi et al. 2015; Zuo et al. 2015). There are many statistical downscaling methods based on various algorithms including automated statistical downscaling (ASD), artificial neural network (ANN), LARS-WG stochastic weather generator, non-homogeneous hidden Markov model (NHMM), statistical downscaling model (SDSM), support vector machine (SVM), Bias Corrected Spatial Disaggregation (BCSD), Asynchronous Regression (AR), LOcalized Constructed Analogs (LOCA), and so on (Bates et al. 1998; Semenov et al. 1998; Wilby et al. 1999; Dettinger et al. 2004; Wood et al. 2004; Coulibaly and Baldwin 2005; Hessami et al. 2008; Chen et al. 2012; Stoner et al. 2013; Pierce et al. 2014).

Basic principle of statistical downscaling is to identify the statistical relationship between the target variable to be downscaled and various coarse resolution causal variables (Beuchat et al. 2012), and then apply the established relationships to downscale the target variable. In general, the target variable

✉ Rajib Maity
rajib@civil.iitkgp.ac.in; rajibmaity@gmail.com

[1] Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

is associated with a large number of climatic variables (Kidson and Thompson 1998; Charles et al. 1999; Wilby et al. 1999), hereafter referred to as *associated variables*. An important aspect of downscaling is to find a subset of associated variables which are potentially useful causal variables, hereafter referred to as the potential causal variables (PCVs), for developing a statistical relationship. This study considers this issue of identifying the PCVs for statistical downscaling models. This is a challenging task because the choice may vary with region and the target variable to be downscaled (Huth 1999; Tomozeiu et al. 2007; Hessami et al. 2008).

As per literature, the commonly used associated variables are air temperature, geo-potential height, specific humidity, pressure, and zonal/meridional wind speed at different pressure levels (Anandhi et al. 2008; Devak and Dhanya 2014; Pichuka and Maity 2016; Chithra and Thampi 2017). Inclusion of all possible associated hydroclimatic variables yields either a prohibitively large number of variables in the causal variables set resulting in highly complex downscaling models which poses serious challenges in parameter estimation, or ends up in possibly incomplete subset of the associated variables. Most of the selection criteria for shortlisting the PCVs are based on the vaguely understood physical processes of the system and/or correlation between the causal and target variable (Grimes et al. 2003; Tatli et al. 2004; Haylock et al. 2006; Hessami et al. 2008). For instance, a widely used approach to select the PCVs is correlation analysis (CA) (Chen et al. 2012; Meenu et al. 2013; Hassan et al. 2014; Pervez and Henebry 2014). Partial correlation analysis (PaCA) is another commonly used method to identify relevant causal variables (Hessami et al. 2008; Liu et al. 2011). This method excludes/partials out the effect of other variables and can thereby reveal the true correlation between two variables of interest (Harpham and Wilby 2005; Yang et al. 2011). Some researchers, such as Huth (1999), Hessami et al. (2008), and Chen et al. (2011), have also used stepwise regression analysis (SRA) to select the causal variables for statistical downscaling. In order to reduce the complexity of the downscaling model, certain studies rely on reducing the dimensionality of the associated variables by using techniques such as principal component analysis (Okkan and Inan 2015), and adaptive nonlinear interaction structures in higher dimensions (Radchenko and James 2010). These techniques are good in cases where a well-defined set of associated variables are known; however, such knowledge is either vague or incomplete. Moreover, it also remains unknown whether the same information is provided by more than one causal variable, also known as redundancy in information. Thereby, the existing techniques used to identify the PCVs are either unable to avoid the redundant information from multiple associated variables or miss out important variables due to the complex nature of association. When multiple variables are associated with a hydroclimatic variable, complete information on the conditional independence structure is essential in order to obtain a well-defined set of PCVs. Only the directly associated variables may be picked out to be used in the downscaling model, leaving out the effect of conditionally independent and independent variables. Here lies the potential of graphical modeling (GM) that provides a means of representing dependence structure among a large number of associated variables (Jordan 2004; Bang-Jensen and Gutin 2007; Ihler et al. 2007; Whittaker 2009). This forms the motivation of this study.

Algorithms that search for such dependence structure are typically represented using graphical models and have been used widely in the fields of statistics, machine learning, and the social and natural sciences (Beal et al. 2003; Lauritzen and Sheehan 2003; Jordan 2004; Bang-Jensen and Gutin 2007; Whittaker 2009; Krumsiek et al. 2011). GM approach may be highly beneficial in the field of hydrology and hydroclimatology where numerous variables are associated in complex ways with incomplete knowledge of the dependence structure. Some recent studies have shown the efficacy of GM in identifying the complete conditional independence structure especially in the field of hydroclimatology (Taeb et al. 2017; Dutta and Maity 2018; Dutta and Maity 2020a, b). The objective of this study is to explore the efficacy of GM, in context of downscaling, to identify the dependence structure among the associated variables for the identification of the PCVs. Different sets of PCVs identified through the proposed and other existing approaches, (i.e., CA, PaCA, and SRA) are used for the downscaling of monthly precipitation. Two different downscaling approaches, namely, statistical downscaling model (SDSM) and support vector regression (SVR)–based downscaling model, are utilized in this study. Performances of different PCV identification approaches are compared by comparing the quality of the downscaled precipitation. Better match between monthwise mean, probability distributions, and variance of observed and downscaled precipitation is considered the better quality of downscaled precipitation, which in turn help to recommend the best PCV identification approach.

2 Methodology

Identification of the PCVs for the statistical downscaling using GM approach (henceforth “proposed GM approach” or simply “proposed approach”) is facilitated by identification of a conditional dependence structure among the associated variables and the target variable to be downscaled. Detailed description of the proposed approach is discussed in the following section. However, the methodologies for other existing approaches are presented in Appendix A.

2.1 Selection of the potential causal variable

A graph can be defined as a mathematical object, $G = (V, E)$, where V is the set of vertices or nodes and E is the set of edges (Whittaker 2009). Each variable is associated with a node and each edge is associated with a pair of nodes. The graph/conditional independence structure of a set of random variables is defined by a set of pairwise conditional independence relationships that determine the edge set of the graphs. Edges in the graph represent dependencies between random variables and presence/absence of edges represents dependence/independence between random variables.

The selection/identification of the dependence structure among the associated variables and target variable is determined using the maximum likelihood approach (Whittaker 2009). For application of this approach, the data should follow normal distribution. In case the data does not follow normal distribution, they can be transformed using some transformation methodology (e.g., Box and Cox 1964) before developing the conditional independence structure. In the maximum likelihood approach, initially a fully interconnected graph structure, referred to as a saturated model is considered where all the nodes are connected to each other. Next, the edge exclusion deviance (EED) is used for testing if an edge can be eliminated from the saturated model (Whittaker 2009). EED is formulated as follows,

$$\mathrm{EED} = -n\log\left(1-\mathrm{corr}_n^2\left(X_i, X_j|\mathrm{rest}\right)\right) \tag{1}$$

where $\mathrm{corr}_n^2\left(X_i, X_j|\mathrm{rest}\right)$ is the partial correlation coefficient between any two random variables X_i, X_j given the rest and n is the sample size. Considering EED to follow chi-square distribution, at 95% confidence level for one degree of freedom (as one edge is removed at a time) the p value is 3.84, so the edges for which the EED does not reach the value of 3.84 (threshold value) are to be excluded.

The generalized likelihood ratio test statistics can be evaluated based on the distance between the observed sample variance and the estimated variance obtained from the graph structure. This test statistic, also known as the deviance (dev), can be used to check the acceptability of the obtained conditional independence structure at a particular confidence level. The deviance of the model can be evaluated using the following equation (Whittaker 2009),

$$\mathrm{dev} = n\left\{ tr\left(S\widehat{V}^{-1}\right) - \log\det\left(S\widehat{V}^{-1}\right) - k \right\} \tag{2}$$

where S is the variance matrix, $\widehat{V}$ is the modified/estimated variance matrix, k is the number of random variables, and n is as stated above. Considering the test statistic/deviance to follow chi-square distribution with d degrees of freedom (number of edges excluded from the saturated graph structure), the significance level (p value) can be computed as $P\left(\chi_p^2 > \mathrm{dev}\right)$. For this study, the acceptable significance level is fixed at 0.05, i.e., the graph structure is acceptable with 95% confidence level if the p value is higher than 0.05. In case the structure fails to meet the acceptability criteria, a new graph structure with lesser number of edges is to be identifed.

Surviving edges for the finally obtained graph structure can be tested for their strength of association, also known as edge strength. The divergence against conditional independence can be used as the edge strength. The edge strength between two nodes in the conditional independence/graph structure can be calculated as follows (Whittaker 2009),

$$\mathrm{Inf}\left(X_i \amalg X_j|\mathrm{rest}\right) = -\frac{1}{2}\log\left(1-\mathrm{corr}_n{}^2\left(X_i, X_j|\mathrm{rest}\right)\right) \tag{3}$$

where $\mathrm{Inf}(X_i \amalg X_j|\mathrm{rest})$ is the edge strength between X_i and X_j given rest.

The conditional independence structure provides the information on dependent (directly connected/parents to the target variable), independent (not connected to the target variable), and conditionally independent (not directly connected to the target variable) causal variables with respect to the target variable. Thereby, it helps to identify the parents of the target variable and these variables are selected as the PCVs for the statistical downscaling model. It may be noted here that the variables were transformed only to develop the graph structure. After selection of the PCVs, the actual data is used in the downscaling model, not the transformed data.

2.2 Assessment of downscaling performance

As mentioned before, two different downscaling approaches are employed to downscale monthly precipitation considering the PCVs identified using the proposed GM approach and other existing approaches. In the downscaling process, the model is calibrated (1960–1995) and validated (1996–2005) separately, and the analysis is carried out for each station using the separately identified PCVs, using different approaches. The first downscaling model, SDSM, is constructed based on multiple regression equations, given the target variable (monthly precipitation) and PCVs (selected coarse resolution atmospheric variables) during calibration period. The parameters of the developed regression model are saved and further used during validation period. Model validation produces 20 ensembles of downscaled monthly precipitation at each station given the respective PCVs and the parameters of the developed regression model. The second downscaling model, SVM for regression, also known as SVR, is used to downscale the monthly rainfall considering the PCVs identified using the different approaches. Details of the methodology on SVR can

Table 1 Details of the meteorological stations along with the basic statistics of the rainfall data at each station

Station ID	Station name	District (state)	Latitude and longitude	Description of location	Basic statistics				
					Mean (mm)	Range (min to max in mm)	Standard deviation (mm)	Coefficient of variation	Skewness
S1	Chiplima	Sambalpur (Orissa)	21.36° N and 83.88° E	Central Highlands	109	0 to 636	148	1.35	1.47
S2	Jaisalmer	Jaisalmer (Rajasthan)	26.90° N and 70.92° E	Dessert	15	0 to 201	28	1.91	2.97
S3	Jalore	Jalore (Rajasthan)	25.38° N and 72.99° E	Dessert	36	0 to 335	65	1.83	2.30
S4	Gulbarga	Gulbarga (Karnataka)	17.35° N and 76.80° E	Deccan Plateau	59	0 to 480	70	1.20	1.69
S5	Gadchiroli	Gadchiroli (Maharashtra)	20.17° N and 79.98° E	Deccan Plateau	88	0 to 453	112	1.27	1.17
S6	Pasighat	East Siang (Arunachal Pradesh)	28.10° N and 95.37° E	North East	155	0 to 640	140	0.91	0.81
S7	Badaun	Badaun (Uttar Pradesh)	28.01° N and 79.11° E	Gangetic Plain	64	0 to 455	98	1.53	1.83
S8	Bardhaman	Bardhaman (West Bengal)	23.21° N and 87.88° E	Gangetic Plain	117	0 to 633	131	1.13	1.24

be found in Chen et al. (2010). While developing the SVR model during calibration period, the goal is to find a function $y = f(x)$ such that any observation (y) does not deviate from the simulated/predicted value ($\hat{y}$) by more than a threshold value ε, known as ε-margin, for the corresponding causal/input data (x). This relation/function is further utilized during model validation to obtain the downscaled precipitation.

Efficacy of the proposed approach is established by comparing the observed and downscaled monthly precipitation in terms of mean, variance and probability distribution. Whereas the variances are compared considering all the months in the year, comparison between mean are undertaken month-wise using the Wilcoxon test. The mean error in downscaled monthly precipitation is evaluated as the absolute difference between the monthly observed and downscaled data. The ensemble mean of the downscaled target variable is used to evaluate the monthly mean error. The mean error is further tested using Wilcoxon rank-sum test, a non-parametric approach to establish the significant difference between two samples of data using magnitude based rank (Johnson and Bhattacharya 2009; Maity 2018). In this study, the Wilcoxon test is used to study the difference between the observed and downscaled monthly precipitation (each ensemble) for each month of analysis. The null hypothesis considered for the test is that the population means (designated as μ_1 and μ_2) are equal and the alternative hypothesis is that the population means are not equal. Assuming the sample size to be large, the rank sum test uses a Z-statistic that follows standard normal distribution. The test statistic can be evaluated as,

$$Z = \frac{W - n_1(N+1)/2}{\sqrt{n_1 n_2 (N+1)/12}} \tag{4}$$

where n_1 and n_2 are the number of data points of μ_1 (mean for downscaled results) and μ_2(mean for observed data) respectively, $N = n_1 + n_2$, and W is the sum of ranks for μ_1. Based on the value of the test statistic, p value is obtained for each ensemble of the downscaled data. The average values across the ensembles are considered the final p value for each month. Considering a significance level (α) of 0.05, the null

Table 2 Details of the entire set of possibly associated variables

Sl. no.	Description of the variables (units)	Code of the variable
1	925 mb air temperature (°K)	ta_925
2	700 mb air temperature (°K)	ta_700
3	500 mb air temperature (°K)	ta_500
4	200 mb air temperature (°K)	ta_200
5	925 mb geopotential height (m)	zg_925
6	500 mb geopotential height (m)	zg_500
7	200 mb geopotential height (m)	zg_200
8	925 mb specific humidity (kg/kg)	huss_925
9	850 mb specific humidity (kg/kg)	huss_850
10	925 mb zonal wind (m/s)	ua_925
11	500 mb zonal wind (m/s)	ua_500
12	200 mb zonal wind (m/s)	ua_200
13	925 mb meridional wind (m/s)	va_925
14	500 mb meridional wind (m/s)	va_500
15	200 mb meridional wind (m/s)	va_200
16	Surface pressure (Pa)	ps
17	Precipitable water (kg/m^2)	pr

Table 3 Potential causal variables (PCVs) selected based on edge strength (ES) using the proposed GM approach

Station ID	PCV	ES
S1	pr	0.84
	ta_500	0.81
	ta_700	0.70
	huss_925	0.13
	huss_850	0.12
	zg_200	0.09
	ua_500	0.08
S2	ta_700	0.95
	zg_500	0.96
	zg_200	0.81
	huss_925	0.71
	ua_500	0.67
	pr	0.52
S3	pr	0.70
	ta_200	0.69
	zg_500	0.14
	zg_200	0.11
	ps	0.10
S4	va_200	0.95
	ps	0.94
	pr	0.80
	ta_500	0.13
	ua_500	0.10
	zg_925	0.08
S5	ua_925	0.92
	ps	0.70
	pr	0.67
	zg_200	0.10
	huss_925	0.10
S6	ta_925	0.93
	ta_500	0.94
	ps	0.70
	va_200	0.69
	zg_925	0.12
	pr	0.09
S7	zg_200	0.97
	va_200	0.95
	ua_925	0.71
	pr	0.71
	ta_200	0.11
S8	ta_700	0.85
	zg_925	0.85
	ua_925	0.17
	pr	0.11
	zg_500	0.10
	va_500	0.07

hypothesis of equality in mean is rejected if the *p* value is less than 0.05.

3 Data used

Eight meteorological stations (designated as S1 through S8) are selected from different climatic regions in India and downscaling is carried out at each station to study the spatial variation in the selection of the PCVs. The time period of the analysis is 46 years where, the first 30 years (1960–1989) is used for model development and the next 16 years (1990–2005) for model testing. Description for all the stations along with the basic statistics (mean, range, standard deviation, coefficient of variation, and skewness) of the rainfall data at each station is provided in Table 1. It is noticed that the range and pattern of rainfall are very different from one another, including the stations lying in the same climatic region. For instance, stations S2 and S3 both lie in the dessert region; however, the range of rainfall for the former is 0–201 mm and the same for the latter is 0–335 mm. Next, considering the stations S4 and S5, the range of rainfall is approximately same. However, for station S4, the variation in the month-wise mean rainfall during the summer monsoon season (June–September) is very low (with the highest monthly mean of ~ 200 mm) and for station S5, the variation in the month-wise mean rainfall during the same season is very high (with the lowest and highest month-wise mean of ~ 180 mm and ~ 300 mm respectively). Lastly, considering stations S7 and S8, the range of rainfall for the former is 0–455 mm (with high variation in month-wise mean during summer monsoon season) and the same for the latter is 0–633 mm. Stations S1 and S6 also have a unique pattern of rainfall with highest month-wise means of ~ 400 mm.

For each station, monthly precipitation is used as the target variable to be downscaled, sourced from the India Water Portal (http://www.indiawaterportal.org/data). The coarse resolution associated variables used are air temperature, geopotential height, specific humidity, zonal wind, meridional wind, surface pressure, and precipitable water at different altitudes, obtained from World Data Center for HadCM3 (historical data) climate model (http://www.ipcc-data.org/sim/gcm_monthly/AR5/Reference-Archive.html). These initially selected variables are based on the physical interaction among the hydroclimatic variables as identified by various studies, primarily considering the Indian domain and the details for the same are given in Table 2. Owing to the varying range and pattern of rainfall as well as the geographical location, it is probable that the set of potential causal variables will be different from one location to another. For instance, the association of lower, middle, and upper tropospheric circulations and rainfall varies with geographical location and season. The high variation in

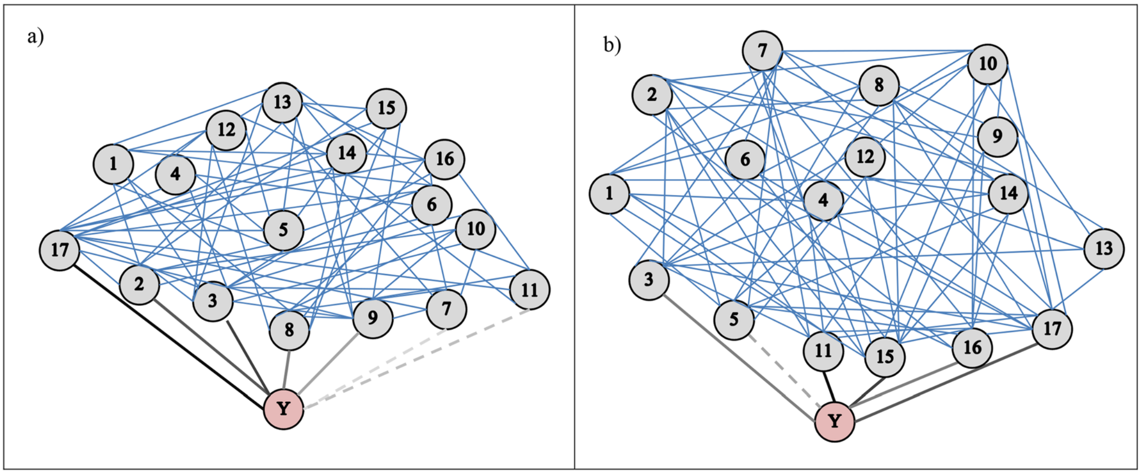

Fig. 1 Conditional independence structure, where numbers 1 to 17 represent the causal variables (details in Table 2) and *Y* is the target variable to be downscaled, **a** obtained for station S1, and **b** obtained for station S4. The edges between the target variable and its parents are shown with varying gradients of gray shade denoting the approximate strength of association. Dotted lines indicate even poorer association. The edges for parents not considered the PCVs (based on edge strength) are shown by dashed lines

summer monsoon rainfall may be caused by meridional wind, zonal wind, and temperature at different pressure levels due to monsoon circulations. Thereby, the causal variables at different altitudes may influence rainfall at a particular location to varying degrees depending of various factors.

4 Results and discussion

4.1 Selection of the PCVs using the proposed GM approach and comparison with other approaches

Two typical examples of conditional independence structures between all the associated and target variables are shown in Fig. 1 for stations S1 and S4. The parent variables, selected as the PCVs, are tabulated in Table 3 for all the stations (S1 to S8) along with the edge strengths. Results indicate that stations in the same zone have many common PCVs which indicates that the proposed approach can appropriately prioritize the causal variables for downscaling. Still there are some differences and different sets of PCVs are identified for stations in different climatic zones. This is expected due to the wide variation in the precipitation characteristics even within a climatic zone. As observed, stations S2 and S3 or S4 and S5 or S7 and S8, though lying in the same climatic zone, do not have the exactly same set of causal variables. Geographically, they are hundreds to thousands kilometer away (S2 to S3 ~ 250 km, S4 and S5 ~ 650 km, and S7 to S8 ~ 1000 km). The broad climatic zones (as shown in Table 1) are as per India-WRIS. However, precipitation varies at very small scale and every station will thereby have its unique features. Further, the range and pattern of rainfall are very different at each station. Following the pattern of monthly rainfall in the different stations (figure not shown), it is observed that the maximum variation in month-wise mean rainfall is observed for the summer monsoon months at stations S1, S2, S3, S5, and S7. However, at stations S4, S6, and S8, the month-wise mean rainfall during the summer monsoon months is similar to that during the pre- and post-monsoon months (October–May). For the former group of stations, GM identifies the upper- and mid-tropospheric circulations as the PCVs. The poleward retraction of the mid-tropospheric circulation and warming of the troposphere suggest that precipitation processes may be more directly linked to upper-tropospheric circulation, as it is common in convective and monsoon regimes. However, for the stations S4, S6, and S8, mostly the lower- and mid-tropospheric circulations show strong association with precipitations due to the equatorward migration of the mid-tropospheric flow during winter months. The moist condition of the mid-tropospheric air is an important factor in precipitation mechanisms, since moist air is associated with vertical motion and convective processes, especially during the warmer months (Cavazos and Hewitson 2005). It is interesting to note that stations S7 and S8, though lying in the same climatic regions, have different sets of PCVs due to different pattern of rainfall in the two stations. Furthermore, for the stations S1, S5, S7, and S8, temperature of the tropospheric layer and low-level zonal wind are identified as PCVs. These stations have high range and variation in month-wise mean rainfall during the monsoon months. Significance of these variables is apparent during this season when it is common for an expansion of the troposphere due to monsoonal circulation. Generally, in tropical convective regions, low-level convergence is accompanied by upper-level divergence (Webster et al. 1998). Though stations S4 and S5 lie in the same climatic region, the meridional wind and pressure also play an important role at station S4 in addition to temperature and zonal wind. This may be primarily due to the high variation in month-wise

Table 4 Potential causal variables (PCVs) selected using the four approaches (GM, PaCA, SRA, and CA). The code of the variables is given in Table 2

Station ID	PCV using different methods			
	GM	PaCA	SRA	CA
S1	ta_700 ta_500 zg_200 huss_925 huss_850 ua_500 pr	ta_700 huss_850 ua_500 ua_200 pr	zg_500 huss_925 huss_850 va_925 ua_500 pr	ta_500 ta_200 zg_200 huss_850 va_200
S2	ta_700 zg_500 zg_200 huss_925 ua_500 pr	ta_925 ta_700 ta_500 ta_200 zg_500 pr	ta_700 ta_200 zg_500 zg_200 ua_500 pr	ta_500 ta_200 zg_200 huss_850 ua_925 va_500 va_200 pr
S3	ta_200 zg_500 zg_200 ps pr	ta_500 zg_925 zg_200 ps pr	ta_700 zg_925 zg_500 ps pr	ta_500 zg_925 zg_500 ps pr
S4	ta_500 zg_925 ua_500 va_200 pr ps	ta_500 zg_500 ua_500 va_200 ps pr	ta_500 ua_500 va_200 pr ps	ta_500 ta_200 zg_200 ua_925 va_200
S5	zg_200 huss_925 ua_925 ps pr	zg_200 huss_925 ua_925 va_200 pr	zg_200 huss_850 ua_925 ps pr	zg_500 huss_850 va_500 va_200 ps
S6	ta_925 ta_500 zg_925 va_200 ps pr	ta_925 ta_500 zg_200 huss_925 ua_200 va_200 pr	ta_925 zg_925 ua_500 ua_200 va_200 pr	ta_925 ta_700 zg_200 huss_850 ua_925 va_500 va_200
S7	ta_200 zg_200 ua_925 va_200 pr	ta_200 zg_200 ua_925 va_925 pr	ta_200 zg_200 ua_925 va_200 pr	ta_500 zg_200 huss_850 ua_925 pr
S8	ta_700 zg_925 zg_500 ua_925 va_500 pr	ta_700 ta_500 zg_200 ua_925 va_925	zg_500 ua_925 ua_200 va_925 va_200 pr	ta_500 zg_200 huss_850 ua_925 va_200

mean rainfall at station S4 during the post-monsoon when precipitation is influenced by surface meridional synoptic systems. Meridional wind is also playing an important role in influencing precipitation at stations S6, S7, and S8, stations with high variation during the post-monsoon months. Temperature at different tropospheric levels is identified as the PCVs for most of the stations. The mid- and upper-tropospheric ridges and troughs are linked to the underlying tropospheric temperature, which serves to intensify developing systems through changes in the thickness advection (Cavazos and Hewitson 2005). Precipitable water is associated with the precipitation at all stations; however, the degree of association may vary depending on the climatic conditions. Thereby, the sets of causal variables are not exactly the same even within a specific climatic zone and these are distinctly different for the stations lying in different climatic zones. Said above, it should also be noted that the aforementioned justifications are not exhaustive. For a complete justification, separate extensive analysis is needed, which is beyond the scope of this study. This study focuses on the effectiveness of the proposed GM approach in identifying the PCVs at a location as compared with its counterparts.

The PCVs selected using the proposed GM approach are compared with the same obtained using the other existing approaches and presented in Table 4. It can be clearly observed that different approaches provide different set of PCVs; however, for a particular station, certain PCVs remain the same for all the approaches. For instance, specific humidity at 850 mb and zonal wind at 925 mb are selected as PCV using all the selection approaches at stations S1 and S8 respectively. Although some such cases may be observed, the overall set of the PCVs significantly changes with the approach and the station considered. Further, it may be observed that for most of the stations, certain common PCVs are selected using GM, PaCA, and SRA, although the entire set may be unique. For instance, at station S4, the variables, air temperature at 500 mb, zonal wind at 500 mb, meridional wind at 200 mb, and surface pressure are selected as the PCVs using all the three approaches. However, CA provides a completely different combination of PCVs, the reason being that CA provides misleading results in case the causal variables are strongly correlated amongst themselves. This is a very common occurrence as most of the coarse scaled atmospheric variables are dependent and this approach is unable to capture the true association among the large pool of associated variables and the target variable. Approaches like PaCA and SRA may provide results closer to GM as the former partials out the effect of other variables and the latter stepwise eliminates the effect of weakly associated variables. However, these approaches do not consider the complete interaction among all the variables, which may lead to redundancy in information. This may lead to a combination of PCVs that is insufficient for developing precise downscaling model. Thereby, proper identification of the PCVs has been increasingly identified as a major issue for the statistical downscaling model and a robust statistical technique needs to be utilized to resolve uncertainties associated with selection of causal variables. It is vital to determine the complete conditional independence structure among the large pool of associated variables and the target variable. The proposed

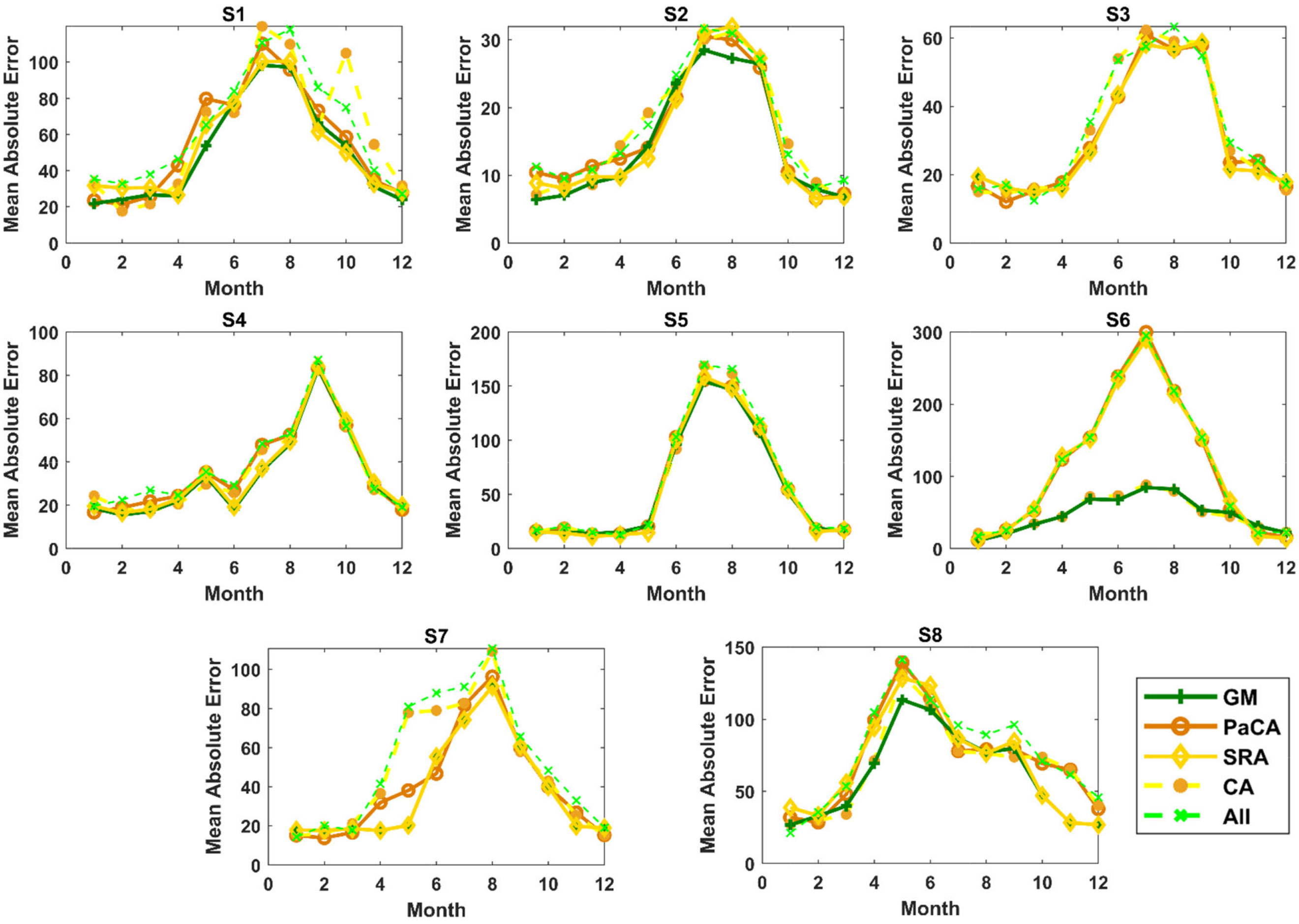

Fig. 2 Month-wise mean absolute errors (in mm) in SDSM downscaled precipitation obtained using PCVs identified by the proposed GM and existing approaches at different stations

GM approach facilitates identification of such structures and can be used effectively for identification of PCVs.

4.2 Relative performance of downscaled precipitation using different sets of PCVs

4.2.1 Performance in terms of month-wise mean

The absolute error in the estimates of mean downscaled precipitation obtained using the proposed approach is compared with the existing approaches (Figs. 2 and 3). In this and subsequent figures, legends are used as per the name of approaches used for selecting the PCVs though the downscaling model remains same, i.e., either SDSM or SVR, as mentioned before. It is noticed that the performance of the proposed GM approach is best or nearest to the best considering both the downscaling models. This is consistent for all the location with different climate regimes. It is further noticed that the performance of the existing methods varies from one location to another. In other words, one method may perform best at a location but exhibits lack of robustness at other locations. However, performance of GM is consistently good (best or near to the best as compared with its counterparts) at all the locations. Particularly, at station S6, GM and CA show distinguishably low errors as compared with the other two selection approaches considering both SDSM and SVR. Moreover, at stations S1, S2, S5, and S8, GM provides comparatively lower error as compared with the other three approaches. At stations S3, S4, and S7, the absolute errors using GM and other existing approaches are completely or closely coinciding for all the months of analysis. Further, downscaling of monthly precipitation is carried out using all the 17 causal variables and the results for the same are shown in Figs. 2 and 3. The results clearly show that the model performance diminishes on using all the causal variables as PCVs. Large numbers of causal variables which are either independent or conditionally independent of precipitation for a certain station/region increases the complexity of the model without providing additional information.

As stated earlier, the error values are tested by a non-parametric Wilcoxon rank-sum test at the 95% confidence level (Figs. 4 and 5). The results obtained using GM and the existing approaches are vastly varying with respect to the station and the month of analysis. At stations S1, S2,

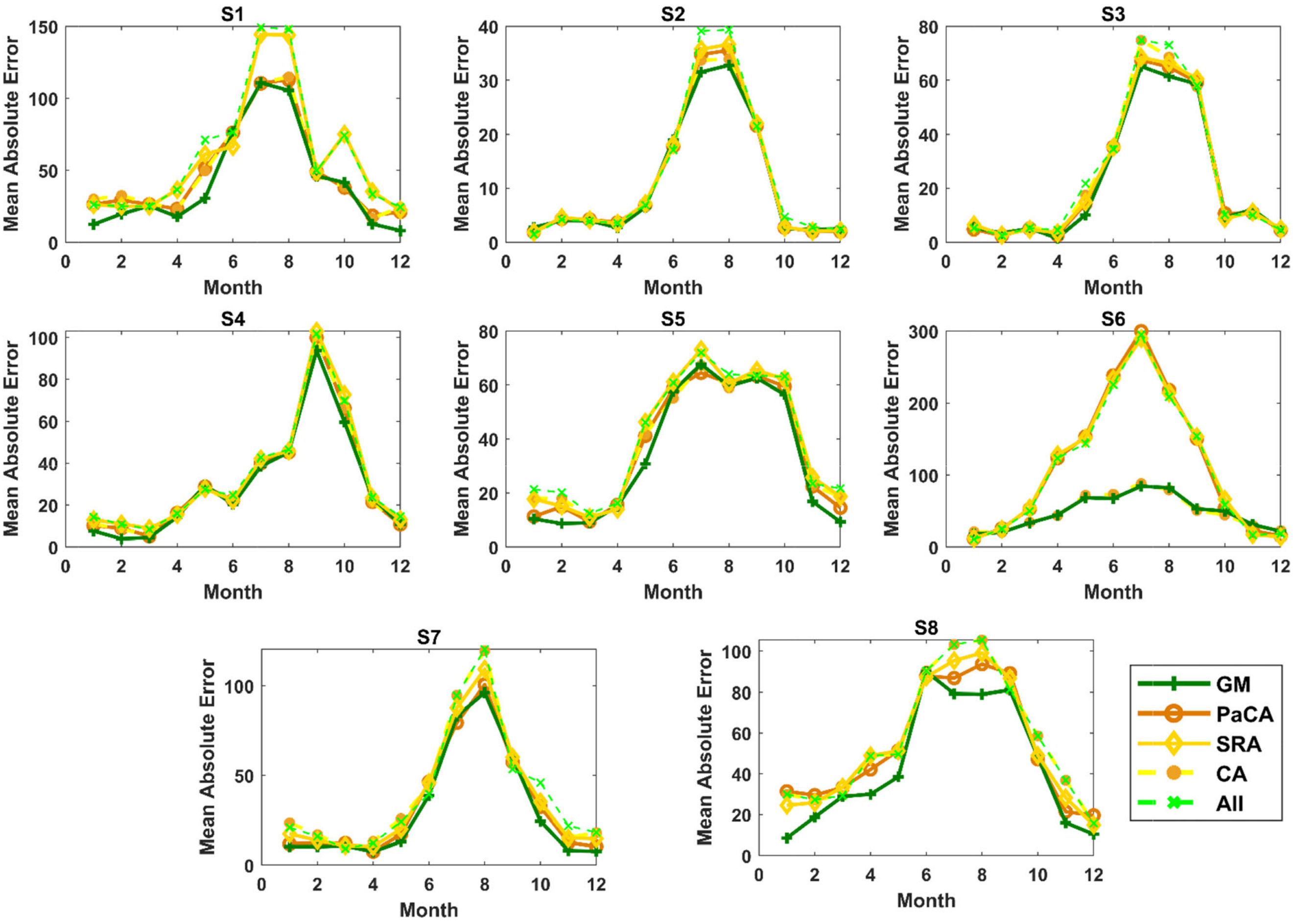

Fig. 3 Month-wise mean absolute errors (in mm) in SVR downscaled precipitation obtained using PCVs identified by the proposed GM and existing approaches at different stations

S3, and S4, considering 95% confidence level the errors are insignificant for all months (i.e., the *p* value is above 0.05) using GM. These results are closely followed by the same obtained using SRA for the abovementioned stations. However, PaCA also shows the second best performance at certain stations. For instance, at station S3 (Figs. 4 and 5), PaCA closely follows the results provided by GM for certain months, whereas SRA shows the second best results for the rest of the months. At stations S6 and S7, GM provides insignificant error for eleven months except the month of May, July and August. At stations S1 and S7, variation in results can be observed using the three existing approaches, as discussed earlier. For example, SRA provides the second best result for almost all the months at station S1, whereas for station S7, SRA and PaCA provide the closest results for some specific months. Similar to Figs. 2 and 3 at station S6, GM followed by CA provides insignificant error for most of the months, whereas the error is significant using PaCA and SRA. For the remaining stations, namely S5 and S8, the performance of GM is better as compared with the other three approaches. The precision of the downscaled results depends on the PCVs and in terms of error analysis of mean, it is clearly evident that different combination of causal variables leads to varying error in the downscaled results. However, it is vital to note the performance of GM is superior at each station considering the monthly means using both the downscaling approaches.

4.2.2 Performance in terms of variance and probability distribution

The variance of the observed and downscaled monthly precipitation is evaluated at each station considering all the months together. The variance of the downscaled monthly precipitation is evaluated individually for each ensemble obtained during model validation and the average is considered for comparison in case of SDSM. The results show that the variance in the observed precipitation is best captured in the downscaled precipitation while using PCVs identified by the proposed approach as compared with using the same obtained through other existing approaches across all the stations (Fig. 6). The second best performance varies spatially among the existing

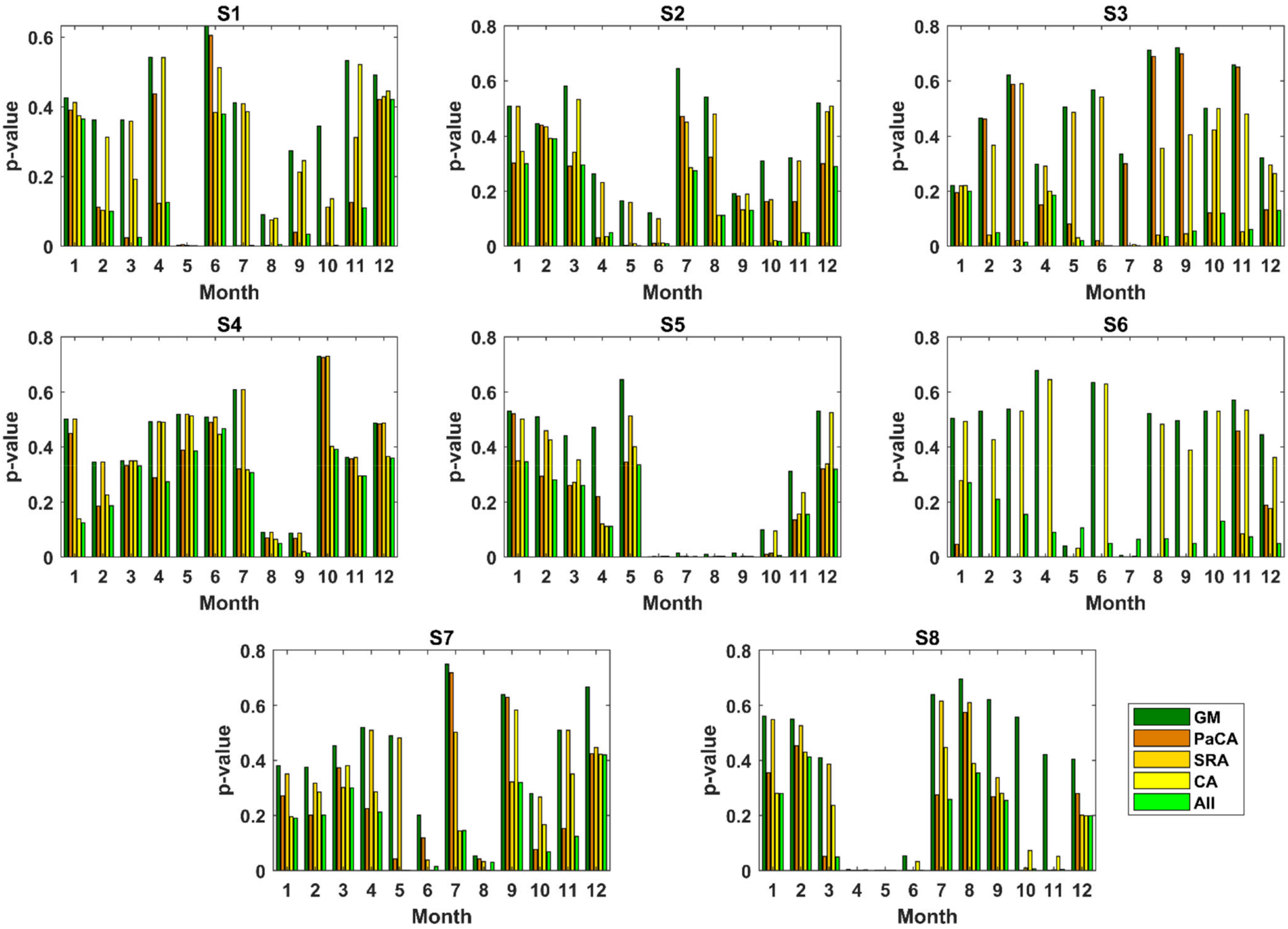

Fig. 4 Month-wise p values of Wilcoxon rank-sum test results for the assessment of difference in means between observed and SDSM downscaled precipitation obtained using PCVs identified by the proposed GM and existing approaches at different stations. The higher the p value, the better is the performance of the model for a particular month

approaches (PaCA, SRA, and CA) establishing the robust performance of the proposed approach.

Apart from the variance, the probability distributions of observed and downscaled monthly precipitation were also compared considering all the months together and KS test is used to evaluate whether the two distributions are essentially same at the 95% confidence level. The test results, shown in Fig. 7, indicate the p values obtained for the abovementioned test. In the case of the proposed GM approach for identifying PCVs, p value exceeds 0.05 for all the stations considering both the downscaling models. Thereby, the downscaled results perform well in reproducing the distribution of monthly precipitation at significance level of 0.05 using the PCVs identified through the proposed GM approach. The comparative results clearly depict that the downscaling models developed considering the PCVs identified through PaCA, SRA, and CA provide varying performance across the different stations. However, GM consistently provides superior results at all the locations with varying climate regime for all the months.

4.2.3 Overall comparison and recommendation

As the downscaling model (SDSM and SVM) remains the same, the PCVs identified using the different approaches are the only factors influencing the performance of the downscaled product. In brief, comparison of mean and variance and probability distribution of downscaled precipitation with that of the observed data establishes the efficacy of the GM approach to identify the PCVs as compared with the other three existing approaches. It is due to the fact that GM utilizes the complete conditional independence structure and eliminates the effect of independent or conditionally independent associated variables. Apart from many other factors, the precision of the downscaled results depends on the combinations of PCVs used as input, and it is clearly evident from the analysis that different combination of PCVs leads to varying performance and quality of the downscaled products. However, the performance of the proposed GM approach in identifying the PCVs is superior at each station

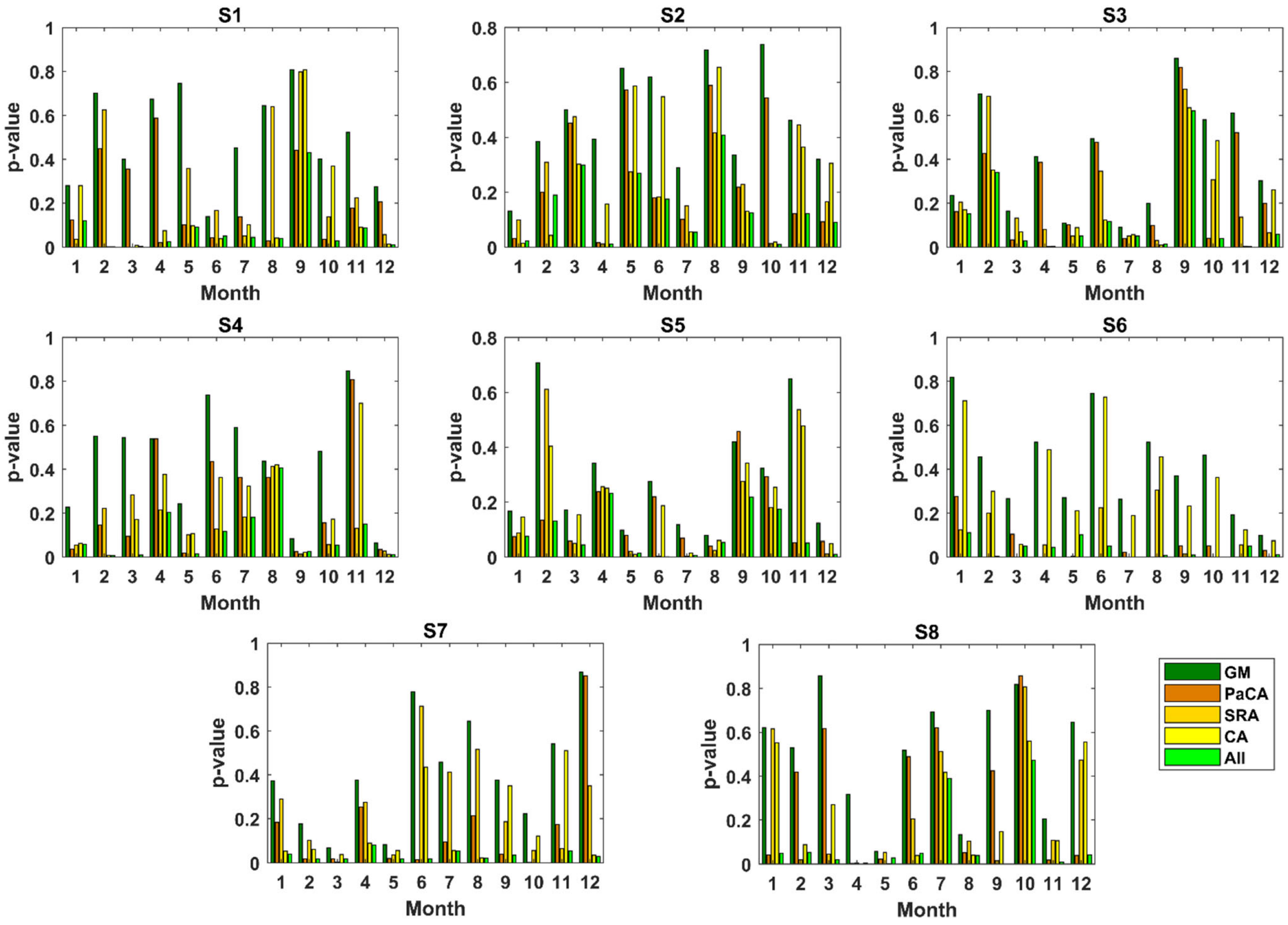

Fig. 5 Month-wise *p* values of Wilcoxon rank-sum test results for the assessment of difference in means between observed and SVR downscaled precipitation obtained using PCVs identified by the proposed GM and existing approaches at different stations. The higher the *p* value, the better is the performance of the model for a particular month

considering all three metrics mentioned before. Considering a significant spatio-temporal variation in hydroclimatic association, the proposed GM approach can be recommended as a promising approach to identify the PCVs for the downscaling models.

5 Conclusion

In this study, the proposed GM approach is established as an effective technique for selection of the PCVs of a statistical downscaling model. It facilitates the development of a

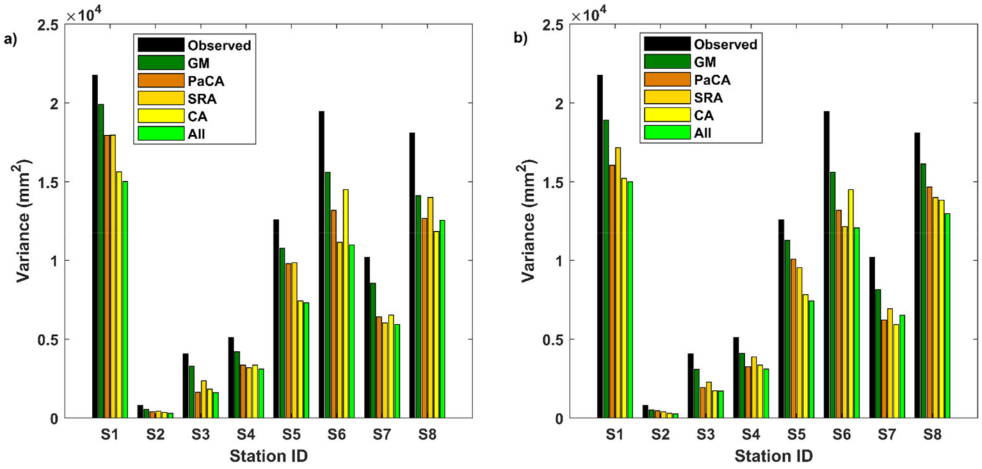

Fig. 6 Comparison of the variance between the observed and downscaled precipitation obtained using PCVs identified by the proposed GM and existing approaches at different stations through **a** SDSM and **b** SVR as the downscaling model

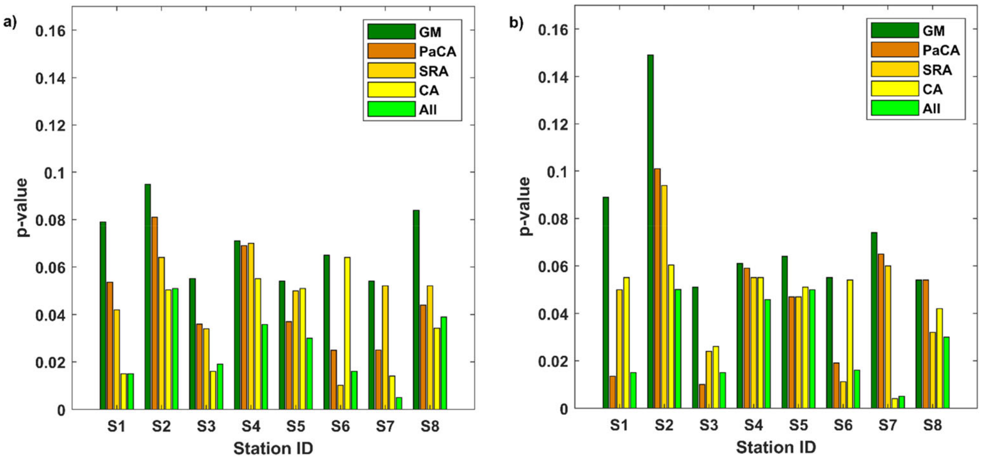

Fig. 7 *p* values (probability scale) of Kolmogorov-Smirnov goodness-of-fit test to compare the distributions of observed and downscaled precipitation obtained using PCVs identified by the proposed GM and existing approaches at different stations. Results are shown for **a** SDSM and **b** SVR as the downscaling model. The higher the *p* value, the better is the performance of the model for a particular month

conditional independence structure which helps to study the detailed association among the large pool of associated variables and target variable. The redundancy in information or possibility of missing out important variables due to complex nature of association is eliminated by considering the detailed conditional independence structure. Comparison of the observed monthly precipitation with the downscaled precipitation obtained using the proposed GM approach, in terms of mean, variance, and probability distribution, shows satisfactory performance for most of the stations.

The efficacy of the proposed approach in identifying the PCVs is established by comparing the downscaled results obtained using the PCVs identified by other existing approaches (CA, PaCA, and SRA). In the case of the existing approaches, the downscaled result shows inconsistent performance. Certain studies comparing the abovementioned existing approaches for different regions have depicted similar results (Yang et al. 2016, 2017). PCVs identified by a particular approach may provide better performance for certain months of analysis and a specific location; however, PCVs identified by no one approach provides consistent performance throughout all months and locations. However, the comparison of mean, variance, and probability distribution of the downscaled and observed data shows that the PCVs identified through the proposed GM approach helps to obtain consistent and robust results considering different seasons at each station located in varying climatic regions. The proposed GM approach provides a complete conditional independence structure, eliminating the information on independent and conditionally independent variables, which can be efficiently used for selection of the PCVs. Thereby, the proposed GM approach can be promising to identify the PCVs for downscaling models to tackle the significant spatio-temporal variation in hydroclimatic association. The error in the downscaled results can be reduced by using different downscaling methodologies. Further studies can be carried out by varying the large pool of associated variables initially considered, using downscaling models with higher efficiency and considering the time-varying association among the casual and target variables.

Funding The work was partially supported by the sponsored projects supported by Department of Science and Technology, Climate Change Programme (SPLICE), Government of India (Ref No. DST/CCP/CoE/79/2017(G)), through a sponsored project.

Appendix 1. Mathematical details of correlation analysis, partial correlation analysis, and stepwise regression analysis

Appendix 1.1. Correlation analysis

The correlation analysis (CA) is the most commonly used approach for selection of the PCVs. Strong correlation of the causal variables, from a pool of possibly associated hydroclimatic variables, with the target variable is the most basic criteria for selection of PCVs. In this approach, the selection is governed by the correlation coefficient between the associated variables and the target variable to be downscaled. A certain value of the correlation coefficient is considered the threshold value and all the associated variables having equal or higher correlation are considered the PCVs for downscaling. Pearson's correlation coefficient is used in this study and the same can be expressed as follows,

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}} \tag{5}$$

where r_{xy} is Pearson's correlation coefficient between the associated variables (X) and predictand (Y), n is the number of observations, x_i and y_i are the observations of X and Y respectively, and $\overline{x}$ and $\overline{y}$ are the means of X and Y respectively. The p value is evaluated, considering the correlation coefficient to follow t distribution at 95 % confidence level with $n-2$

degrees of freedom. The causal variables with p value greater than 0.05 are recommended to select as the PCVs of the statistical downscaling model.

Appendix 1.2. Partial correlation analysis

Partial correlation is the measure of association between two variables (a particular associated variable and target variable), while controlling the effect of other associated variables. The partial correlation analysis (PaCA) can be used to identify the PCVs for downscaling as it adjusts the effect of other associated variables. The partial correlation coefficient between two variables controlling the third variable can be expressed as follows,

$$r_{xy,z} = \frac{r_{xy}-r_{xz}r_{yz}}{\sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)}} \tag{6}$$

where $r_{xy,\ z}$ is the partial correlation between two variables X and Y when the third variable Z is controlled and r_{xy}, r_{xz}, r_{yz} is the correlation coefficient between X and Y, X and Z, and Y and Z respectively. The p value is evaluated, considering the partial correlation coefficient to follow t distribution at 95% confidence level with $n-3$ degrees of freedom. The causal variables with p value greater than 0.05 are recommended to select as the PCVs of the statistical downscaling model.

Appendix 1.3. Stepwise regression analysis

The stepwise regression analysis (SRA) is a method of fitting a regression model by stepwise removal of the least significant variables until all the remaining variables are significant. This method is often used for selection of PCVs when a large number of associated variables are available and to deal with issues related to multi-collinearity. In this technique, initially all the causal variables are considered in the model. At each step of the analysis, a variable is included or excluded from the model usually based on the partial F-tests. If F is greater than the critical F value, the causal variables can be included in the equation. The partial F statistic can be expressed as follows,

$$F = \frac{\left(R_q^2-R_{q-1}^2\right)(n-q-1)}{\left(1-R_q^2\right)} \tag{7}$$

where R is the correlation coefficient between a criteria variable and prediction equation, q is the number of causal variables in the equation, and n is as defined before. If the test statistic is less than the critical F value at 95% confidence level with degree of freedom $(n-q-1)$, the causal variables should be excluded from the equation. The causal variables with p value greater than 0.05 are recommended to select as the PCVs of the statistical downscaling model.

References

Anandhi A, Srinivas VV, Nanjundiah RS, Nagesh Kumar D (2008) Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. Int J Climatol 28:401–420. https://doi.org/10.1002/joc.1529

Bang-Jensen J, Gutin G (2007) Digraphs: theory, algorithms and applications. Softw Testing, Verif Reliab 12:59–60. https://doi.org/10.1002/stvr.240

Bates BC, Charles SP, Hughes JP (1998) Stochastic downscaling of numerical climate model simulations. Environ Model Softw 13:325–331. https://doi.org/10.1016/S1364-8152(98)00037-1

Beal MJ, Jojic N, Attias H (2003) A graphical model for audiovisual object tracking. IEEE Trans Pattern Anal Mach Intell 25:828–836. https://doi.org/10.1109/TPAMI.2003.1206512

Bergströms, Carlsson B, Gardelin M et al (2001) Climate change impacts on runoff in Sweden-assessments by global climate models, dynamical downscalling and hydrological modelling. Clim Res 16:101–112. https://doi.org/10.3354/cr016101

Beuchat X, Schaefli B, Soutter M, Mermoud A (2012) A robust framework for probabilistic precipitations downscaling from an ensemble of climate predictions applied to Switzerland. J Geophys Res Atmos 117:1–16. https://doi.org/10.1029/2011JD016449

Box GE, Cox DR (1964) An analysis of transformations. J R Stat Soc 26:211–252

Cavazos T, Hewitson BC (2005) Performance of NCEP-NCAR reanalysis variables in statistical downscaling of daily precipitation. Clim Res 28:95–107

Charles SP, Bates BC, Whetton PH, Hughes JP (1999) Validation of downscaling models for changed climate conditions: case study of southwestern Australia. Clim Res 12:1–14. https://doi.org/10.3354/cr012001

Chen ST, Yu PS, Tang YH (2010) Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. J Hydrol 385:13–22. https://doi.org/10.1016/j.jhydrol.2010.01.021

Chen J, Brissette FP, Leconte R (2011) Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. J Hydrol 401:190–202. https://doi.org/10.1016/j.jhydrol.2011.02.020

Chen H, Xu CY, Guo S (2012) Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. J Hydrol 434–435:36–45. https://doi.org/10.1016/j.jhydrol.2012.02.040

Chithra NR, Thampi SG (2017) Downscaling future projections of monthly precipitation in a catchment with varying physiography. ISH J Hydraul Eng 23:144–156. https://doi.org/10.1080/09715010.2016.1264895

Coulibaly P, Baldwin CK (2005) Nonstationary hydrological time series forecasting using nonlinear dynamic methods. J Hydrol 307:164–174. https://doi.org/10.1016/j.jhydrol.2004.10.008

Dettinger MD, Cayan DR, Meyer MK, Jeton A (2004) Simulated hydrologic responses to climate variations and change in the Merced, Carson, and American River basins, Sierra Nevada, California, 1900-2099 *. Clim Change 62:283–317. https://doi.org/10.1023/B:CLIM.0000013683.13346.4f

Devak M, Dhanya CT (2014) Downscaling of precipitation in Mahanadi Basin, India. Int J Civ Eng Res 5:111–120

Dutta R, Maity R (2018) Temporal evolution of hydroclimatic teleconnection and a time-varying model for long-lead prediction of Indian summer monsoon rainfall. Sci Rep 8:10778. https://doi.org/10.1038/s41598-018-28972-z

Dutta R, Maity R (2020a) Spatial variation in long-lead predictability of summer monsoon rainfall using a time-varying model and global climatic indices. Int J Climatol. https://doi.org/10.1002/joc.6556

Dutta R, Maity R (2020b) Temporal networks-based approach for non-stationary hydroclimatic modeling and its demonstration with streamflow prediction. Water Resour Res 56:e2020WR027086. https://doi.org/10.1029/2020WR027086

Fowler HJ, Blenkinsop S, Tebaldi C (2007) Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int J Climatol 27:1547–1578. https://doi.org/10.1002/joc.1556

Grimes DIF, Coppola E, Verdecchia M, Visconti G (2003) A neural network approach to real-time rainfall estimation for Africa using satellite data. J Hydrometeorol 4:1119–1133. https://doi.org/10.1175/1525-7541(2003)004<1119:ANNATR>2.0.CO;2

Gutmann E, Pruitt T, Clark M (2014) An intercomparison of statistical downscaling methods used for water resource assessments in the United States. Water Resour Res:1–20. https://doi.org/10.1002/2014WR015559.Received

Harpham C, Wilby RL (2005) Multi-site downscaling of heavy daily precipitation occurrence and amounts. J Hydrol 312:235–255. https://doi.org/10.1016/j.jhydrol.2005.02.020

Hassan Z, Shamsudin S, Harun S (2014) Application of SDSM and LARS-WG for simulating and downscaling of rainfall and temperature. Theor Appl Climatol 116:243–257. https://doi.org/10.1007/s00704-013-0951-8

Haylock MR, Cawley GC, Harpham C, Wilby RL, Goodess CM (2006) Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. Int J Climatol 26:1397–1415. https://doi.org/10.1002/joc.1318

Hessami M, Gachon P, Ouarda TBMJ, St-Hilaire A (2008) Automated regression-based statistical downscaling tool. Environ Model Softw 23:813–834. https://doi.org/10.1016/j.envsoft.2007.10.004

Huth R (1999) Statistical downscaling in central Europe: evaluation of methods and potential predictors. Clim Res 13:91–101. https://doi.org/10.3354/cr013091

Ihler AT, Kirshner S, Ghil M, Robertson AW, Smyth P (2007) Graphical models for statistical inference and data assimilation. Phys D Nonlinear Phenom 230:72–87. https://doi.org/10.1016/j.physd.2006.08.023

Johnson AR, Bhattacharya KG (2009) Statistics: principles and methods, sixth. John Wiley & Sons, Inc., United States of America

Jordan MI (2004) Graphical Models. Stat Sci 19:140–155. https://doi.org/10.1214/088342304000000026

Kidson JW, Thompson CS (1998) A comparison of statistical and model-based downscaling techniques for estimating local climate variations. J Clim 11:735–753. https://doi.org/10.1175/1520-0442(1998)011<0735:ACOSAM>2.0.CO;2

Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst Biol 5:21. https://doi.org/10.1186/1752-0509-5-21

Lauritzen SL, Sheehan NA (2003) Graphical models for genetic analyses. Stat Sci 18:489–514. https://doi.org/10.1214/ss/1081443232

Liu Z, Xu Z, Charles SP, Fu G, Liu L (2011) Evaluation of two statistical downscaling models for daily precipitation over an arid basin in China. Int J Climatol 31:2006–2020. https://doi.org/10.1002/joc.2211

Maity R (2018) Statistical methods in hydrology and hydroclimatology. Springer Nature, Singapore

Meenu R, Rehana S, Mujumdar PP (2013) Assessment of hydrologic impacts of climate change in Tunga-Bhadra river basin, India with HEC-HMS and SDSM. Hydrol Process 27:1572–1589. https://doi.org/10.1002/hyp.9220

Okkan U, Inan G (2015) Statistical downscaling of monthly reservoir inflows for Kemer watershed in Turkey: use of machine learning methods, multiple GCMs and emission scenarios. Int J Climatol 35:3274–3295. https://doi.org/10.1002/joc.4206

Pervez MS, Henebry GM (2014) Projections of the Ganges-Brahmaputra precipitation-downscaled from GCM predictors. J Hydrol 517:120–134. https://doi.org/10.1016/j.jhydrol.2014.05.016

Pichuka S, Maity R (2016) Spatio-temporal downscaling of projected precipitation in the 21st century: indication of a wetter monsoon over the Upper Mahanadi Basin, India. Hydrol Sci J 62:1–16. https://doi.org/10.1080/02626667.2016.1241882

Pierce DW, Cayan DR, Thrasher BL (2014) Statistical downscaling using localized constructed analogs (LOCA)*. J Hydrometeorol 15:2558–2585. https://doi.org/10.1175/JHM-D-14-0082.1

Pinto JG, Neuhaus CP, Leckebusch GC, Reyers M, Kerschgens M (2010) Estimation of wind storm impacts over Western Germany under future climate conditions using a statistical-dynamical downscaling approach. Tellus, Ser A Dyn Meteorol Oceanogr 62:188–201. https://doi.org/10.1111/j.1600-0870.2009.00424.x

Radchenko P, James GM (2010) Variable selection using adaptive nonlinear interaction structures in high dimensions. J Am Stat Assoc 105:1541–1553. https://doi.org/10.1198/jasa.2010.tm10130

Schoof JT, Shin DW, Cocke S, LaRow TE, Lim YK, O'Brien JJ (2009) Dynamically and statistically downscaled seasonal temperature and precipitation hindcast ensembles for the southeastern USA. Int J Climatol 29:243–257. https://doi.org/10.1002/joc.1717

Semenov MA, Brooks RJ, Barrow EM, Richardson CW (1998) Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. Clim Res 10:95–107. https://doi.org/10.3354/cr010095

Stoner AMK, Hayhoe K, Yang X, Wuebbles DJ (2013) An asynchronous regional regression model for statistical downscaling of daily climate variables. Int J Climatol 33:2473–2494. https://doi.org/10.1002/joc.3603

Taeb A, Reager JT, Turmon M, Chandrasekaran V (2017) A statistical graphical model of the California Reservoir System. Water Resour Res. 53:9721–9739. https://doi.org/10.1002/2017WR020412

Tatli H, Dalfes HN, Menteş ŞS (2004) A statistical downscaling method for monthly total precipitation over Turkey. Int J Climatol 24:161–180. https://doi.org/10.1002/joc.997

Tatsumi K, Oizumi T, Yamashiki Y (2015) Effects of climate change on daily minimum and maximum temperatures and cloudiness in the Shikoku region: a statistical downscaling model approach. Theor Appl Climatol 120:87–98. https://doi.org/10.1007/s00704-014-1152-9

Tomozeiu R, Cacciamani C, Pavan V, Morgillo A, Busuioc A (2007) Climate change scenarios for surface temperature in Emilia-Romagna (Italy) obtained using statistical downscaling models. Theor Appl Climatol 90:25–47. https://doi.org/10.1007/s00704-006-0275-z

Webster PJ, Magaña VO, Palmer TN, Shukla J, Tomas RA, Yanai M, Yasunari T (1998) Monsoons: processes, predictability, and the prospects for prediction. J Geophys Res Ocean 103:14451–14510. https://doi.org/10.1029/97JC02719

Whittaker J (2009) Graphical models in applied multivariate statistics. Wiley Publishing

Wilby RL, Hay LE, Leavesly HH (1999) A comparison of downscaled and raw output: implications for climate change scenarios in the San Juan river basin, Colorado. J Hydrol 225:67–91. https://doi.org/10.1016/S0022-1694(99)00136-5

Wilby R, Dawson C, Barrow E (2002) Sdsm—a decision support tool for the assessment of regional climate change impacts. Environ

Model Softw 17:145–157. https://doi.org/10.1016/S1364-8152(01)00060-3

Wood AW, Leung LR, Sridhar V, Lettenmaier DP (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Clim Change 62:189–216. https://doi.org/10.1023/B:CLIM.0000013685.99609.9e

Yang J, Li L, Wang A (2011) A partial correlation-based Bayesian network structure learning algorithm under linear SEM. Knowledge-Based Syst 24:963–976. https://doi.org/10.1016/j.knosys.2011.04.005

Yang C, Wang N, Wang S, Zhou L (2016) Performance comparison of three predictor selection methods for statistical downscaling of daily precipitation. Theor Appl Climatol 131:43–54. https://doi.org/10.1007/s00704-016-1956-x

Yang C, Wang N, Wang S (2017) A comparison of three predictor selection methods for statistical downscaling. Int J Climatol 37:1238–1249. https://doi.org/10.1002/joc.4772

Zuo D, Xu Z, Zhao J, Abbaspour KC, Yang H (2015) Response of runoff to climate change in the Wei River basin, China. Hydrol Sci J 60: 508–522. https://doi.org/10.1080/02626667.2014.943668

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.