Downscaling of precipitation on a lake basin: evaluation of rule and decision tree induction algorithms

Manish Kumar Goyal and C. S. P. Ojha

ABSTRACT

We investigate the performance of existing state-of-the-art rule induction and tree algorithms, namely Single Conjunctive Rule Learner, Decision Table, M5 Model Tree, Decision Stump and REPTree. Downscaling models are developed using these algorithms to obtain projections of mean monthly precipitation to lake-basin scale in an arid region in India. The effectiveness of these algorithms is evaluated through application to downscale the predictand for the Lake Pichola region in Rajasthan state in India, which is considered to be a climatically sensitive region. The predictor variables are extracted from (1) the National Centre for Environmental Prediction (NCEP) reanalysis dataset for the period 1948–2000 and (2) the simulations from the third-generation Canadian Coupled Global Climate Model (CGCM3) for emission scenarios A1B, A2, B1 and COMMIT for the period 2001–2100. M5 Model Tree algorithm was found to yield better performance among all other learning techniques explored in the present study. The precipitation is projected to increase in future for A2 and A1B scenarios, whereas it is least for B1 and COMMIT scenarios using predictors.

Key words | downscaling, IPCC SRES, M5, machine learning algorithms, precipitation

Manish Kumar Goyal (corresponding author)
Department of Civil and Environmental Engineering,
University of Waterloo,
Waterloo N2L3G1,
Canada
and
Department of Civil and Environmental Engineering,
Nanyang Technological University,
Singapore 639798,
Singapore
E-mail: *mkgoyal@uwaterloo.ca*; *mkgoyal@tu.edu.sg*

Manish Kumar Goyal
C. S. P. Ojha
Department of Civil Engineering,
Indian Institute of Technology,
Roorkee 247667,
India

INTRODUCTION

General circulation models (GCMs) are an important tool in the assessment of climate change and have been demonstrated to reproduce observed features of recent climate and past climate changes. This climate change information is available on a global/spatial scale as the output of GCMs at very coarse resolution (Prudhomme *et al.* 2003). However, they remain relatively coarse in resolution and are unable to resolve significant sub-grid scale features such as topography, clouds and land use (Grotch & MacCracken 1991). There is therefore a need to convert the GCM outputs into hydrologic variables (e.g. precipitation) at a local scale. The methods used to convert GCM outputs into local meteorological variables required for reliable hydrological modelling are usually referred to as 'downscaling' techniques. In the past couple of decades several downscaling approaches have been proposed in order to deal with the problem of spatial scale mismatch, and these approaches have been widely used in their cradles such as USA and Europe (Wilks 1989; Murphy 2000; Hellström *et al.* 2001; Hanssen-Bauer *et al.* 2005; Fowler & Wilby 2007; Vrac *et al.* 2007; Wetterhall *et al.* 2007; Chu *et al.* 2010). According to the latest IPCC assessment: 'Research on SDM [statistical downscaling models] has shown an extensive growth in application, and includes an increased availability of generic tools for the impact community' (Christensen *et al.* 2007).

Two fundamental approaches exist for the downscaling of large-scale GCM output to a finer spatial resolution. The first of these is a dynamical approach where a higher resolution climate model is embedded within a GCM. The second approach is to use statistical methods to establish empirical relationships between GCM-resolution climate variables and local climate. Dynamic downscaling approach has superior capability in complex terrain or with changed land cover (Kite 1997; Wang *et al.* 2004). However, this method entails higher computation cost and relies strongly on the boundary conditions provided by GCMs (Chu *et al.* 2010). In contrast, statistical downscaling gains local

doi: 10.2166/nh.2012.040

predictands by appropriate statistical or empirical relationships with surface or troposphere atmospheric features predictors (Wilby & Wigley 1997; Xu 1999; Fowler *et al.* 2007). As this method is comparatively cheap and computationally efficient, and is as powerful as its dynamic competitor, it has been widely employed in climate change impact assessments. However, its drawback is that it needs a reliable observed historical data series for calibration and builds the appropriate statistical relationship (Chu *et al.* 2010).

Statistical downscaling methods are generally classified into three groups: (i) regression models; (ii) weather typing schemes; and (iii) weather generators (WGs). Each group covers a range of methods, all relying on the fundamental concept that regional climates are largely a function of the large-scale atmospheric state (Fowler *et al.* 2007). This relationship may be expressed as a stochastic and/or deterministic function between large-scale atmospheric variables (predictors) and local climate variables (predictands).

In regression models, the transfer function is used to describe methods that directly quantify a relationship between the predictand and a set of predictor variables (Giorgi & Hewitson 2001; Kang *et al.* 2007).

Weather typing or classification schemes relate the occurrence of particular 'weather classes' to local climate. Weather classes may be defined synoptically, typically using empirical orthogonal functions from pressure data (Goodess & Palutikof 1998), by indices from sea-level pressure (SLP) data (Conway *et al.* 1996) or by applying cluster analysis (Fowler *et al.* 2000) or fuzzy rules (Bardossy *et al.* 2005) to atmospheric pressure fields (Fowler *et al.* 2007).

Weather generators in the simplest form are stochastic models, based on daily precipitation with a two state first-order Markov chain dependent on transition probabilities for simulating precipitation occurrence, and a gamma distribution for precipitation amounts (Wilks 1992), although second-order (Mason 2004) and third-order (Dubrovsky *et al.* 2004) Markov chain models have now been developed that are better able to reproduce precipitation occurrence or persistence (Fowler *et al.* 2007).

Several studies (Fowler *et al.* 2007; Anandhi *et al.* 2008; Vimont *et al.* 2010) have shown that the statistical downscaling method is simple to handle and generally has superior capability; it is therefore widely applied (Wilby & Harris 2006). A statistical downscaling technique that produces future scenarios based on statistical relationships between larger-scale climate features and hydrologic variables (such as precipitation) is therefore considered here.

Precipitation is an important parameter for climate change impact studies. A proper assessment of probable future precipitation and its variability is to be made for various water resources planning, management and hydro-climatology scenarios. Downscaling of precipitation has found wide application in hydro-climatology for scenario construction and simulation/prediction of: (i) daily and monthly precipitation (Wetterhall *et al.* 2005); (ii) monthly precipitation (Tripathi *et al.* 2006); (iii) monthly precipitation (Benestad *et al.* 2007); (iv) seasonal precipitation (Chu *et al.* 2008); (v) monthly precipitation and temperatures (Anandhi *et al.* 2008, 2009); and (vi) annual precipitation (Vimont *et al.* 2010).

In India, statistical downscaling methods have gradually started to receive increasing attention (Ghosh & Mujumdar 2006, 2008; Anandhi *et al.* 2008, 2009). Ghosh & Mujumdar (2006) presented a methodology for the prediction of future monthly rainfall scenarios in a meteorological subdivision in Orrisa using the fuzzy clustering approach on the GCM outputs. Mean sea-level pressure and geopotential height were used as predictors. Principal component analysis (PCA) is used to reduce the dimensionality of the dataset. Fuzzy clustering technique is applied to classify the principal components identified by the PCA and the fuzzy membership values are used in the model, with an assumption that the effects of circulation patterns on precipitation in different clusters are different. The developed model was observed to produce a good prediction of rainfall with a high goodness-of-fit (R^2) value.

Ghosh & Mujumdar (2008) investigated a methodology of statistical downscaling based on sparse Bayesian learning and Relevance Vector Machine (RVM) to model streamflow at river basin scale on the Mahannadi River in Orrisa. This study considered 2 m surface air temperature, mean sea-level pressure (MSLP), 500 hPa geopotential height and surface specific humidity as the predictors for modelling Mahanadi streamflow in monsoon season. A decreasing trend is observed for monsoon streamflow of Mahanadi due to high surface warming in future, with the Center for

Climate System Research (CCSR), Tokyo and National Institute for Environmental Studies (NIES) GCM and B2 scenario. It is also concluded in this study that, due to an increase in temperature, the water yield in the river is adversely affected.

Anandhi *et al.* (2008, 2009) and carried out downscaling of maximum and minimum temperatures as well as monthly precipitation to river basin scale using the support vector machine (SVM) method. For downscaling precipitation, predictors were chosen based on a physical meaningful relationship with precipitation. Seasonal stratification was performed to facilitate the development of a separate downscaling model to capture the relationship between predictors and predictand for each season. For downscaling temperature, the predictor variables are classified into three groups. The performance of the SVM models that are developed (one for each combination of predictor group, predictand, calibration period and location-based stratification i.e. land, land and ocean of climate variables) was evaluated. The results of the validation indicate that the SVM model is a feasible choice for downscaling the predictands.

In this paper, we explore existing state-of-the-art rule induction and tree algorithms; namely: (a) Single Conjunctive Rule Learner, (b) Decision Table, (c) M5 Model Tree, (d) Decision Stump, and (e) REPTree as a downscaling methodology to study climate change impact over the Lake Pichola basin in an arid region. A number of downscaling methods such as regression (Kilsby *et al.* 1998), canonical correlation analysis (Heyen *et al.* 1996; Xoplaki *et al.* 2000), *K*-nearest neighbour (Gangopadhyay *et al.* 2005; Goyal *et al.* 2011), artificial neural networks (Hewitson & Crane 1994; Gardner & Dorling 1998; Cannon & Lord 2000; Schoof & Pryor 2001; Ojha *et al.* 2010) and SVMs (Anandhi *et al.* 2008) have been used in the past in several regions across the globe.

In recent decades, machine learning and data mining research has developed rapidly and, as one of the most successful branches of Artificial Intelligence, is playing a more and more important role in real-world applications. Rule induction and decision tree algorithms are alternative approaches that are quite transparent and do not need optimization of network geometry and internal parameters. These methods have been applied in rainfall–runoff modelling (Solomatine & Dulal 2003), flood forecasting (Solomatine & Yunpeng 2004), modelling water level discharge relationship (Bhattacharya & Solomatine 2005) and sediment transport (Bhattacharya *et al.* 2007).

There is no apparent evidence of any study in the literature dealing with simultaneous evaluation of various decision learning approaches for downscaling precipitation. In the light of this, the objective of this study is: (i) to rank various rule induction and tree algorithms; and (ii) to downscale mean monthly precipitation using the best available approach from simulations of CGCM3 for the latest Intergovernmental Panel on Climate Change (IPCC) scenarios. The scenarios which are studied in this paper are relevant to the IPCC's fourth assessment report (AR4) released in 2007.

Study region

The area of this study is the Lake Pichola catchment in Rajasthan state in India, located at 72.5–77.5° E, 22.5–27.5° N and 587 m a.m.s.l. It receives an average annual precipitation of 597 mm based on last decade (1990–2000). It has a tropical monsoon climate where most of the precipitation is confined to a few months of the monsoon season; this indicates heavy dependence on the monsoon rainfall.

In Rajasthan state, rainfall in this region occurs mainly during June–September through the monsoon wind; non-monsoon rainfall is limited and irregular. The state is characterized by a non-nucleated, dispersed pattern of settlement, with diverse physiography ranging from desert and semi-arid regions of western Rajasthan to the greener belt east of the Aravallis, and the hilly tribal tracts in the south-east (Goel & Singh 2006).

The southwest (summer) monsoon has warm winds blowing from the Indian Ocean, causing copious amount of precipitation during June–September months and may act as a potential causative factor for the spatial and temporal variability of precipitation in the region. This monsoon, which has its beginning in the last week of June, may last until mid-September. Pre-monsoon showers begin towards the middle of June and post-monsoon rains occasionally occur in October. In the winter season, there is also sometimes a little rainfall associated with the passing western distribution over the region. (Note that a western distribution is an eastwards-moving extra-tropical upper air

trough in the subtropical westerlies, often extending down to the lower atmospheric level of the north Indian latitude during the winter months; Das *et al.* 2010.)

Of the various water bodies of Udaipur region, Lake Pichola is the oldest and biggest water body. This lake was created by a Banjara Chief during the time of Maharana Lakhaji (1382–1418 AD) by building a small earthen dam across the river Sisrama. The lake has an irregular and sub-triangular shape. The water spread area of the lake varies considerably from season to season and also annually. Initially the water of the lake was used for irrigation purposes. However, as the population grew, the lake became the major source of drinking water supply to Udaipur city and irrigation was stopped. It has been reported that, until recently, about 80% of the total water supply to Udaipur city is provided by Lake Pichola alone. Recently however, the water availability in the lake is unable to meet the demand and water is being brought to Udaipur from Lake Jaisamand (located about 40 km from Udaipur). Besides drinking water supply, the lake has many other uses. A dense population residing along the northern shores of the lake uses the lake water directly for bathing, washing of clothes and socio-cultural activities. However, the most important aspect of Lake Pichola is tourism, upon which the economy of the region is dependent. The lake ecosystem once used to support a wide variety of fauna and flora, most of which has disappeared now (Khobragade 2009).

During the past several decades, the streamflow regime in the lake catchment has changed considerably resulting in water scarcity, low agriculture yield and degradation of the ecosystem in the study area. Regions with arid and semi-arid climates could be sensitive to even insignificant changes in climatic characteristics (Linz *et al.* 1990). Temperature affects the evapotranspiration (Jessie *et al.* 1996), evaporation and desertification processes and is also considered as an indicator of environmental degradation and climate change. Understanding the relationships between the hydrologic regime, climate factors and anthropogenic effects is important for the sustainable management of water resources in the entire catchment; this study area was therefore chosen because of the above-mentioned reasons. Figure 1 depicts the location map of the study region. The observed monthly precipitation is depicted in Figure 2 for the study period (1975–2000).

DATA EXTRACTION

The monthly mean atmospheric variables were derived from the National Center for Environmental Prediction (NCEP) reanalysis dataset (Kalnay *et al.* 1996) for the period of January 1975 to December 2000. The data have a horizontal resolution of 2.5° latitude × 2.5° longitude and 17 constant pressure levels in the vertical. The atmospheric variables

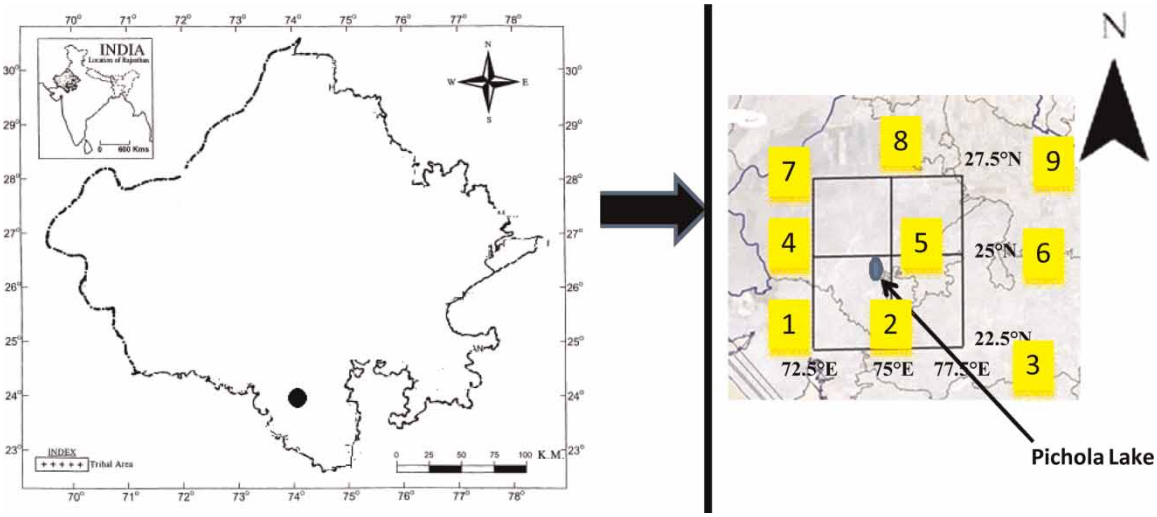

Figure 1 | Location map of the study region in Rajasthan state of India with NCEP grid.

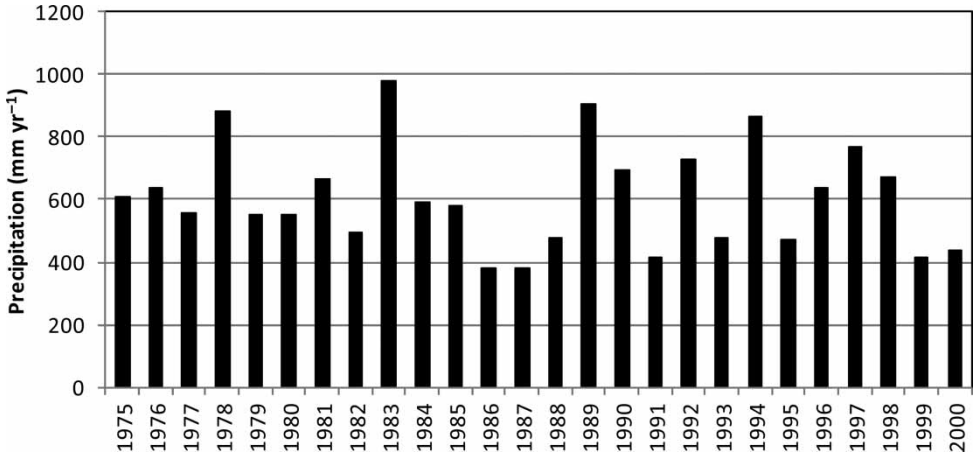

Figure 2 | Observed precipitation for the study region for the period 1975–2000.

are extracted for nine grid points whose latitude ranges from 22.5 to 27.5° N and longitude ranges from 72.5 to 77.5° E at a spatial resolution of 2.5°. The meteorological data (i.e. precipitation) at a monthly timescale was obtained from records available for Lake Pichola, located in Udaipur at 24°34′ N latitude and 73°40′ E longitude. The data are available for the period January 1975 to December 2000 (Khobragade 2009).

The Canadian Center for Climate Modeling and Analysis (CCCma) (http://www.cccma.bc.ec.gc.ca) provides GCM data for a number of surface and atmospheric variables for the CGCM3 T47 version which has a horizontal resolution of roughly 3.75° latitude × 3.75° longitude and a vertical resolution of 31 levels. CGCM3 is the third version of the CCCMA Coupled Global Climate Model which makes use of a significantly updated atmospheric component AGCM3 and uses the same ocean component as in CGCM2. The data comprise of present-day (20C3M, i.e. climate of the 20th century) and future simulations forced by four emission scenarios, namely A1B, A2, B1 and COMMIT. The climate data are extracted at a monthly timescale for the period from January 1975 to December 2100. Data used in this study were obtained for CGCM3 climate of the 20th Century (20CM3) experiments.

The nine grid points surrounding the study region, as shown in Figure 1, are selected as the spatial domain of the predictors to adequately cover the various circulation domains of the predictors considered in this study. The GCM data is re-gridded to a common 2.5° using inverse square interpolation technique (Willmott *et al.* 1985). The utility of this interpolation algorithm was examined in previous downscaling studies (Shannon & Hewitson 1996; Crane & Hewitson 1998; Tripathi *et al.* 2006; Ghosh & Mujumdar 2008; Goyal & Ojha 2011a). The development of downscaling models for all the months for the predictand variable precipitation begins with selection of potential predictors, followed by application of rule induction and tree algorithms downscaling model. The developed model is then used to obtain projections of precipitation from simulations of CGCM3.

Introduction to rule induction and tree algorithms, different error norms and selection of predictors

Rule induction and tree algorithms

Rule induction and tree algorithms are among the most widely used machine learning algorithms. They perform a general to specific search of a feature space, adding the most informative features to a rule/tree structure as the search proceeds by generating a set of rules and by constructing a decision tree where each internal node is a feature. The extraction of important information from a large pile of data and its correlations is often the advantage of using machine learning (Nilsson 1965). These algorithms provide a common method of organizing data to make predictions and to provide a basis for model construction, second only to regression. The objective is to select a

minimal set of features that efficiently partitions the feature space into classes of observations and assemble them into a rule/tree (Goyal & Ojha 2011b).

Single Conjunctive Rule Learner

Single Conjunctive Rule Learner is one of the data-mining learning algorithms and is normally known as inductive learning. The goal of rule induction is generally to induce a set of rules from data that captures all generalizable knowledge within that data, while being as small as possible (Cohen 1995). Classification in rule-induction classifiers is typically based on the firing of a rule on a test instance, triggered by matching feature values at the left-hand side of the rule (Clark & Niblett 1989). Rules can be of various normal forms and are typically ordered; with ordered rules, the first rule that fires determines the classification outcome and halts the classification process (Othman & Yau 2007). For classification, the information of one antecedent is the weighted average of the entropies of both the data covered and not covered by the rule. For regression, the information is the weighted average of the mean-squared errors of both the data covered and not covered by the rule. In pruning, weighted average of the accuracy rates on the pruning data is used for classification while the weighted average of the mean-squared errors on the pruning data is used for regression (Ibrahim *et al.* 2006).

Decision Table

Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table. This is done using a best-first search. The Decision Table algorithm builds a decision rule using a simple decision table majority classifier, as proposed by Kohavi (1995). It summarizes the dataset with a 'decision table' which contains the same number of attributes as the original dataset. A new data item is then assigned a category by finding the line in the decision table that matches the non-class values of the data item. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table (Kohavi 1995).

M5 Model Tree

The Model Tree technique provides a structural representation of the data and a piecewise linear fit of the class (Quinlan 1992). These have a conventional decision tree structure but use linear functions at the leaves instead of discrete class labels, as shown in Figure 3. Like conventional decision tree learners, M5 builds a tree by splitting the data based on the values of predictive attributes. Instead of selecting attributes by an information theoretic metric, M5 chooses attributes that minimize intra-subset variation in the class values of instances that go down each branch. The variability is measured by the standard deviation of the values that reach that node from the root through the branch by calculating the expected reduction in error as a result of testing each attribute at that node. The attribute which maximizes the expected error reduction is chosen. The splitting stops if the values of all instances that reach a node vary slightly or only a few instances remain. The standard deviation reduction (SDR) is calculated as:

$$\mathrm{SDR} = \mathrm{sd}(T) - \sum_i \frac{|T_i|}{|T|} \times \mathrm{sd}(T_i) \tag{1}$$

where T represents a set of examples that reach the node; T_i denotes the sets of examples that have the ith outcome of the potential set; and sd represents the standard deviation (Wang & Witten 1997). After the tree has been grown, M5 computes a linear multiple regression model for every interior node. The data associated with that node and only the attributes tested in the sub-tree rooted at that node are used in the regression. The attributes will

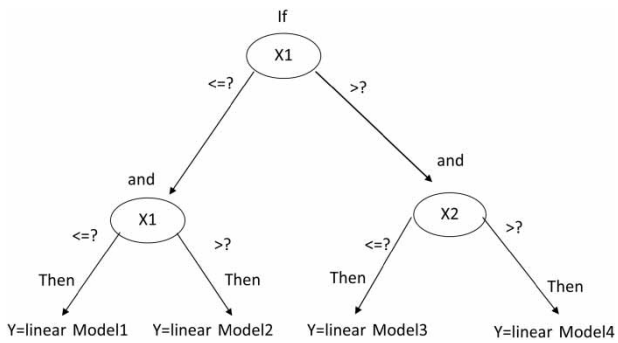

Figure 3 | The M5 Model Tree.

be dropped one by one if they lower the estimated error. The tree is then pruned of the leaves that result in a lower expected estimated error. In the final stage, a smoothing process is performed to compensate for the sharp discontinuities that will inevitably occur between adjacent linear models at the leaves of the pruned tree, particularly for some models constructed from a smaller number of training examples (Shahidi & Mahjoobi 2009).

Decision Stump

Decision stumps are simple classifiers where the final decision is made by only a single hypothesis or feature. A decision stump is simply a one-node decision tree based on a co-occurrence feature, while the majority classifier assigns the most frequent sense in the training data to every occurrence of that word in the test data. The Decision Stump algorithm builds simple binary decision 'stumps' for both numeric and nominal classification problems (Witten *et al.* 1999). This is a decision tree with a single node that determines how to classify inputs based on a single hypothesis. It contains one leaf for each possible hypothesis value, specifying class levels that should be assigned to inputs whose features have that value. In order to build that value, the hypothesis to be used must be chosen carefully. The easiest method is simply to build a decision stump for each possible hypothesis and determine which achieves the highest accuracy with the training data. Once the hypothesis is known, a decision stump can be built by assigning a level to each leaf based on the most frequent level for selected examples in the training set (Bird *et al.* 2009).

REPTree

REPTree algorithm is a fast decision tree learner. It builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back-fitting). The algorithm only sorts values for numeric attributes once (Daud & Corne 2007). REPTree uses standard techniques from C4.5 and classification and regression trees (CART) methods (Breiman *et al.* 1984; Quinlan 1996).

We use the WEKA (Witten & Frank 2005) implementations of the various learning algorithms; WEKA is written in Java and is freely available from www.cs.waikato.ac.nz/~ml.

Different error norms

The different statistical parameters of each model are calculated during calibration to get the best statistical agreement between observed and simulated meteorological variables. For this purpose, various statistical performance measures such as coefficient of correlation (CC), root mean square error (RMSE) and mean absolute error (MAE) were used to measure the performance of various models. These measures are defined below.

The coefficient of correlation (CC) is defined:

$$\mathrm{CC} = \frac{\sum_{i=1}^{N}(Y_o - \bar{Y}_o)(Y_c - \bar{Y}_c)}{\left[\sum_{i=1}^{N}(Y_o - \bar{Y}_o)^2 \cdot \sum_{i=1}^{N}(Y_c - \bar{Y}_c)^2\right]^{1/2}} \quad (2)$$

where N represents the number of feature vectors prepared from the NCEP record, Y_o and Y_c denote the observed and the simulated values of predictand, respectively, and $\bar{Y}_o$ and σ_{obs} are the mean and the standard deviation of the observed predictand.

The RMSE between observed and computed outputs is defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(Y_c - Y_o)^2}{N}} \quad (3)$$

Mean absolute error (MAE) is defined as follows (Johnson *et al.* 2003):

$$\mathrm{MAE} = 1 - \frac{\sum_{i=1}^{N}|Y_c - Y_o|}{\sum_{i=1}^{N}|Y_o - \bar{Y}_o|} \quad (4)$$

Selection of predictors

The selection of appropriate predictors is one of the most important steps in a downscaling exercise for downscaling predictands. The predictors are chosen by the following

criteria: (1) predictors are skilfully predicted by GCMs; (2) they should represent important physical processes in the context of the enhanced greenhouse effect; and (3) they should not be strongly correlated to each other (Hewitson & Crane 1996; Hellström *et al.* 2001; Wilby *et al.* 2002; Cavazos & Hewitson 2005). Various authors (e.g. Dibike & Coulibaly 2006; Tripathi *et al.* 2006; Anandhi *et al.* 2008; Goyal & Ojha 2011a,c) have used large-scale atmospheric variables – namely air temperature (at 925, 500 and 200 mb pressure levels), geopotential height (at 500 and 200 mb pressure levels), zonal (u) and meridional (v) wind velocities (at 925 and 200 mb pressure levels) – as the predictors for downscaling GCM output to mean monthly precipitation over a catchment. These atmospheric variables have therefore been used as potential predictors in this study.

Cross-correlations are in use to select predictors to understand the presence of non-linearity/linearity trend in dependence structure (Dibike & Coulibaly 2006). These cross-correlations between each of the predictor variables in NCEP and GCM datasets are useful to verify if the predictor variables are realistically simulated by the GCM. Cross-correlations are computed between the predictor variables in NCEP and GCM datasets (Table 1). The cross-correlations are estimated using three measures of dependence (namely product moment correlation, Spearman's rank correlation and Kendall's tau scatter plots) and cross-correlations between each of the predictor variables in NCEP and GCM datasets are useful to verify if the predictor variables are realistically simulated by the GCM (Anandhi *et al.* 2008).

The product moment correlation which measures the linear relationship between probable predictor and predictand is given by Pearson (1896). Spearman's rank correlation and Kendall's tau are the two non-parametric correlations used in this study which are estimated based on ranks assigned to data points in predictor and predict and datasets. The advantage of these rank correlations over the linear correlation stems from the use of ranks rather than numerical values of the predictor and the predictand variables for estimation of the correlations (Press *et al.* 1992).

Spearman's rank correlation ρ is computed using the difference between the ranks of contemporaneous values of predictor and predictand, D_i (Spearman 1904a, b):

$$\rho = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N(N^2-1)} \tag{5}$$

Kendall's τ is calculated as (Kendall 1951):

$$\tau = \frac{4\lambda}{(1/2)N(N-1)} - 1 \tag{6}$$

where λ is the difference between the number of concordant pairs and the number of discordant pairs.

Downscaling of GCM models

Downscaling models make use of a strong observed empirical relationship between one or several large-scale predictors and a variable of interest at a regional scale, the predictand. Rule induction and tree algorithms, as mentioned earlier, are used to downscale mean monthly precipitation in this study. The data of potential predictors is first standardized. Standardization is widely used prior to statistical downscaling to reduce bias (if any) in the mean and the variance of GCM predictors with respect to that of NCEP-reanalysis data. Standardization is done for a baseline period of 1948–2000; it is of sufficient duration to establish a reliable climatology, yet not too long nor too contemporary to include a strong global change signal

Table 1 | Cross-correlation computed between probable predictors in NCEP and GCM datasets. *P*, *S* and *K* represent product moment correlation, Spearman's rank correlation and Kendall's tau, respectively

	Ta925	Ua925	Va925	Va200	Ta20	Zg200	Ua200	Ta500	Zg500
P	0.83	0.79	0.67	–0.18	0.66	0.81	0.23	0.81	0.60
S	0.68	0.56	0.43	–0.14	0.46	0.64	0.57	0.64	0.39
K	0.87	0.76	0.61	–0.20	0.68	0.85	0.73	0.85	0.59

(Ghosh & Mujumdar 2008). The procedure typically involves subtraction of mean and division by standard deviation of the predictor variable for a predefined baseline period for both NCEP and GCM output.

Multi-dimensionality of the predictors may lead to a computationally complicated and large-sized model with high multi-collinearity (high correlation between the explanatory variables/regressors). To reduce the dimensionality of the explanatory dataset, PCA is performed. The use of PCs as input to a downscaling model helps in making the model more stable and, at the same time, reduces its computational burden. The data of standardized NCEP predictor variables is then processed using principal component analysis to extract PCs which are orthogonal. A feature vector is formed for each month of the record using the PCs. The feature vector is the input to the models and the contemporaneous value of predictand is the output.

To develop downscaling models, the feature vectors (i.e. predictors) which are prepared from the NCEP record are partitioned into a training set and a validation set. Feature vectors in the training set are used for calibrating the model, and those in the validation set are used for validation. The 26-year mean monthly observed precipitation data were broken up into a calibration period and a validation period. Table 2 summarizes certain details of the models. The models were calibrated and validated using data from the periods 1975–1989 and 1990–2000, respectively. The various error criteria are used as an index to assess the performance of the model. Based on the latest IPCC scenario, namely A1B, A2, B1 and COMMIT (IPCC 2007), models for mean monthly precipitation were evaluated based on the accuracy of the predictions for the validation dataset.

Table 2 | Different downscaling model variants used in the study to obtain projections of predictand (precipitation) at monthly timescale

Model	Period of downscaling	Length of the record	Approach
1	1975–2100	1975–2000	Single Conjunctive Rule Learner
2	1975–2100	1975–2000	Decision Table
3	1975–2100	1975–2000	M5 Model Tree
4	1975–2100	1975–2000	Decision Stump
5	1975–2100	1975–2000	REPTree

RESULTS AND DISCUSSION

Downscaling models were developed following the methodology described in the previous section.

Potential predictor selection

The most relevant probable predictor variables necessary for developing the downscaling models are identified by using the three measures of dependence following the procedure. The cross-correlations enable the reliability of the simulations of the predictor variables to be verified by the GCM, and are shown in Table 1. A pool of potential predictors is identified by specifying threshold values for the computed cross-correlations. To aid in this task, a threshold of 0.6 is chosen for product moment correlation to segregate high and low correlations. However, the choice of threshold is subjective. A higher threshold results in the selection of a few of the probable predictors as potential predictors. In contrast, a very low threshold results in the selection of almost all the probable predictors as potential predictors. In general, most of the predictor variables are realistically simulated by the GCM. It is noted that air temperature at 925 mb (Ta925) is the most realistically simulated variable with a CC greater than 0.8, while meridional wind at 200 mb (Va200) is the least correlated variable between NCEP and GCM datasets (CC = −0.17). It is clear from Table 1 that air temperature at 925 mb (Ta925), 500 mb (Ta500) and 200 mb (Ta200), meridional wind at 925 mb (Va 925), zonal wind at 925 mb (Ua925), geopotential height at 200 mb (Zg200) and 500 mb (Zg500) are better correlated than meridional wind at 200 mb (Va200) and zonal wind at 200 mb (Ua200).

Downscaling and performance of GCM models

Seven predictor variables – namely air temperature at 925, 500 and 200 mb, zonal wind at 925 mb, meridional wind at 925 mb and geopotential height at 500 and 200 mb at

nine NCEP grid points with a dimensionality of 63 – are used as the standardized data of potential predictors.

PCA is performed to transform the set of correlated N-dimensional predictors ($N = 63$) into another set of N-dimensional uncorrelated vectors (called principal components) by linear combination, such that most of the information content of the original dataset is stored in the first few dimensions of the new set. It is observed that the four leading PCs of the PCA method explained about 97% of the information content (or variability) of the original predictors. Figure 4 shows the percentage of variance explained by the principal components. PCs are therefore extracted to form feature vectors from the standardized data of potential predictors. These feature vectors are provided as input to the various downscaling models. Results of the different models (namely, models 1–5) for predictand as discussed in Table 2 are tabulated in Table 3.

It can be seen from Table 3 that the coefficient of correlation (CC) was in the range of 0.80–0.93, RMSE was in the range 32.48–74.68 and MAE was in the range 19.16–50.82 for models (namely, models 1–5) for the training and validation set for the predictand. It can also be observed from Table 3 that the performance of model 3 using M5 Model Tree algorithm for mean monthly precipitation is clearly superior to that of Single Conjugative Rule Learner, Decision Table, Decision Stump and REPTree for both the training and validation dataset. M5 Model Tree, in contrast to others, divides the input space into a number of subspaces; a separate specialized model is built for each subspace.

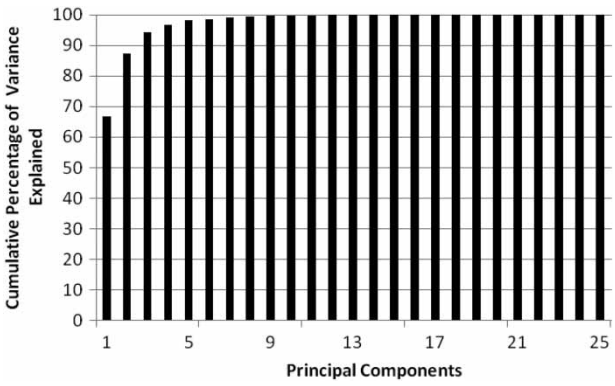

Figure 4 | Percentage of variance explained by the principal components.

A comparison of mean monthly observed precipitation with precipitation simulated for model 3 using M5 Model Tree algorithm is shown in Figure 5 for the validation period. The Appendix shows pruned model trees obtained by using the M5 Model Tree algorithm for the predictand.

Once the downscaling models have been calibrated and validated, the next step is to use these models to downscale the scenarios simulated by the GCM. The GCM simulations are run through the calibrated and validated model (3) to obtain future simulations of the predictand. The predictand (precipitation) patterns are analyzed with box plots for 20-year time slices. The middle line of the box gives the median, whereas the upper and lower edges give the 75 percentile and 25 percentile of the dataset, respectively. The difference between the 75 percentile and 25 percentile is known as the inter-quartile range (IQR). The two bounds of a box plot outside the box denote the value to be 1.5 times IQR lower than the third quartile or minimum value, whichever is high, and 1.5 times higher than the third quartile or the maximum value, whichever is less.

Typical results of downscaled predictand (precipitation) obtained from the predictors are presented in Figure 6. In Figure 6(a), raw output from the GCM is compared to the results from downscaled data from the GCM along with the observed precipitation for the study region using box

Table 3 | Various performance statistics of models using various approaches

Model	CC		MAE		RMSE	
	Training	Validation	Training	Validation	Training	Validation
1	0.80	0.40	38.52	50.82	54.28	70.22
2	0.82	0.49	31.69	49.70	51.07	74.68
3	0.93	0.77	19.16	32.66	32.48	50.04
4	0.80	0.40	36.69	50.66	54.08	72.24
5	0.92	0.67	20.14	35.05	36.09	61.28

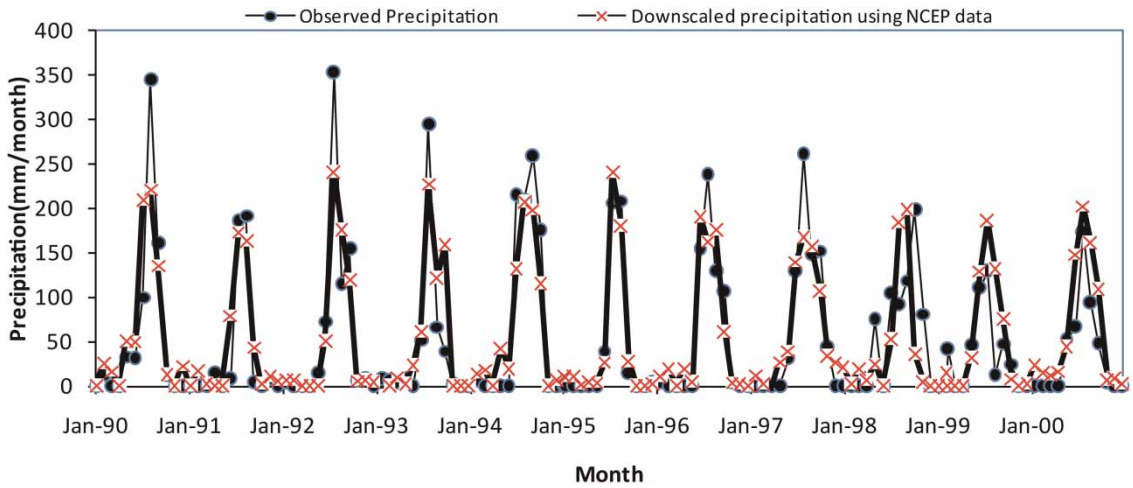

Figure 5 | Typical results comparing monthly observed precipitation with precipitation simulated for model 3 using M5 Model Tree for NCEP data for validation period 1990–2000.

plots. It can be inferred that the mean of the precipitation is underestimated by about 17% using raw output of the GCM directly, while the mean of the precipitation is overestimated by about 2% using the downscaled precipitation of the same GCM data. This study therefore confirms that statistical downscaling adds value for regional/local climate prediction when compared to GCM results. The projected precipitation for 2001–2020, 2021–2040, 2041–2060, 2061–2080 and 2081–2100 for the four scenarios A1B, A2, B1 and COMMIT are shown in Figure 6(b), (e), respectively.

From the box plots of downscaled predictand (Figure 6), it can be observed that precipitations are projected to increase in future for A1B, A2 and B1 scenarios. The projected increase of precipitation is highest for A1B and A2 scenarios and least for the B1 scenario. The annual amount of precipitation would increase by about 6 and 8% for A1B and A2 scenarios; for the B1 scenario, it would be about 2%. This is because of the scenarios considered, A1B and A2 have the highest concentration of atmospheric carbon dioxide (CO_2) equal to 720 and 850 ppm; CO_2 concentrations for B1 and COMMIT scenarios are about 550 and 370 ppm, respectively. The rise in concentration of CO_2 in the atmosphere causes the Earth's average temperature to increase which in turn causes an increase in evaporation, especially at lower latitudes. The evaporated water would eventually precipitate (Anandhi *et al.* 2008). In the COMMIT scenario, where the emissions are held the same as for the year 2000, no significant trend in the pattern of projected future precipitation could be discerned. The overall results show that the projections obtained for precipitation are indeed robust.

One of the concerns in the literature is whether atmospheric temperature, zonal winds, meridional winds and geopotential height (at range of different heights) are best predictors for precipitation. This aspect has been addressed by assessing correlation; accordingly, the variables having higher values of coefficient of correlation have been considered. In the absence of this approach, there does not appear to be a better way of handling the selection of variables, as evidenced in the literature.

Comparison with previous downscaling studies

While this is the first study to compare various types of decision tree algorithms for precipitation projections in the Rajasthan, India, there have been a few studies using other methods in other parts of India. It is therefore worthwhile to relate the performance of the models presented here to those presented in other studies that closely relate to this study.

Anandhi *et al.* (2008) developed downscaling models using a SVM to obtain projections of monthly precipitation to river basin scale in India for the catchment of the Malaprabha reservoir. The results of validation indicate that the SVM model is a feasible choice for downscaling the predictand (i.e. precipitation). The resulting models produced

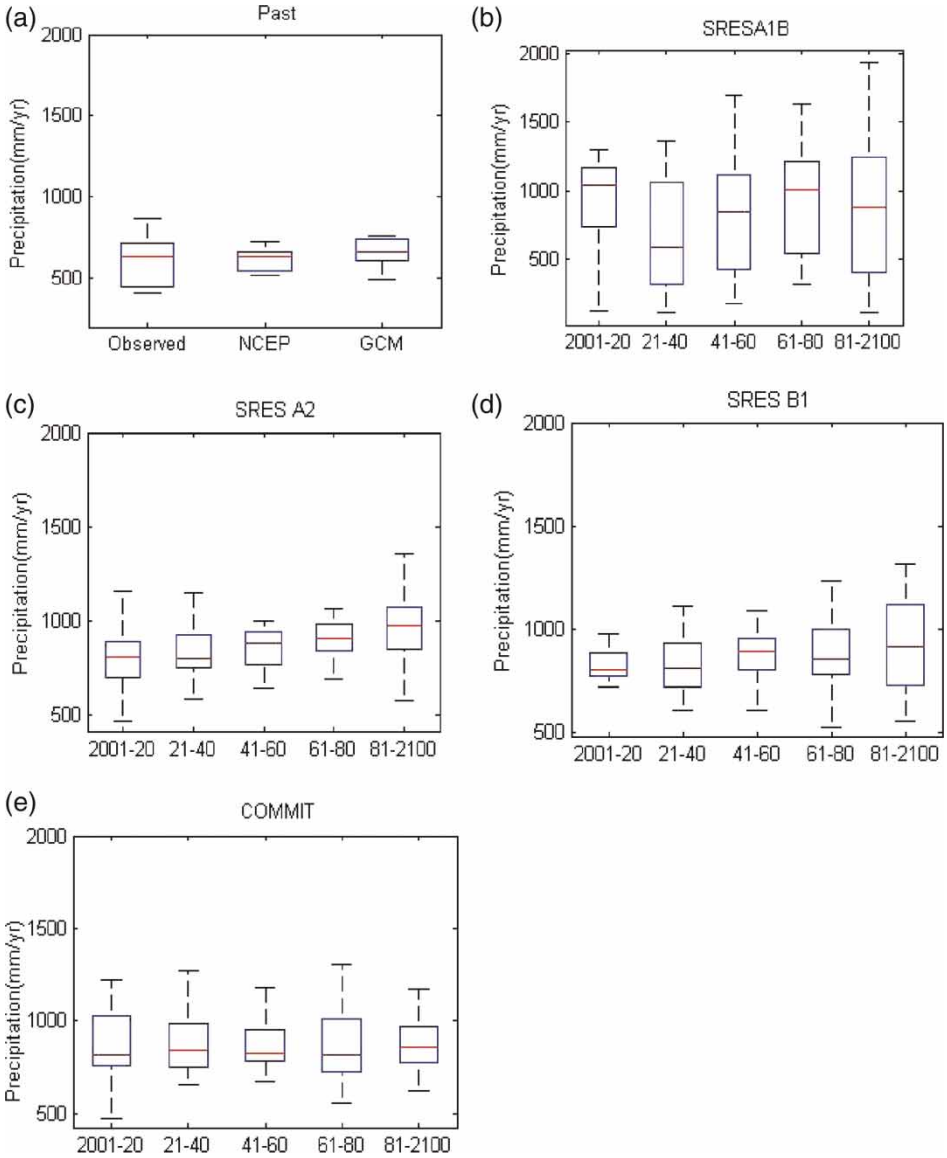

Figure 6 | Box plot results of downscaling model 3 using the M5 Model Tree for the predictand precipitation.

similar results to those of this study; for example, the results of downscaling show that precipitation is projected to increase in the future for almost all the scenarios considered.

CONCLUSIONS

This paper explores the suitability of various available rule and decision tree learning algorithms (namely Single Conjunctive Rule Learner, Decision Table, M5 Model Tree, Decision Stump and REPTree) to downscale mean monthly precipitation from GCM output to local scale. The predictand is downscaled from simulations of CGCM3 for four IPCC scenarios, namely Special Report of Emission Scenarios (SRES) A1B, A2, B1 and COMMIT through the application of lake catchment in arid regions in India. Five models based on algorithms have been developed and the performance of the models is evaluated using the statistical measures CC, RMSE

and MAE. The overall conclusions of this evaluation study are as follows.

- M5 Model Tree algorithm performed best. The ranking of these approaches in order of decreasing performance for predictand precipitation is: M5 Model Tree, REPTree, Decision Table, Single Conjunctive Rule Learner and Decision Stump.
- The results of downscaling models using M5 Model Tree algorithm show that precipitation is projected to increase most in the future for the A2 and A1B scenarios, whereas it is least for B1 and COMMIT scenarios.
- The study clearly indicates that decision tree algorithms have potential for application in statistical downscaling.

ACKNOWLEDGEMENTS

The authors are grateful to two unknown reviewers for their insightful comments and suggestions which improved the paper.

REFERENCES

Anandhi, A., Srinivas, V. V., Nanjundiah, R. S. & Kumar, D. N. 2008 Downscaling precipitation to river basin for IPCC SRES scenarios using support vector machines. *International Journal of Climatology* **28**, 401–420.

Anandhi, A., Srinivas, V. V., Kumar, D. N. & Nanjundiah, R. S. 2009 Role of predictors in downscaling surface temperature to river basin in India for IPCC SRES scenarios using support vector machine. *International Journal of Climatology* **29**, 583–603.

Bardossy, A., Bogardi, I. & Matyasovszky, I. 2005 Fuzzy rule-based downscaling of precipitation. *Theoretical and Applied Climatology* **82**, 119–129.

Benestad, R. E., Hanssen-Bauer, I. & Førland, E. J. 2007 An evaluation of statistical models for downscaling precipitation and their ability to capture long-term trends. *International Journal of Climatology* **27** (5), 649–655.

Bhattacharya, B. & Solomatine, D. P. 2005 Neural networks and M5 model trees in modelling water level–discharge relationship. *Neurocomputing* **63**, 381–396.

Bhattacharya, B., Price, R. K. & Solomatine, D. P. 2007 Machine learning approach to modeling sediment transport. *Journal of Hydraulic Engineering* **133** (4), 440–450.

Bird, S., Klein, E. & Loper, E. 2009 *Natural Language Processing with Python*. O'Reilly Media, Inc., Sebastopol.

Breiman, L., Friedman, J. H., Ohlsen, R. A. & Stone, J. C. 1984 *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Cannon, A. J. & Lord, E. R. 2000 Forecasting summertime surface-level ozone concentrations in the Lower Fraser Valley of British Columbia. An ensemble neural network approach. *Journal of the Air and Waste Management Association* **50**, 322–339.

Cavazos, T. & Hewitson, B. C. 2005 Performance of NCEP variables in statistical downscaling of daily precipitation. *Climate Research* **28**, 95–107.

Christensen, J., Hewitson, B., Busuioc, A., Chen, A., Gao, X., Jones, R., Kolli, R., Kwon, W.-T., Magaña Rueda, V., Mearns, L., Meñendez, C., Raisanen, J., Rinke, A., Sarr, A. & Whetton, P. 2007 *Climate change 2007: the physical science basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change: Regional Climate Projections. Cambridge University Press, Cambridge, United Kingdom and New York, NY (Chapter 11).

Chu, J. L., Kang, H., Tam, C. Y., Park, C. K. & Chen, C.-T. 2008 Seasonal forecast for local precipitation over northern Taiwan using statistical downscaling. *Journal of Geophysical Research* **113**, D12118.

Chu, J. T., Xia, J., Xu, C.-Y. & Singh, V. P. 2010 Statistical downscaling of daily mean temperature, pan evaporation and precipitation for climate change scenarios in Haihe River, China. *Theoretical and Applied Climatology* **99**, 149–161.

Clark, P. & Niblett, T. 1989 The CN2 rule induction algorithm. *Machine Learning* **3**, 261–284.

Cohen, W. 1995 Fast effective rule induction. In *Proceedings of 12th International Conference on Machine Learning*, Morgan Kaufmann, pp. 115–123.

Conway, D., Wilby, R. L. & Jones, P. D. 1996 Precipitation and air flow indices over the British Isles. *Climate Research* **7**, 169–183.

Crane, G. & Hewitson, B. C. 1998 Doubled CO_2 precipitation changes for the Susquehanna basin: down-scaling from the genesis general circulation model. *International Journal of Climatology* **18**, 65–76.

Das, M. R., Mukhopadhyay, R. K., Dandekar, M. M. & Kshirsagar, S. R. 2010 Pre-monsoon western disturbances in relation to monsoon rainfall, its advancement over NW India and their trends. *Current Science* **82** (11), 1320–1321.

Daud, M. N. R. & Corne, D. W. 2007 Human readable rule induction in medical data mining: a survey of existing algorithms. In *Proceedings of WSEAS European Computing Conference*, Athens, Greece.

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

Dubrovsky, M., Buchtele, J. & Zalud, Z. 2004 High-frequency and low frequency variability in stochastic daily weather generator and its effect on agricultural and hydrologic modelling. *Climatic Change* **63**, 145–179.

Fowler, H. J. & Wilby, R. L. 2007 Beyond the downscaling comparison study. *International Journal of Climatology* **27**, 1543–1545.

Fowler, H. J., Kilsby, C. G. & O'Connell, P. E. 2000 A stochastic rainfall model for the assessment of regional water resource systems under changed climatic conditions. *Hydrology and Earth System Sciences* **4**, 261–280.

Fowler, H. J., Blenkinsop, S. & Tebaldi, C. 2007 Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology* **27**, 1547–1578.

Gangopadhyay, S., Clark, M. & Rajagopalan, B. 2005 Statistical downscaling using K-nearest neighbors. *Water Resources Research* **41**, W02024.

Gardner, M. W. & Dorling, S. R. 1998 Artificial neural networks (the multi layer perceptron): a review of applications in the atmospheric sciences. *Atmospheric Environment* **32**, 2627–2636.

Ghosh, S. & Mujumdar, P. P. 2006 Future rainfall scenario over Orissa with GCM projections by statistical downcasting. *Current Science* **90** (3), 396–404.

Ghosh, S. & Mujumdar, P. P. 2008 Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Advances in Water Resources* **31**, 132–146.

Giorgi, F. & Hewitson, B. C. 2001 Regional climate information – evaluation and projections. In: *Climate Change 2001: The Scientific Basis* (J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dia, K. Maskell & C. A. Johnson, eds.). Cambridge University Press, Cambridge.

Goel, A. & Singh, R. B. 2006 Climatic variability and drought in Rajasthan. Advances in Geosciences. Vol. 4. *Hydrological Sciences*. World Scientific Publishing Co. Ltd., Singapore.

Goodess, C. M. & Palutikof, J. 1998 Development of daily rainfall scenarios for southeast Spain using a circulation-type approach to downscaling. *International Journal of Climatology* **18**, 1051–1083.

Goyal, M. K. & Ojha, C. S. P. 2011a Downscaling of surface temperature for lake catchment in arid region in India using linear multiple regression and neural networks. *International Journal of Climatology* in press.

Goyal, M. K. & Ojha, C. S. P. 2011b Estimation of Scour downstream of a ski – jump bucket using. Support vector and M5 model tree. *Water Resources Management* **25** (9), 2177–2195.

Goyal, M. K. & Ojha, C. S. P. 2011c PLS regression based Pan evaporation and minimum-maximum temperature projections for an arid lake basin in India. *Theoretical and Applied Climatology* **105** (3–4), 403–415.

Goyal, M. K., Ojha, C. S. P. & Burn, D. H. 2011 Nonparametric statistical downscaling of temperature, precipitation and evaporation for semi-arid region in India. *Journal of Hydrologic Engineering* in press.

Grotch, S. L. & MacCracken, M. C. 1991 The use of general circulation models to predict regional climatic change. *Journal of Climate* **4**, 286–303.

Hanssen-Bauer, I., Achberger, C., Benestad, R. E., Chen, D. L. & Forland, E. J. 2005 Statistical downscaling of climate scenarios over Scandinavia. *Climate Research* **29**, 255–268.

Hellström, C., Chen, D., Achberger, C. & Räisänen, J. 2001 Comparison of climate change scenarios for Sweden based on statistical and dynamical downscaling of monthly precipitation. *Climate Research* **19**, 45–55.

Hewitson, B. C. & Crane, R. G. (eds.). 1994 *Neural Nets Applications in Geography*. Kluwer Academic Publishers, Dordrecht.

Hewitson, B. C. & Crane, R. G. 1996 Climate downscaling: techniques and application. *Climate Research* **7**, 85–95.

Heyen, H., Zorita, E. & von Storch, H. 1996 Statistical downscaling of monthly mean North Atlantic air-pressure to sea level anomalies in the Baltic Sea. *Tellus* **48A**, 312–323.

Ibrahim, F., Abu Osman, N. A., Usman, J. & Kadri, N. A. (eds.). 2006 Biomed 06, IFMBE Proceedings 15, pp. 520–523.

Intergovernmental Panel on Climate Change (IPCC) 2007 Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (S. D. Solomon, M. Qin, Z. Manning, M. Chen, K. B. Marquis, M. Averyt, H. L. Tignor, eds.). Cambridge University Press, Cambridge.

Jessie, C. R., Antonio, R. M. & Stahis, S. P. 1996 *Climate Variability, Climate Change and Social Vulnerability in the Semi-arid Tropics*. Cambridge University Press, Cambridge.

Johnson, M. S., Coon, W. F., Mehta, V. K., Steenhuis, T. S., Brooks, E. S. & Boll, J. 2003 Application of two hydrologic models with different runoff mechanisms to a hillslope dominated watershed in the northeastern US: a comparison of HSPF and SMR. *Journal of Hydrology* **284**, 57–76.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R. & Joseph, D. 1996 The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77** (3), 437–471.

Kang, H. W., An, K. H., Park, C. K., Solis, A. L. S. & Stitthichivapak, K. 2007 Multimodel output statistical downscaling prediction of precipitation in the Philippines and Thailand. *Geophysical Research Letters* **34**, L15710.

Kendall, M. G. 1951 Regression structure and functional relationship Part I. *Biometrika* **38**, 11–25.

Khobragade, S. D. 2009 *Studies on evaporation from open water surfaces in tropical climate*. PhD Thesis, Indian Institute of Technology, Roorkee, India.

Kilsby, C. G., Cowpertwait, P. S. P., O'Connell, P. E. & Jones, P. D. 1998 Predicting rainfall statistics in England and Wales using atmospheric circulation variables. *International Journal of Climatology* **18**, 523–539.

Kite, G. W. 1997 Simulating Columbia river flows with data from regional-scale climate models. *Water Resources Research* **33** (6), 1275–1285.

Kohavi, R. 1995 The power of decision tables. *Machine Learning* **914**, 174–189.

Linz, H., Shiklomanov, I. & Mostefakara, K. 1990 *Chapter 4 Hydrology and water: likely impact of climate change*. IPCC WGII report, WMO/UNEP, Geneva.

Mason, S. J. 2004 Simulating climate over western North America using stochastic weather generators. *Climatic Change* **62**, 155–187.

Murphy, J. 2000 Predictions of climate change over Europe using statistical and dynamical downscaling techniques. *International Journal of Climatology* **20**, 489–501.

Ojha, C. S. P., Goyal, M. K. & Adeloye, A. J. 2010 Downscaling of precipitation for lake catchment in arid region in India using linear multiple regression and neural networks. *The Open Journal of Hydrology* **4**, 122–136.

Othman, M. F. b. & Yau, T. M. S. 2007 Comparison of different classification techniques using WEKA for breast cancer. *Biomed 06: IFMBE Proceedings* **15**, 520–523.

Nilsson, J. N. 1965 *Introduction to Machine Learning*. McGraw-Hill, New York.

Pearson, K. 1896 Mathematical contributions to the theory of evolution III regression heredity and panmixia. *Philosophical Transactions of the Royal Society of London Series* **187**, 253–318.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. 1992 *Numerical Recipes in Fortran 77: The Art of Scientific Computing*. Cambridge University Press, New York.

Prudhomme, C., Jakob, D. & Svensson, C. 2003 Uncertainty and climate change impact on the flood regime of small UK catchments. *Journal of Hydrology* **277**, 1–23.

Quinlan, J. R. 1992 Learning with continuous classes. In *Proceeding of the Fifth Australian Joint Conference on Artificial Intelligence*. World Scientific, Singapore, pp. 343–348.

Quinlan, J. R. 1996 Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research* **4**, 77–90.

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

Shahidi, A. E. & Mahjoobi, J. 2009 Comparison between M5 model tree and neural networks for prediction of significant wave height in Lake Superior. *Ocean Engineering* **36**, 1175–1181.

Shannon, D. A. & Hewitson, B. C. 1996 Cross-scale relationships regarding local temperature inversions at Cape Town and global climate change implications. *South African Journal of Science* **92** (4), 213–216.

Solomatine, D. P. & Dulal, K. N. 2003 Model tree as an alternative to neural network in rainfall–runoff modelling. *Hydrological Sciences Journal* **48** (3), 399–411.

Solomatine, D. P. & Yunpeng, X. 2004 M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering* **9** (6), 491–501.

Spearman, C. E. 1904a General intelligence objectively determined and measured. *American Journal of Psychology* **5**, 201–293.

Spearman, C. E. 1904b Proof and measurement of association between two things. *American Journal of Psychology* **15**, 72–101.

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.

Vimont, J. D., Battisti, D. S. & Naylor, R. L. 2010 Downscaling Indonesian precipitation using large-scale meteorological fields. *Int. J. Climatol.* **30** (11), 1706–1722.

Vrac, M., Stein, M. L., Hayhoe, K. & Liang, X. Z. 2007 A general method for validating statistical downscaling methods under future climate change. *Geophysical Research Letters* **34**, L18701.

Wang, Y. & Witten, I. H. 1997 Induction of model trees for predicting continuous lasses. In *Proceedings of the Poster Papers of the European Conference on Machine Learning*. University of Economics, Faculty of Informatics and Statistics, Prague.

Wang, Y. Q., Leung, L. R., Mcgregor, J. L., Wang, W. C., Ding, Y. H. & Kimura, F. 2004 Regional climate modeling: progress, challenges, and prospects. *Journal of the Meteorological Society of Japan* **82** (6), 1599–1628.

Wetterhall, F., Halldin, S. & Xu, C. Y. 2005 Statistical precipitation downscaling in central Sweden with the analogue method. *Journal of Hydrology* **306**, 136–174.

Wetterhall, F., Halldin, S. & Xu, C.-Y. 2007 Seasonality properties of four statistical-downscaling methods in central Sweden. *Theoretical and Applied Climatology* **87** (1–4), 123–137.

Wilby, R. L. & Harris, I. 2006 A framework for assessing uncertainties in climate change impacts: low-flow scenarios. *Water Resources Research* **42**, W02419.

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 climate change impacts. *Environmental Modelling and Software* **17**, 147–159.

Wilks, D. S. 1989 Conditioning stochastic daily precipitation models on total monthly precipitation. *Water Resources Research* **25** (6), 1429–1439.

Wilks, D. S. 1992 Adapting stochastic weather generation algorithms for climate change studies. *Climate Change* **22**, 67–84.

Willmott, C. J., Rowe, C. M. & Philpot, W. D. 1985 Small-scale climate map: a sensitivity analysis of some common assumptions associated with the grid-point interpolation and contouring. *American Cartographer* **12**, 5–16.

Witten, I. H. & Frank, E. 2005 *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S. J. 1999 WEKA: Practical machine learning tools and techniques with java implementations. In *Proceedings of ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Info. Systems*, 192–196.

Xoplaki, E., Luterbacher, J., Burkard, R., Patrikas, I. & Maheras, P. 2000 Connection between the large-scale

500 hPa geopotential height fields and precipitation over Greece during wintertime. *Climate Research* **14**, 129–146.

Xu, C.-Y. 1999 From GCMs to river flow: a review of downscaling methods and hydrologic modelling approaches. *Progress in Physical Geography* **23** (2), 229–249.

First received 9 May 2010; accepted in revised form 19 January 2011

APPENDIX

Pruned model tree obtained from M5 modelling approach for precipitation

In the following, 'pc' represents 'principal component' and the number of rules is 4.

pc1 < =− 0.623: LM1 (36/4.07%)
pc1 > −0.623:
|pc2 < =0.324: LM2 (19/18.38%)
|pc2 > 0.324:
| | pc3 < =− 1.145: LM3 (4/65.264%)
| | pc3 > −1.145: LM4 (13/68.017%)

LM1
$$\text{Precipitation} = (2.6785 \times \text{pc1}) + (6.9488 \times \text{pc2}) - (2.3238 \times \text{pc3}) + (7.0986 \times \text{pc4}) + 17.3294$$

LM2
$$\text{Precipitation} = (3.76 \times \text{pc1}) + (17.2173 \times \text{pc2}) - (2.3238 \times \text{pc3}) + (15.6429 \times \text{pc4}) + 50.5933$$

LM3
$$\text{Precipitation} = (2.6785 \times \text{pc1}) + (15.9142 \times \text{pc2}) - (13.3619 \times \text{pc3}) + (16.1769 \times \text{pc4}) + 126.1765$$

LM4
$$\text{Precipitation} = (2.6785 \times \text{pc1}) + (15.9142 \times \text{pc2}) - (9.814 \times \text{pc3}) + (16.1769 \times \text{pc4}) + 112.2409$$