Theor Appl Climatol (2012) 108:85–104
DOI 10.1007/s00704-011-0511-z

ORIGINAL PAPER

Daily relative humidity projections in an Indian river basin for IPCC SRES scenarios

Aavudai Anandhi · V. V. Srinivas · D. Nagesh Kumar · Ravi S. Nanjundiah

Received: 29 June 2010 / Accepted: 18 August 2011 / Published online: 9 September 2011

Abstract A two-stage methodology is developed to obtain future projections of daily relative humidity in a river basin for climate change scenarios. In the first stage, Support Vector Machine (SVM) models are developed to downscale nine sets of predictor variables (large-scale atmospheric variables) for Intergovernmental Panel on Climate Change Special Report on Emissions Scenarios (SRES) (A1B, A2, B1, and COMMIT) to R_H in a river basin at monthly scale. Uncertainty in the future projections of R_H is studied for combinations of SRES scenarios, and predictors selected. Subsequently, in the second stage, the monthly sequences of R_H are disaggregated to daily scale using *k*-nearest neighbor method. The effectiveness of the developed methodology is demonstrated through application to the catchment of Malaprabha reservoir in India. For downscaling, the probable predictor variables are extracted from the (1) National Centers for Environmental Prediction reanalysis data set for the period 1978–2000 and (2) simulations of the third-generation Canadian Coupled Global Climate Model for the period 1978–2100. The performance of the downscaling and disaggregation models is evaluated by split sample validation. Results show that among the SVM models, the model developed using predictors pertaining to only land location performed better. The R_H is projected to increase in the future for A1B and A2 scenarios, while no trend is discerned for B1 and COMMIT.

A. Anandhi · V. V. Srinivas · D. N. Kumar
Department of Civil Engineering, Indian Institute of Science,
Bangalore 560 012, India

A. Anandhi
CUNY Institute for Sustainable Cities,
City University of New York,
New York, NY 10065, USA

D. N. Kumar (✉)
Center for Earth Sciences, Indian Institute of Science,
Bangalore 560 012, India
e-mail: nagesh@civil.iisc.ernet.in

R. S. Nanjundiah
Centre for Atmospheric & Oceanic Sciences,
Indian Institute of Science,
Bangalore 560012, India

1 Introduction

Humidity is a decisive factor (besides wind speed and temperature) in determining the amount and rate of evaporation and transpiration. Besides this, humidity is found to have substantial effect on plant growth, as it affects rate of leaf emergence, plant height, leaf area, leaf blade length, the number of roots, total root length, and dry matter production (Hirai et al. 2000). Surface humidity regulates evaporation and transpiration processes and so has obvious connections to both hydrological and surface energy budgets (Gaffen and Ross 1999). This makes it an important variable in hydrological and agronomy studies, especially in regions with an arid and semi-arid climate.

Information on spatio-temporal variations in relative humidity (R_H) finds use in (1) evapotranspiration estimation models (e.g., Penman–Monteith; Monteith 1981), (2) crop simulation models (e.g., Erosion Productivity Impact Calculator model; Brown and Rosenberg 1997; Easterling et al. 1992), (3) mesoscale models to parameterize the cloud radiative effects (Koch et al. 1997), (4) modeling greenhouse (Stanghellini and de Jong 1995), and (5) hydrologic models. Relative humidity is an important factor in determining the distribution and occurrence of clouds (Wright et al. 2010). Also, surface humidity is the principal

source for the free troposphere where water vapor has important implications for earth's radiation and energy budgets and therefore climate sensitivity (Willett et al. 2010). Hence, it is necessary to assess the implications of projected climate change on R_H at river basin scale.

The present study is motivated to develop a two-stage methodology to obtain future projections of daily R_H in a river basin for various climate change scenarios. The remainder of this paper is structured as follows: Sect. 2 presents the background of the study and objectives set for the study. Section 3 provides details of data considered. The description of the study region is provided in Sect. 4. Section 5 describes the methodology developed in the study to arrive at daily sequences of R_H at river basin scale for different climate change scenarios. Section 6 presents the results and discussion. Finally, Sect. 7 provides summary of the work presented and conclusions drawn from the study.

2 Background

General climate models (GCMs) simulate climatic conditions on earth, hundreds of years into the future for plausible future climates referred to as scenarios. However, as GCMs are generally run at coarser scale to cover the whole globe, they are ineffective in simulating local climatic conditions on earth and in providing consistent estimates of climate change on a river basin scale (Houghton et al. 2001). Techniques such as downscaling and disaggregation are used to translate the large-scale information provided by GCM to the spatial scales that are relevant to regional impact studies (Houghton et al. 2001). The downscaling methods in vogue in literature, the advantages of using Support Vector Machine (SVM) (Vapnik 1995, 1998) for downscaling, and the fundamental principle of SVM and its formulation are discussed in Appendix 1.

Review of latest literature on statistical downscaling of R_H and the methods available for temporal disaggregation of R_H are presented in Tables 1 and 2, respectively. From these tables, it can be observed that in the past, very few studies have addressed downscaling and disaggregation of R_H though it is one of the important variables for climate change impact assessment. In the present study, a two-stage methodology based on SVM is advocated for downscaling R_H. In the first stage of the two-stage model, a SVM model is developed to downscale the large-scale atmospheric variables to R_H in a river basin at monthly scale. Subsequently, the monthly sequences of R_H are disaggregated to daily scale using k-nearest neighbor (k-NN) method. The idea behind developing the two-stage methodology is that the monthly sequences of atmospheric variables simulated by the GCM are more reliable than those simulated at daily scale. Further, the SVM requires large computational effort to directly downscale daily sequences of large-scale atmospheric variables to daily sequences of hydrometeorological variables in a river basin.

Owing to the availability of a number of GCMs, predictors, scenarios, downscaling methods, etc., there is uncertainty in the future scenarios of R_H. Hence, there is a need to study the changes in R_H of a region by the various alternatives available.

The following objectives are set for this study: (1) To investigate whether the relationships between the predictor variables and R_H vary with season and location, (2) to investigate the performance of season/location-specific SVM downscaling models to identify the best model, (3) to study the uncertainties in downscaled R_H to predictors selected and Intergovernmental Panel on Climate Change Special Report on Emissions Scenarios (IPCC SRES), (4) to study projected trends of monthly R_H over the study region for IPCC SRES scenarios, and (5) to demonstrate the effectiveness of the developed disaggregation model over that based on triangular distribution method.

3 Study region

The catchment of Malaprabha reservoir in the Karnataka State of India was considered for the study. It has an area of 2,564 km^2 situated between 15°30′ N and 15°56′ N latitudes and 74°12′ E and 75°15′ E longitudes. The catchment lies between 792.5 and 480.0 m above mean sea level. The Malaprabha reservoir is a major source of water for the arid region of north Karnataka, which is possibly the largest arid region in India outside the Thar Desert. The reservoir supplies water for an irrigable area of 218,191 ha in north Karnataka. The location map of the study region is shown in Fig. 1.

4 Data

The observed records of R_H were available at daily time scale for a gauging station near the region. It is situated in Gadag District (administrative division) at 15°25′ N latitude and 75°38′ E longitude and has records for the period January 1978 to December 2000. The primary source of the data is the Water Resources Development Organization, Government of Karnataka State, India. One gauging station for downscaling was used for the entire basin since the distribution of relative humidity stations around the world is fewer in number (5,110) when compared to the number of stations measuring precipitation (27,075) and temperature (12,783) (New et al. 2002). A number of studies state that the principal cause of the relatively small number of published papers on air humidity is the lack of long-term

Table 1 Literature on predictor selection for statistical downscaling of R_H using transfer functions

SN	Predictor	Predictand and (time scale)	Data	Technique	Region	Reference
1	Surface and 925 hpa temperature, latent heat, specific humidity at 925 hpa.	Relative humidity (monthly)	NCEP-NCAR reanalysis data sets, CGCM3	Multi-step linear regression	India (river basin)	Anandhi (2011)
2	Daily sea level pressure	Maximum, minimum, and average daily temperature, precipitation amount and occurrence, relative humidity, cloudiness, and sunshine duration	ECMWF ERA-40 data set	Using 24 objective and 2 subjective atmospheric circulation classifications collected and developed within the COST733 Action	Czech (21 stations)	Cahynová and Huth (2010)
3	Air temperature and specific humidity at 1,000; 800; 600; and 400 hPa from NCEP-NCAR reanalysis	Observed air temperature and specific humidity (6-hourly)	NCEP-NCAR reanalysis data sets for March 2004 to May 2006	Least square regression	Peru (Cordillera Blanca)	Hofer et al. (2010)
4	Geopotential height at 500, 850, and 1,000 hpa; wind speed and vorticity at 500 and 850 hpa; temperature, relative humidity, and specific humidity at 850 hpa	Relative humidity, water vapor pressure, dew-point temperature, and dew point deficit (07, 14, and 21 h local time for summer months June to August)	NCEP-NCAR reanalysis data sets	Multi-step linear regression	Czech Republic (4 point stations	Huth (2005)
5	(a) Geopotential height at 500, 700, and 1,000 hpa; geopotential height thickness 500–700; 500–1,000; and 700–1,000 hpa; for classification	Relative humidity (daily)	Control Run 591 (C591) from ECHAM4 at T42 resolution	(a) Classification by cluster analysis to identify significant circulation patterns	Germany (52 weather stations grouped into 6 climate zones)	Enke and Spekat (1997)
	(b) geopotential height, geostrophical wind zonal component and meridional component, scalar wind, and vorticity at 500; 700; and 1,000 hpa; scalar wind thickness 500–700; 500–1,000; and 700–1,000 hpa, humidity at 700 hpa			(b) downscaling using multiple conditional (weather pattern dependent) screening regression analysis		

hygrometric or psychrometric measurements as well as the methodological difficulties related to the data analysis (Wypych 2010; Wang and Gaffen 2001; Gaffen and Ross 1999).

In India, the year has been divided into four seasons, namely winter (January, February), pre-monsoon or summer (March to May), southwest monsoon or summer monsoon season (June to September), and post-monsoon season (October to December) (Parthasarathy et al. 1994). The period from June to September is categorized in India as the Indian Summer Monsoon Rainfall season because rainfall in this season is caused by evaporation during summer. In summer, the weather is dry, and the humidity is low. The mean monthly daytime R_H in the basin during monsoon season ranges from 83% to 87%, while the same during post-monsoon (October to January) and summer (February to May) seasons ranges from 67% to 78% and 57% to 74%, respectively (Fig. 2).

The reanalysis data of the monthly mean atmospheric variables and latent heat flux in the study region were extracted from database prepared by the National Centers for Environmental Prediction (NCEP; Kalnay et al. 1996) for the period January 1978 to December 2000. The data of atmospheric variables were collated for nine grid points whose latitudes range from 12.5° N to 17.5° N and longitudes range from 72.5° E to 77.5° E at a spatial

Table 2 Literature on methods for disaggregation of R_H

SN	Method	Reference
1	Method based on actual and saturated vapor pressure	Debele et al. (2007)
2	Approximate regression quantile method	Green and Kozek (2003)
3	Triangular distribution method with adjustment for clear/overcast conditions	Neitsch et al. (2001)

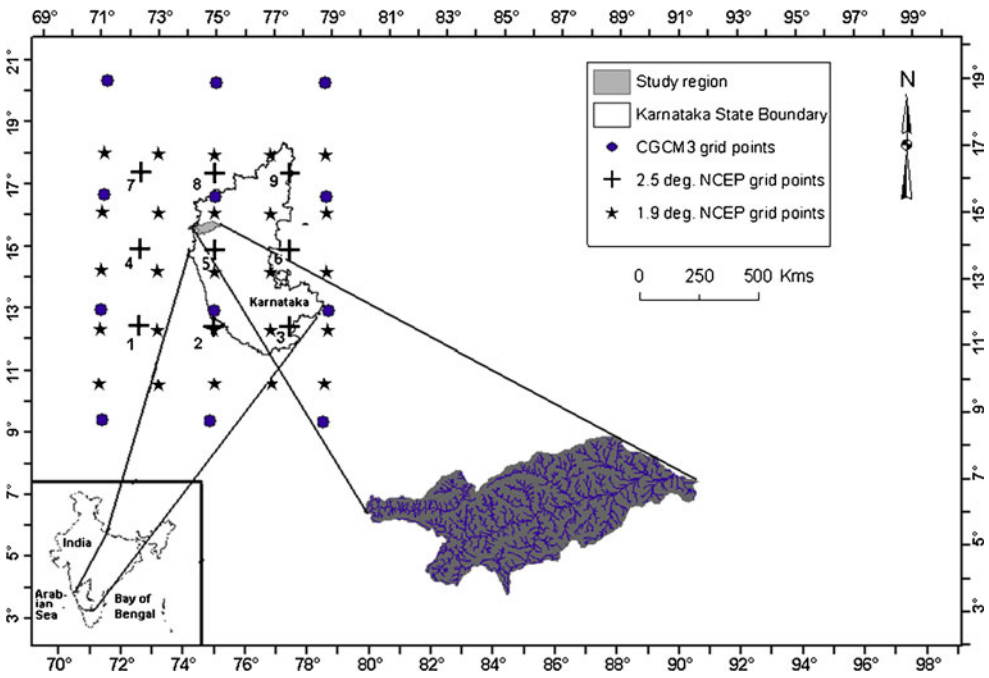

Fig. 1 Location of the study region in Karnataka State of India. The latitude, longitude, and scale of the map refer to the Karnataka State. The data are extracted at 12 CGCM3 grid points. Information on LH obtained from reanalysis data at 25 grid points spaced at 1.9° are re-gridded to the nine 2.5° NCEP grid points. Among the nine grid points, 1, 4, and 7 are on the Arabian Sea, and the remaining points are on land

resolution of 2.5°. The data of latent heat flux were extracted for 25 grid points whose latitudes range from 10.48° N to 18.1° N and longitudes range from 71.25° E to 78.75° E at a spatial resolution of approximately 1.9°.

For obtaining future projections of R_H, monthly simulations of the third-generation Canadian Coupled Global Climate Model (CGCM3) were considered for four scenarios (A1B, A2, B1, and COMMIT) that are relevant to IPCC's fourth assessment report (AR4) (Alley et al. 2007). The CGCM3 data were obtained (through website http://www.cccma.bc.ec.gc.ca/) for the period January 1978 to December 2100, for 12 grid points whose latitudes range from 9.28° N to 20.41° N and longitudes range from 71.25° E to 78.75° E. The spatial resolution of CGCM3 output is 3.75° along the longitude and approximately 3.75° along the latitude.

The GCM data and the information on atmospheric flux were re-gridded to a common 2.5° NCEP grid using Grid Analysis and Display System (GrADS; Doty and Kinter 1993).

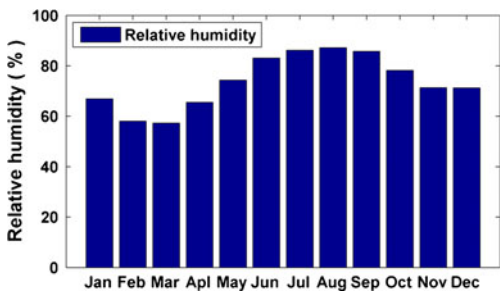

Fig. 2 Mean monthly daytime relative humidity in the study region for the period 1978–2000

5 Methods of analysis

This section presents the methodology proposed for the development of the SVM model for spatial downscaling of atmospheric variables (simulated by NCEP and CGCM3) to R_H in a river basin at monthly scale and for temporal disaggregation of the downscaled R_H from monthly to daily scale.

The relationships between predictors in NCEP and GCM data sets and predictor–predictand relationships were investigated using scatterplots and three measures of dependence, namely product moment correlation, Spearman's rank correlation, and Kendall's tau (Helsel and Hirsch 2002). The performance of the developed models was evaluated using three statistical measures, namely Nash–Sutcliffe error estimate (E_f), mean absolute error (MAE), and product moment correlation (P). The formulae and description of these measures can be found in Anandhi et al. (2009).

5.1 Development of downscaling model

The development of downscaling model consists of selection of probable and potential predictors, stratification of seasons and location, and finally training and validation of the downscaling model. The developed model is then used to obtain future projections of R_H from simulations of CGCM3.

5.1.1 Methodology for selection of probable predictors

The selection of appropriate predictors for downscaling a predictand is one of the most important steps in a

downscaling exercise (Anandhi et al. 2009; Teutschbein et al. 2011). The choice of predictors could vary with region, season, and the predictand to be downscaled (Kostopoulou et al. 2007; Wetterhall et al. 2007; Frei et al. 2005). Any type of variable or index can be used as predictor as long as it is reasonable to expect that there exists a relationship between the predictor and the predictand (Wetterhall et al. 2005). Often, in climate impact studies, the predictors chosen are as variables that are: (1) reliably simulated by GCMs and readily available from archives of GCM output and reanalysis data sets, (2) physically related with the predictand, (3) based on previous studies, and (4) carry climate change information (Wetterhall et al. 2009; Wilby et al. 1999; Hofer et al. 2010). The predictors selected have to satisfy one or more of these criteria. For example, a predictor with less dependence structure may be selected because it may carry climate change information (Wetterhall et al. 2009; Wilby et al. 1999), to account for changing variables that affect the local climate (Hofer et al. 2010).

In the present study, large-scale atmospheric variables, namely air temperature and specific humidity at 925 mb, surface air temperature, and latent heat flux were considered as the probable predictors. They are denoted by Ta 925, Hus 925, Ta sur, and LH, respectively. These are chosen as they have a physically meaningful relationship with the predictand (R_H). Humidity represents concentration of water vapor in the air. Temperature and humidity are associated with local thermodynamic stability and hence are useful as predictors. In the previous studies, temperature and specific humidity are considered as potential predictors to downscale humidity (Anandhi 2011; Enke and Spekat 1997; Huth 2005; Hofer et al. 2010). Changes in surface humidity are reported to be caused by, or at least linked to, changes in atmospheric circulation patterns. Further, the changes in the longwave patterns, dominant air mass types, or strength or position of climatological "centers of action" will have important influence on local humidity and temperature regimes (Gaffen and Ross 1999).

Temperature affects the moisture-holding capacity and the pressure at a location. The pressure gradient affects the circulation, which in turn affects the moisture brought into the place and hence the humidity. At 925 mb pressure height, the boundary layer (near surface effect) is important. LH indicates the amount of moisture going from the surface to the atmosphere. The amount of moisture held in the atmosphere is related to temperature through "Clausius–Clapeyron" equation. A number of studies have shown strong relationship between humidity and temperature (Dai 2006; Gaffen and Ross 1999; Wypych 2010), and the strength of the relationship is found to vary from region to region (Wypych 2010). Because atmospheric water vapor provides a strong positive feedback to greenhouse gas-induced global warming, a realistic humidity–temperature relationship is vital for climate models to correctly simulate future climate change (Dai 2006).

Scatterplots were prepared, and cross-correlations were computed between the predictor variables in NCEP and GCM data sets to verify if these variables are reliably simulated by GCM. Similar plots and statistics were computed between the NCEP predictor variables and the predictand to investigate the presence of nonlinearity/linearity in their relationship. The details are provided in Sect. 6.

5.1.2 Stratification based on seasons and location (land/sea)

The relationship of R_H with a predictor variable may be sensitive to the location of its grid point (whether on land or over sea) (Dai 2006) as well as the season. To verify this, season- and location-based stratifications of predictors were considered, and nine groups were formed. The groups include one for each season (wet and dry seasons), location (land and sea), and combinations of both (wet season and land, dry season and land, wet season and sea, dry season and sea). The ninth group constitutes the case with no predictor stratification. The stratification of climate of the study region into wet and dry seasons is based on precipitation. The seasons identified in Anandhi et al. (2008) for the four IPCC scenarios are adopted for this study to downscale R_H.

5.1.3 Development of SVM downscaling model

The procedure developed for downscaling R_H is explained briefly in this section. It is based on the ideas that were explained in detail in Anandhi et al. (2008). Separate SVM models were developed to downscale R_H from potential predictors pertaining to each of the aforementioned nine groups.

Nine NCEP grid points covering and surrounding the study region were selected as the spatial domain of the four probable predictors (m_1=4), to adequately cover the various circulation domains of the predictors. From the m_2 probable predictors (m_2=4×9; i.e., m_1×number of NCEP grid points considered), the m_3 potential predictors for downscaling R_H were selected by specifying two threshold values (T_{ng} and T_{np}). The T_{ng} is for cross-correlation between predictors in NCEP and GCM data sets, whereas T_{np} denotes the same between predictors in NCEP data set and R_H. From the NCEP data on potential predictors, the orthogonal principal components (PCs) which preserve more than 98% of the variance in the data were extracted,

and the corresponding principal directions were noted. The PCs of GCM predictor data were extracted along the principal directions of NCEP predictor data. The PCs account for most of the variance in the predictor data and also remove redundancy, if any, among the input data. The use of PCs in the analysis makes the model more stable and reduces the computational load. Feature vector for each month was formed using the PCs extracted for the month. These feature vectors, formed using the PCs extracted from NCEP data, were used for development (calibration and validation) of the SVM models. Different SVM models were developed for various combinations of stratification and predictor variables. The performance of the models developed was evaluated using the three statistical measures (E_f, MAE, and P), and the best SVM model was identified.

5.2 Development of the disaggregation model

The disaggregation problem considered in this paper entails the temporal disaggregation of R_H over the Malaprabha catchment from monthly to daily scale using two models. The first model (model-1) is based on the k-NN technique, and the second model (model-2) is based on triangular distribution method of Soil and Water Assessment Tool (SWAT; Neitsch et al. 2001).

5.2.1 Disaggregation model based on k-NN technique (model 1)

Statistical relationship is developed between the observed mean values of predictand at monthly and daily scales. The relationship is then used to disaggregate the projections of the predictand obtained at monthly scale using a downscaling model.

Let the historical (past) and projected (future) values of predictand be denoted by $u^h_{\upsilon h,\tau,j}$ and $u^p_{\upsilon p,\tau,j}$, respectively, where the subscripts υh and υp are indices for the past and future years (υh=1,…, N_h; υp=1,…, N_p), and τ denotes the index for the month within the year (τ=1, 2,…, t,…, ω), j refers to the index for the day within the month τ. N_h refers to the number of years of historical record (herein, N_h=23 for data from 1978 to 2000), N_p refers to the projected period (herein, N_p=100 for data from 2001 to 2100), and ω represents the number of months (=12) in a year. Further, let $u^h_{\upsilon h,\tau,\bullet}$ denote the monthly mean value of the predictand computed using the observed daily values of the predictand in the month τ of year υh. Similarly, let the monthly mean value projected for the predictand in the month τ of year υp be $u^p_{\upsilon p,\tau,\bullet}$.

$$u^h_{\upsilon h,\tau,\bullet} = \frac{\sum_{j=1}^{D_\tau} u^h_{\upsilon h,\tau,j}}{D_\tau} \qquad \tau = 1,\ldots,t,\ldots,\omega \quad \upsilon h = 1,\ldots,N_h \tag{1}$$

where D_τ denotes the number of days in month τ. For the calibration period (which could be less than or equal to the length of historical record), the observed value of the predictand on day j in month τ is expressed as a ratio of the monthly mean value of the predictand as:

$$w_{\upsilon h,\tau,j} = \frac{u^h_{\upsilon h,\tau,j}}{u^h_{\upsilon h,\tau,\bullet}} \quad j = 1,\ldots,D_\tau \quad \tau = 1,\ldots,\omega \quad \upsilon h = 1,\ldots,N_h \tag{2}$$

Let $\mathbf{W}^h_{\upsilon h,\tau} = \{w_{\upsilon h,\tau,1},\ldots,w_{\upsilon h,\tau,D_\tau}\}$ denote the vector containing the ratios of the daily values of predictand in month τ of year υh. The following are the key steps of the proposed algorithm to generate daily values of the predictand $u^p_{\upsilon p,\tau,j}$ for the projected period 2001–2100.

1. For every projected value of the predictand, identify the calendar month τ.
2. Form the conditioning set z_τ for each month τ. It comprises the observed monthly values of the predictand for the calendar months falling in the season window of size M centered on the month τ. For a window of size 1 month, the conditioning set is $\mathbf{z}_\tau = \{u^h_{1,\tau,\bullet},\ldots,u^h_{\upsilon h,\tau,\bullet},\ldots,u^h_{N_h,\tau,\bullet}\}$. The window size depends on the projected trend of the downscaled variable at monthly scale. A window of size 1 month could be considered if no change in trend is projected by the SVM model. In the presence of trend, a larger size window (e.g., 3 months) is recommended to explore the temporal relationships between a wide range of monthly and daily values of the predictand for use in disaggregation.
3. To disaggregate the value of predictand in month τ, $u^p_{\upsilon p,\tau,\bullet}$, select its k-nearest neighbors from the conditioning set $\mathbf{Z}_\tau$ based on the Euclidean distance $\xi_{\upsilon h,\upsilon p}$ between $u^p_{\upsilon p,\tau,\bullet}$ and $u^h_{\upsilon h,\tau,\bullet}$ expressed as
$$\xi_{\upsilon h,\upsilon p} = \left\| u^p_{\upsilon p,\tau,\bullet} - u^h_{\upsilon h,\tau,\bullet} \right\|, \text{ for } \upsilon h = 1,\ldots,N_h \tag{3}$$
$\|\ \|$ represents the Euclidean norm. The number of neighbors k is a smoothing parameter. Lall and Sharma (1996) suggest using k equal to $\sqrt{N}$, as a rule of thumb. The k-nearest neighbors considered are to be the most similar to the projected value of the predictand for the month τ.
4. The conditional probability $p(i)$ is determined based on the k-nearest neighbors. Assign weights to each of the k-nearest neighbors using the discrete probability mass function $p(i)$ (Lall and Sharma 1996). This discrete kernel was developed through a Poisson approximation of the probability distribution function

of state space neighbors. Randomly select a nearest neighbor to the projected value of the predictand by constructing cumulative density function using $p(i)$ values.

$$p(i) = \frac{1/i}{\sum_{j=1}^{k}(1/j)} \qquad i = 1, \ldots, k \tag{4}$$

It is to be noted that $p(i)$ is the same for all the months in the projected period. Let $u^{h}_{\upsilon h,\tau,\bullet}$ denote the nearest neighbor.

5. The projected daily values of the predictand for the month τ in year υp are obtained by multiplying the projected monthly predictand value $u^{p}_{\upsilon p,\tau,\bullet}$ with $\mathbf{W}^{h}_{\upsilon h,\tau}$ corresponding to the nearest neighbor $u^{h}_{\upsilon h,\tau,\bullet}$.

5.2.2 Disaggregation model based on triangular distribution (model 2)

The model based on triangular distribution method generates daily relative humidity values from monthly relative humidity values. This model requires four inputs: monthly relative humidity and a random number between 0.0 and 1.0 (inputs given by user), maximum and minimum relative humidity values allowed in a month (estimated from mean monthly relative humidity). The relevant equations are as follows.

$$\begin{aligned}&\text{if } \mathrm{rand}(0,1) \le \left(\frac{R_{\mathrm{hmon}} - R_{\mathrm{hLmon}}}{R_{\mathrm{hUmon}} - R_{\mathrm{hLmon}}}\right) \text{then}\\ &R_{\mathrm{h}} = R_{\mathrm{hLmon}} + \left[\mathrm{rand}(0,1) \times (R_{\mathrm{hUmon}} - R_{\mathrm{hLmon}}) \times (R_{\mathrm{hmon}} - R_{\mathrm{hLmon}})\right]^{0.5}\end{aligned} \tag{5}$$

$$\begin{aligned}&\text{else if } \mathrm{rand}(0,1) \ge \left(\frac{R_{\mathrm{hmon}} - R_{\mathrm{hLmon}}}{R_{\mathrm{hUmon}} - R_{\mathrm{hLmon}}}\right) \text{then}\\ &R_{\mathrm{h}} = R_{\mathrm{hUmon}} - (R_{\mathrm{hUmon}} - R_{\mathrm{hmon}})\\ &\quad \times \left[\frac{R_{\mathrm{hUmon}} \times \langle 1 - \mathrm{rand}(0,1)\rangle - R_{\mathrm{hLmon}} \times \langle 1 - \mathrm{rand}(0,1)\rangle}{R_{\mathrm{hUmon}} - R_{\mathrm{hmon}}}\right]^{0.5}\end{aligned} \tag{6}$$

$$R_{\mathrm{hUmon}} = R_{\mathrm{hmon}} + (1 - R_{\mathrm{hmon}}) \times \exp(R_{\mathrm{hmon}} - 1) \tag{7}$$

$$R_{\mathrm{hLmon}} = R_{\mathrm{hmon}} \times (1 - \exp(-R_{\mathrm{hmon}})) \tag{8}$$

where R_{hUmon} and R_{hLmon} are the largest and the smallest relative humidity values that were recorded in a month. R_{h} and R_{hmon} are the daily and monthly R_{H}, respectively and rand(0, 1) is a random number between 0.0 and 1.0.

6 Results and discussion

Downscaling and disaggregation models were developed following the methodology described in Sect. 5. The results are presented and discussed in this section.

6.1 Probable predictors

Reanalysis data set is often held as an example of the best possible historical GCM output (Reichler and Kim 2008; Maurer et al. 2010). Also the coherence, accessibility, and completeness of the NCEP-National Center for Atmospheric Research (NCAR) data sets make them attractive for comparative analyses of GCM performance (Schoof and Pryor 2003). Hence, in this study, the reliability of the GCM in simulating the chosen predictor variables is examined by comparing the simulations of GCM with the NCEP/NCAR reanalysis data sets using correlation and scatterplots (Figs. 3 and 4). The reliability of the variables in the reanalysis data sets is classified into four classes (A to D), depending on the relative influence of the observational data and the model on the variable. As per these classifications, among the predictors chosen for this study, temperature at 925 mb is in the most reliable class (class A), surface temperature and humidity are in class B, while latent heat flux is in the less reliable class (class C).

Scatterplots and cross-correlation bar plots were prepared to verify the reliability of the simulations of the predictor variables by the GCM and to study the predictor–predictand relationships (Figs. 5 and 6). Results of the analysis show that the correlation between predictor variables in NCEP/NCAR and GCM simulations decreases from predictors in class A to class C. Among the predictors, temperature at 925 mb (class A variable) has highest correlation, and latent heat flux (class C variable) has the least correlation, while the correlations of surface temperature and specific humidity at 925 mb (class B variables) are in between (reasonably large). Since caution should be exercised in interpreting results of reanalysis especially for categories B and C (Kalnay et al. 1996), it is speculated that low correlation in predictors need not indicate less reliability in GCM simulation as this could be due to the contribution of uncertainty in simulations from reanalysis. In the past, downscaling approaches have used some predictors of class C (Hofer et al. 2010; Widmann et al. 2003; Schmidli et al. 2006; Doblas-Reyes et al. 2006). In an earlier study, Bader (2004) evaluated ten coupled models by comparing the simulations of specific humidity with NCEP-NCAR and ERA15 reanalysis data sets. They found that the accuracy of specific humidity estimates decreases substantially with height, and at 850 hPa, the mean over the ten models agrees better with NCEP-NCAR reanalysis than it does with ERA15. In this study, this specific humidity at 925 mb (Hus 925) is used as one of the predictors.

Further, from the plots between the probable predictor variables in NCEP data and the predictand (Figs. 5 and 6), the dependence structure between the predictor and the predictand (R_{H}) can be observed. Ta sur and Ta 925 are negatively correlated with the predictand (Fig. 6), which is

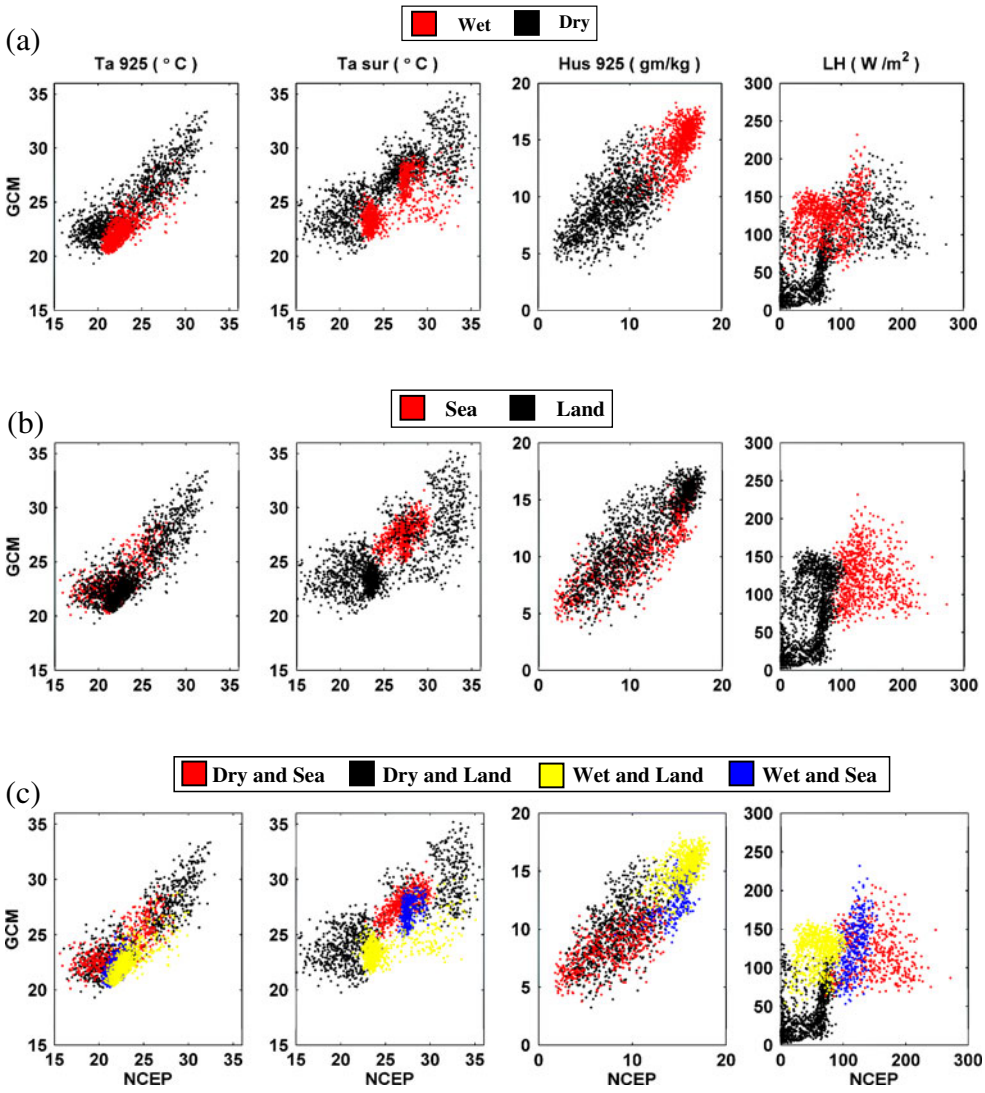

Fig. 3 Scatterplots prepared to investigate dependence structure between probable predictor variables in NCEP and GCM data sets for the period 1978–2000. **a**, **b,** and **c** denote plots for predictors stratified based on season, location, and their combinations, respectively. In each plot, the ordinate denotes the GCM value of the predictor variable, whereas the abscissa represents the NCEP value of the predictor variable

related to the fact that R_H is the ratio of the actual moisture content in the air to the maximum moisture content it could hold when saturated. If the moisture in the air remains unchanged and the temperature rises, the maximum amount of moisture that the air could hold increases, indicating a drop in R_H. Thus, the R_H is low when temperature of earth's

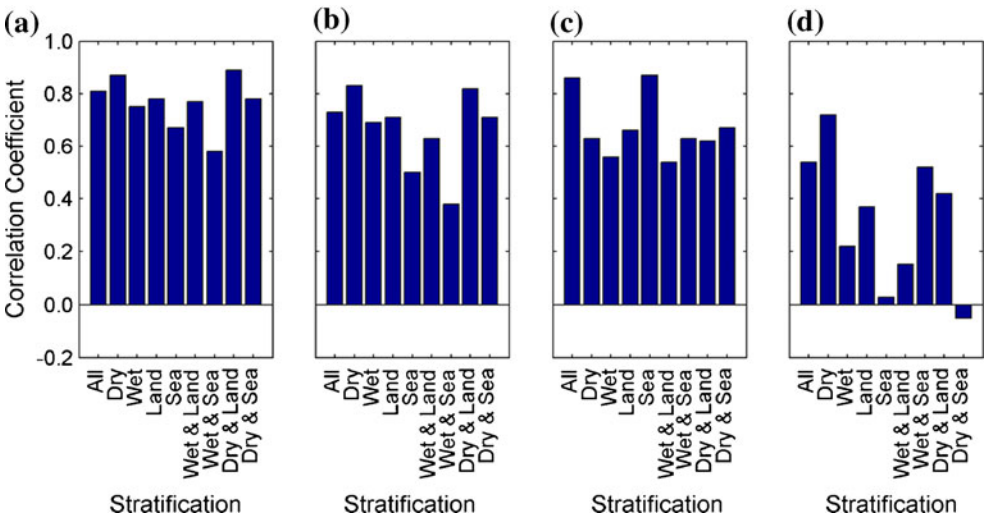

Fig. 4 Bar plots showing cross-correlations computed between each of the four probable predictor variables in NCEP and GCM data sets, for each of the nine stratifications considered for the period 1978–2000. **a**, **b**, **c,** and **d** denote plots for each of the four probable predictor variables, namely temperature at 925 mb, temperature at earth's surface, specific humidity at 925 mb, and latent heat flux, respectively. Cross-correlations computed using product moment correlation

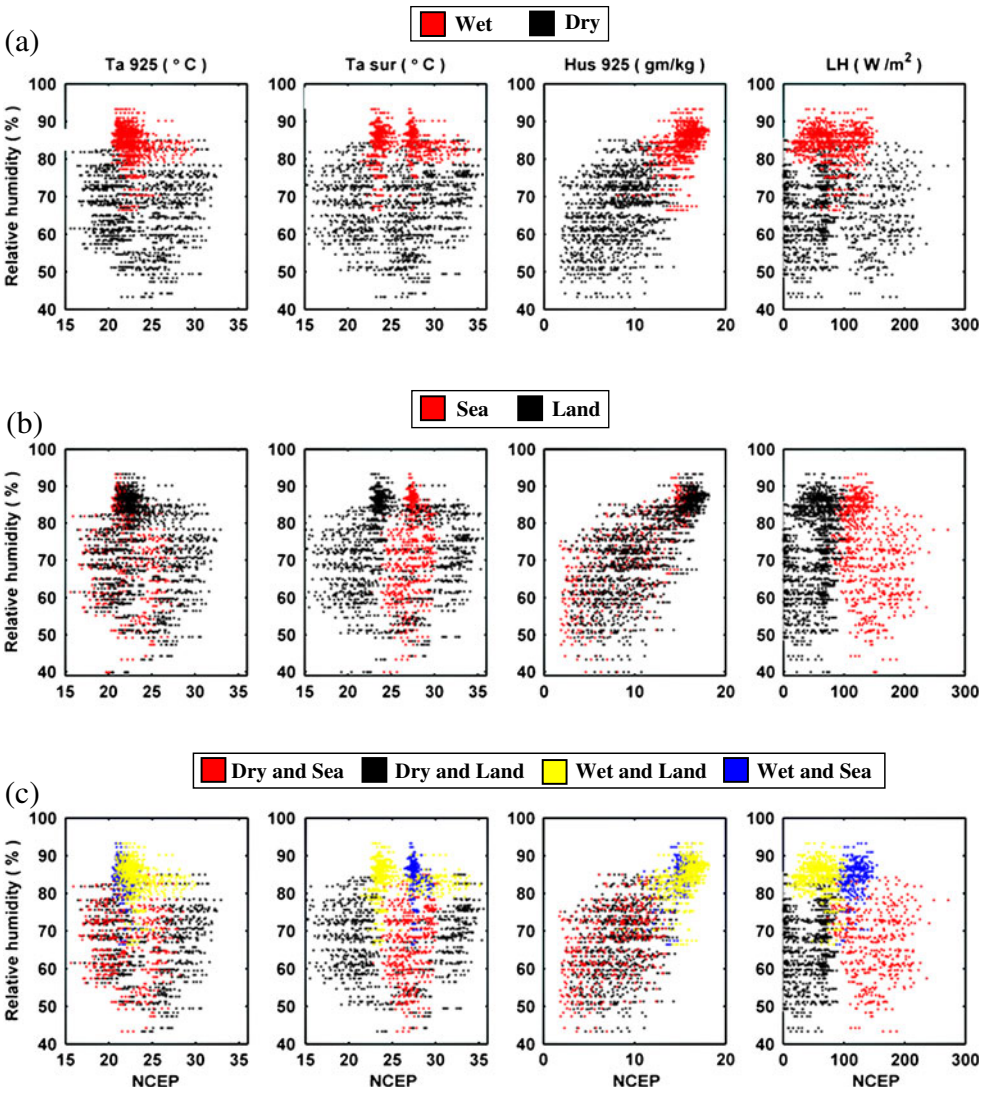

Fig. 5 Scatterplots prepared to investigate dependence structure between probable predictor variables in NCEP data and predictand for the period 1978–2000. **a**, **b,** and **c** denote plots for predictors stratified based on season, location, and their combinations, respectively. In each plot, the ordinate denotes the predictand whereas the abscissa represents the NCEP value of the predictor variable

surface is high. Similar negative correlation was also observed in selected sites in Australia (Jones et al. 2010). Further, the surface evaporation over the arid/semi-arid regions is limited by soil wetness, and it often cannot meet atmospheric demand to maintain a constant R_H as air temperature increases (Dai 2006). The results of this study

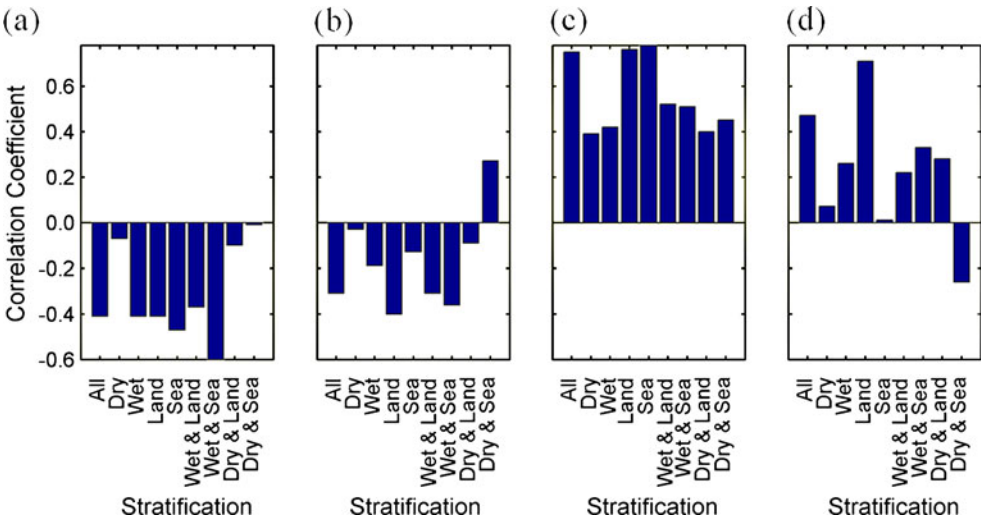

Fig. 6 Bar plots showing cross-correlations computed between each of the four probable predictor variables in NCEP data and the predictand, for each of the nine stratifications considered for the period 1978–2000. **a**, **b**, **c,** and **d** denote plots for each of the four probable predictor variables, namely temperature at 925 mb, temperature at earth's surface, specific humidity at 925 mb, and latent heat flux, respectively. Cross-correlations computed using product moment correlation

do not show a very strong correlation between R_H and temperatures, and the correlation is found to further decrease at annual scales. These are consistent with the results shown in Dai (2006). However, the results in their work show an insignificant positive correlation (0 to 0.2) between R_H and the temperatures of areas in and around the study region. It can be safely speculated that the possible difference in the direction of correlation could be due to the heterogeneities within the grid cell used in their study. The remaining two predictors, Hus 925 and LH, are positively correlated with the predictand (Fig. 6). The correlation coefficient between Hus 925 and R_H is strong (>0.6) as expected because any increases in Hus (e.g., due to increased soil wetness from precipitation) should raise the local R_H (Dai 2006). In this study, the predictors such as latent heat flux and temperature at the earth's surface have less correlation with the predictand for all the stratifications except latent heat in wet season stratification. However, these were selected as predictors since they satisfy criteria two and four. For example, latent heat flux helps to control surface temperatures, with important implications for regional climate characteristics such as the intensity and duration of heat waves (Jung et al. 2010).

6.2 Stratification

For each of the nine groups of predictors described in Subsect. 5.1.2., scatter plots and cross-correlation bar plots were prepared to verify the reliability of the simulation of the predictor variables by the GCM (Figs. 3 and 4) and to study the predictor–predictand relationships (Figs. 5 and 6). Predictors ranked based on product moment correlation as well as rank correlations (Spearman's and Kendall) lead to the same conclusion. Henceforth, values of product moment correlation only are mentioned, for brevity.

From a perusal of the scatterplots and correlation bar plots between the probable predictor variables in NCEP and GCM data sets (Figs. 3 and 4), it can be seen that the ability of the GCM to simulate the predictor variables varies with the predictor, season, and location. Ta 925 and Ta sur in NCEP data are highly correlated with those in GCM data for the grid points on land during dry season (correlations are 0.89 and 0.83, respectively), and the same are less correlated at the grid points on sea during wet season (correlations are 0.58 and 0.38, respectively). These results are consistent with the results in Dai (2006). Further, Hus 925 in NCEP and GCM data sets is highly correlated for grid points on sea (correlation is 0.87) and is less correlated for the grid points on land during wet season (correlation is 0.54). LH in the data sets is highly correlated during dry season (correlation is 0.72) than in wet season, and it is less correlated for the grid points on sea during dry season (correlation is −0.03). The variables that are highly correlated between NCEP and GCM data sets are considered to be realistically simulated by GCM.

From the above results, it can be inferred that there is a high variability in (1) the ability of the GCM to simulate the predictor variables with location and season and (2) dependence structure between the predictor variables and R_H for the various stratifications considered in this study. Hence, seasonal and location-based stratifications are necessary for selecting the potential predictors for downscaling.

6.3 Selection of potential predictors for downscaling

The potential predictor variables were identified from the probable predictors by analyzing the scatterplots and the cross-correlation bar plots prepared for each of the nine predictor groups at each of the nine NCEP grid points. Typical plots prepared for one scenario (where no stratification is considered) are shown in Figs. 7, 8, and 9. Results indicate that the selected potential predictors depend on the chosen predictor group and the thresholds given for correlations between the predictors in NCEP and GCM data (T_{ng}) and between the predictors in NCEP data and the predictand (T_{np}).

Scatterplots in Fig. 7 were prepared to investigate the relationship between the probable predictor variables in NCEP and GCM data sets for the various predictor groups considered in this study, while the scatterplots in Fig. 8 were prepared to investigate the relationship between the probable predictor variables in NCEP data and the predictand data.

The potential predictors selected for a threshold of 0.6 are presented in Table 3 for brevity. Only Hus 925 is selected as a potential predictor for dry season, from the predictor groups pertaining to land and/or sea grid points. In addition to Hus 925, Ta 925 and Ta sur are selected as potential predictors from the predictor groups pertaining to wet season, and wet season and land.

The scatterplots and correlation bar plots between the probable predictor variables in NCEP data sets and the predictand (Figs. 3 and 6) show that the linear dependence structure between each of the predictor variables and the predictand varies with the stratification.

In general, the choice of thresholds T_{ng} and T_{np} is subjective and can theoretically vary from 0 to 1. In this study, the threshold is varied between 0.01 and 0.84 for scrutinizing its effect on the selection of potential predictors. A low threshold (say 0.01) resulted in selecting all the predictors, which include variables that are not realistically simulated by GCM and have less dependence structure with the predictand. For example, the LH with stratifications such as wet, sea, wet and sea, dry and sea are not realistically simulated by GCM. These predictors have less

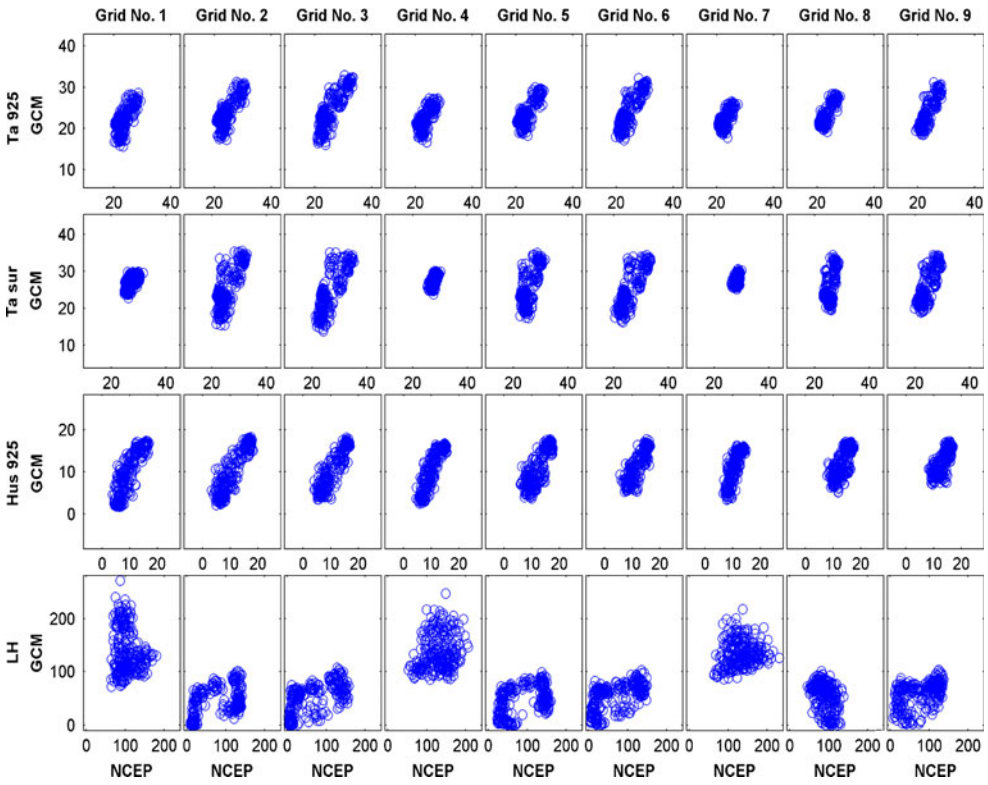

Fig. 7 Scatterplots prepared to investigate dependence structure between predictor variables in NCEP and GCM data sets, for selecting potential predictors for the period 1978–2000. The ordinate denotes the GCM value of the predictor variable, whereas the abscissa represents the NCEP value of the same

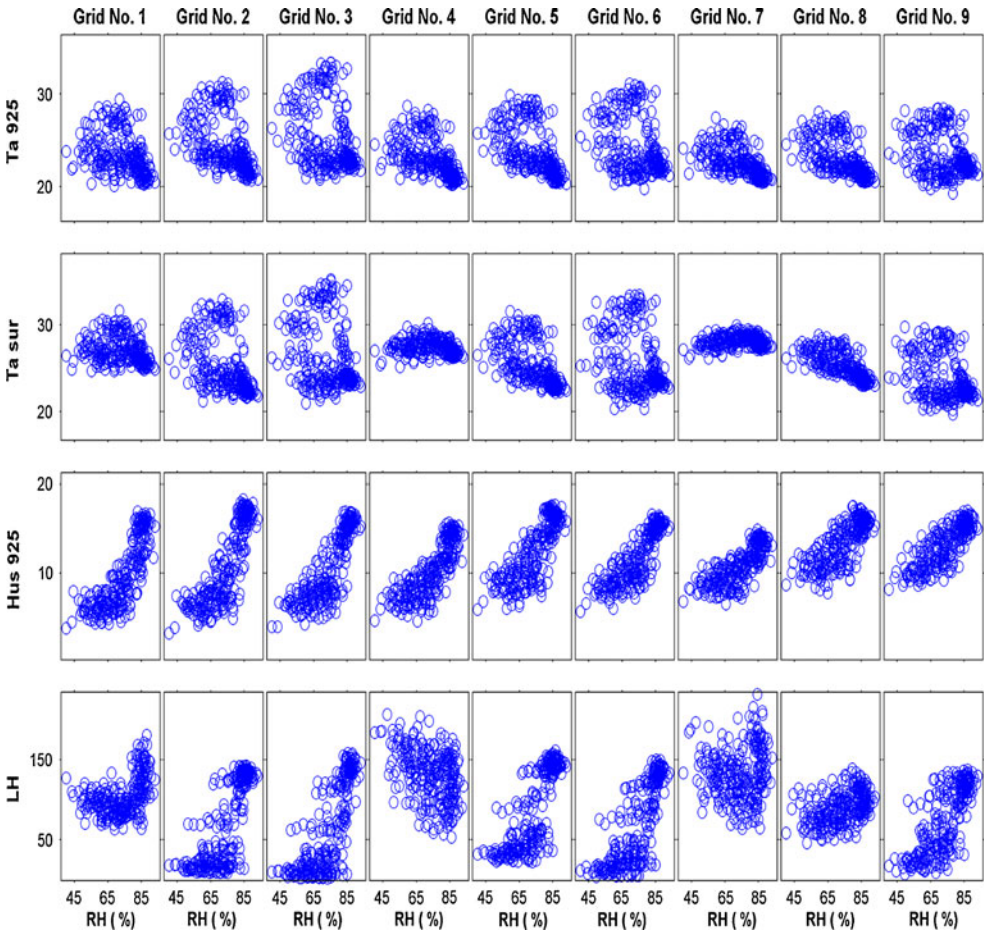

Fig. 8 Scatterplots prepared to investigate dependence structure between predictor variables in NCEP and predictand, for selecting potential predictors for the period 1978–2000. The value of the NCEP predictor variable is plotted as ordinate, whereas the value of the predictand is plotted as abscissa

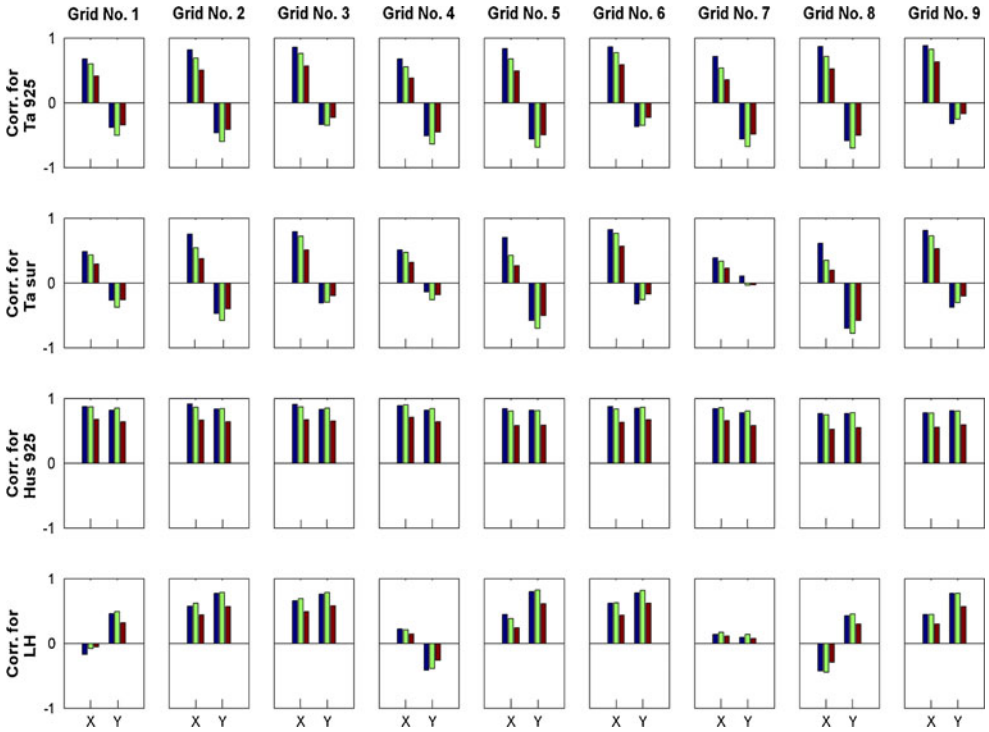

Fig. 9 Correlation bar plots prepared for selecting potential predictors for the period 1978–2000. *X* denotes correlation between predictor variables in NCEP and GCM data sets. The correlation between predictor variables in NCEP and the predictand is represented as *Y*. The *blue*, *green*, and *red bars* in *X* and *Y* represent product moment correlation, Spearman's rank correlation, and Kendall's tau, respectively

correlation with the predictand. In contrast, a high threshold resulted in selecting only a few of the probable predictors as the potential predictors, for example, Hus 925. In general, Hus 925 is found to be highly correlated with the predictand, and it is followed by Ta 925 and Ta sur. LH is found to be the least correlated potential predictor. For a threshold of 0.8, Hus 925 over grid points 1 to 6 is selected as the potential predictor. Among these, points 1 and 4 are on the ocean, whereas points 2, 3, 5, and 6 are on land.

From the foregoing discussion (Sect. 6.1 to 6.3), it can be inferred that the relationship between the predictor variables and the R_H varies with season and location. The seasonal cycle is brought out well by the models, which gives confidence in the model ability to simulate the large-

Table 3 List of potential predictors selected for developing downscaling models for each of the nine groups of predictors, for a threshold of 0.6 between probable predictors in NCEP data and the predictand

Group number	Stratification type	Potential predictors selected	
		Names	NCEP grid points
1	No stratification	Ta_sur	8
		Hus_925	1, 2, 3, 4, 5, 6, 7, 8, 9
		LH	3, 6
2	Dry season	Hus_925	1, 2, 3, 4, 5, 6, 7, 8, 9
3	Wet season	Ta_925	1, 2, 3, 4, 5, 6, 7, 8, 9
		Ta_sur	1, 2, 3, 4, 5, 6, 7, 8, 9
		Hus_925	1, 2, 3, 4, 6, 7
4	Land location	Hus_925	2, 3, 5, 6, 8, 9
		LH	6, 9
5	Sea location	Hus_925	1, 4, 7
6	Wet season and land	Ta_925	2, 3, 5, 6, 8, 9
		Ta_sur	2, 3, 5, 6
		Hus_925	2, 3, 6
7	Wet season and sea	Ta_925	1, 4, 7
		Hus_925	4, 7
8	Dry season and land	Hus_925	2, 6, 9
9	Dry season and sea	Hus_925	1, 4

The numbers shown for NCEP grid points correspond to Fig. 1

scale features. From among the nine groups of predictors considered, for a threshold, the number of feature vectors extracted may be inadequate to discern any relationship for the three stratifications namely "dry season and land," "wet season and sea," and "dry season and sea." Further, predictors based on sea stratification alone cannot be effective, as the study region is far from sea. A downscaling model was not developed for "wet and land" stratification, since developing a model for only "wet and land" stratification will provide downscaled results for wet period alone. There are no values for the dry and land, as a separate downscaling model was not developed for reasons mentioned earlier. Hence, SVM models have been developed to downscale R_H from each of the four stratification groups (wet season, dry season, land, and no stratification). This addresses objective 1 mentioned in Sect. 2.

6.4 Development of SVM downscaling models

Nine SVM downscaling models were developed for various combinations of stratification groups and predictor variables chosen for downscaling based on thresholds. Set (1) consists of all the four predictors (Hus 925, Ta 925, Ta sur, and LH) for land grid points, set (2) has three predictors (Hus 925, Ta 925, and Ta sur) for land grid points, set (3) comprises Ta 925 and Ta sur for land grid points, whereas set (4) has only Hus 925 for land grid points. Sets (5) to (8) have the same four (Hus 925, Ta 925, Ta sur, and LH), three (Hus 925, Ta 925, Ta sur), two (Ta 925, Ta sur), and one (Hus 925) numbers of predictors, but there is no stratification, and all the nine grids were considered. Set (9) comprises predictors which resulted in the best SVM downscaled model from among those developed using the nine predictor groups listed in Subsect. 5.1.2. The first set was chosen to study the overall effect of using all the predictors, the second set enables studying the effect of removing the predictor LH that is least correlated between NCEP and GCM data sets, while the third set was useful to investigate the effect of using only predictors that are negatively correlated with the predictand, and finally, the fourth set was considered to assess the impact of using only the most realistically simulated predictor as input to SVM. Sets (1) to (9) were analyzed to study the uncertainty due to stratification.

The optimal ranges of SVM parameters, namely kernel width (σ) and penalty term (C), are obtained using the grid search procedure. Typical results of the domain search performed to estimate the optimal ranges of the parameters are shown in Fig. 10. From this figure, the range of σ and C having the least Normalized Mean Square Error (NMSE) is selected as the optimal parameter range. The NMSE values are indicated in the barcode provided close to the figure. The optimal value of each parameter is obtained from its optimum range using genetic algorithm.

The performance of the nine SVM models (one for each set) was evaluated using three statistical measures, and results are shown in Fig. 11. From the figure, it can be observed that among the models considered, the performance of the model developed using potential predictors pertaining to land grid points (set 9) is better. The optimal values of σ and C parameters for the best SVM model are 129 and 511, respectively. Results of this model for the validation period (January 1994 to December 2000) are shown in Fig. 12. This addresses objective 2 listed in Sect. 2.

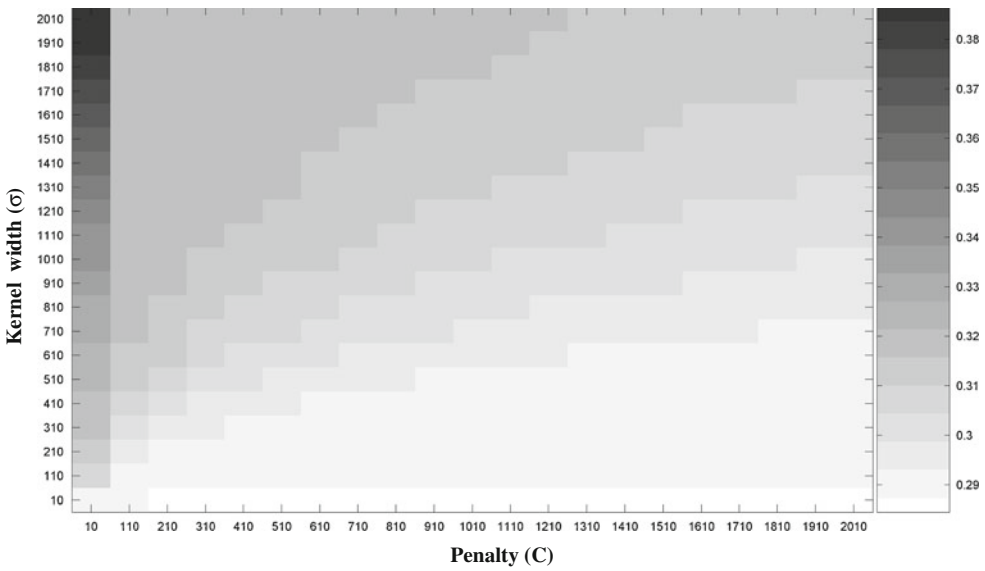

Fig. 10 Typical results of the domain search to determine optimal ranges of the parameters (kernel width, σ; penalty, C) for the SVM models developed in the study. The *bar code* indicates the NMSE values. The parameters giving the least NMSE values are estimated from the optimal ranges using genetic algorithm

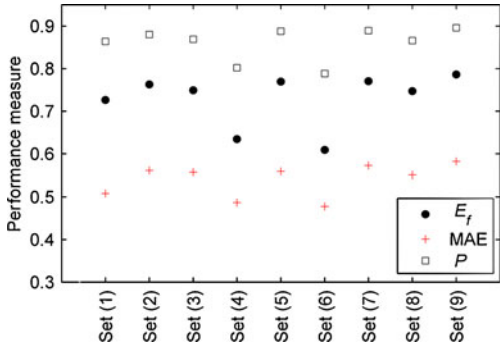

Fig. 11 Uncertainty of SVM models developed for each of the nine sets of predictors and stratifications mentioned in Sect. 6.4. Downscaled R_H values are compared with observed R_H values for the period 1978 to 2000, to estimate the three performance measures (E_f, MAE, and P). Maximum value of E_f, MAE, and P indicates better performance

6.5 Uncertainty in predictors selected

To address objective 3, the sensitivity of the SVM outputs to variations in their inputs has been examined using the nine sets of predictors.

The uncertainty in the monthly sequences of R_H generated by the nine SVM downscaling models is represented in two ways, (1) using three performance measures (*Ef*, MAE, and *P*) (Fig. 11) and (2) using cumulative distribution functions (CDFs) (Fig. 13).

In Fig. 13 a, the uncertainty of R_H downscaled for the period January 1978–December 2000 using predictors in NCEP data sets is brought out for the study region. The uncertainties of R_H obtained using GCM simulations of predictors are shown in part b of the figure.

For different threshold values (T_{ng} and T_{np}) mentioned in Subsect. 5.1.3, the predictors selected for downscaling varied. A SVM downscaling model was developed with the selected predictors for a threshold. The performance of all the developed SVM models was compared. The results indicate that the performance of the developed SVM models was good with thresholds greater than 0.3. However, the performance decreased considerably with a drop in the threshold below 0.3. This is because predictors with low dependence get selected as potential predictor for low threshold, which affects the model performance. The LH is not realistically simulated by GCM. Further, the NCEP reanalysis gridded field of LH is highly model dependent (Kalnay et al. 1996). The use of Ta 925 and Ta sur as predictors did not improve the model performance but affected the projected trend of the predictand, which will be discussed in detail in the next subsection. Among all the developed models, the model based on single predictor of set (4) showed better performance than the model developed using the four predictors in set (1). But its performance was inferior to the model developed using three predictors in set (2) which does not contain LH. This could be because only one predictor chosen for set (4) may be insufficient to represent the thermodynamics and dynamics of circulation in the region. On the other hand, selecting all the predictors creates noise, which affects the performance of the developed model.

6.6 Assessment of trend in the downscaled predictand

The statistical significance of the trend in the projected R_H scenarios is assessed using null hypothesis considering 99% confidence level. This addresses objective 4 set for the study. For the test, it is assumed that the variances of past and projected R_H are unknown and unequal. In the null hypothesis, the mean R_H for the past (1978–2000) and the projected future periods (2001–2100) are assumed to be equal. The test statistic, T (Kottegoda and Rosso 1998), is computed for the various IPCC scenarios for the nine sets.

From Fig. 14, contradictory trends (positive, negative, and neutral) can be observed for SRES A2, followed by A1B, B1, and the least for COMMIT scenarios. Results show that selecting predictors with increasing trend and with positive correlation with R_H produces a significant increasing trend in R_H, while selecting predictors with increasing trend and with negative correlation with R_H produces a significant decreasing trend in R_H. For example,

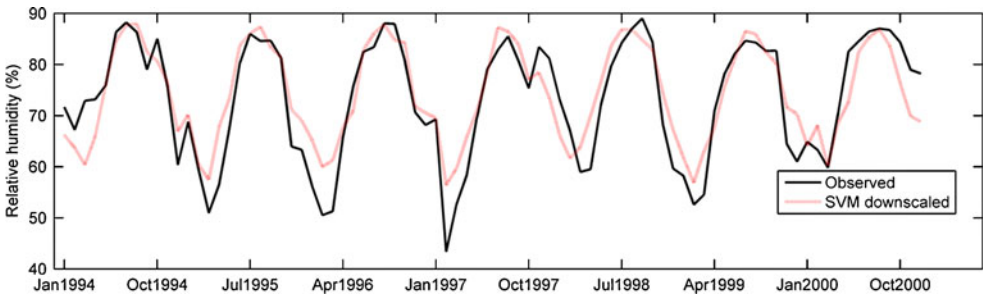

Fig. 12 Typical results from the best SVM downscaling model for the validation period (January 1994 to December 2000). Observed sequence of monthly relative humidity (R_H) is compared with R_H downscaled using each of the models. The best SVM model downscaled R_H from potential predictors pertaining to land location in NCEP data

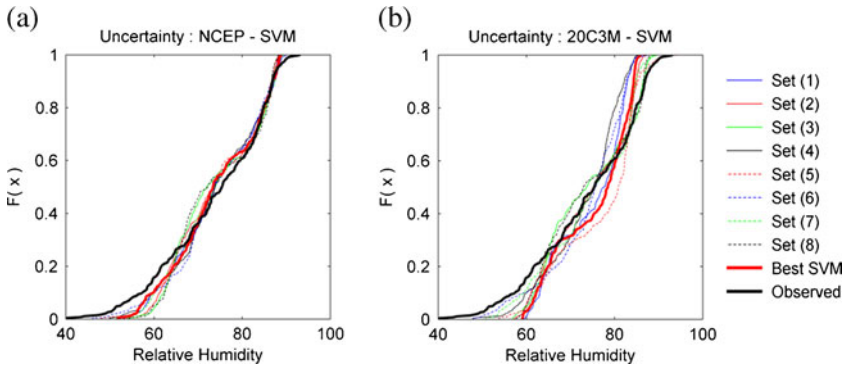

Fig. 13 The CDFs showing the uncertainty in the R_H downscaled using SVM method for nine sets of predictor and stratification combinations for the period 1978–2000

the projections for the predictand based only on the predictors Ta 925 and Ta sur showed a significant negative trend for A1B, A2, and B1 scenarios, whereas no trend is discerned with COMMIT. This is due to the fact that temperature is projected to increase in the future for A1B, A2, and B1 scenarios (highest for A2 and least for B1), and it is negatively correlated to R_H. Hence, it can be inferred that the trend in the downscaled predictand depended on the trend of predictors (positive, negative, or no trend), the predictor–predictand correlation (positive, negative, or no correlation), and the noise inherent in the selected predictors. The conflict in trends is more pronounced for the A2 scenario as the predictors have highest increasing trends and predictor–predictand correlation.

From the significance tests, it can be observed that in the overall context, for the best SVM model, the R_H is projected to increase in the future for A1B and A2 scenarios, whereas no trend is discerned for B1 and COMMIT scenarios. This increase in R_H, temperature (Anandhi et al. 2009), and cloud cover (Anandhi 2008), observed at a smaller scale in this basin in India, is in agreement with the results in Dai (2006). Dai (2006) also showed an increase in R_H over India during the period 1976–2004 and attributed it to the increase in the surface specific humidity coupled with moderate warming and increase in low clouds over these regions. The study by O'Gorman and Muller (2010) also showed a slight increase in multi-model mean R_H for this region for A1B emissions scenario during the period 2080–2099. They used outputs from 12 GCMs participating in the IPCC AR4 report. Simmons et al. (2010) showed a small reduction in relative humidity over land/Asia over a period of 10 years ending with 2008, based on monthly anomalies in the surface air temperature and humidity from comprehensive European Centre for Medium-Range Weather Forecasts reanalyses (ERA-40 and ERA-Interim) and from Climatic Research Unit and Hadley Centre analyses of monthly station temperature data (CRUTEM3) and the synoptic humidity observations (Hadley Centre and Climate Research Unit global surface humidity, HadCRUH). They point to faster warming over the land relative to the oceans as a causal mechanism for the decreasing RH from 2001 to 2008. Willett et al. (2010) showed an increase in humidity in the Asian region when observed changes in the HadCRUH global land surface specific humidity and CRUTEM3 surface temperature from 1973 to 1999 are compared with CMIP3 archive climate model simulations with 20th century forcings. They found that the variability over Southern Asia is biased high in the majority of models, especially in June–July–August season.

The uncertainty in the future projections of R_H was discerned from the CDFs shown in Fig. 14 for the various combinations of scenarios, predictor sets, and downscaling methods.

6.7 Development of disaggregation models

The monthly values of R_H obtained from the SVM downscaling model are disaggregated to daily values using each of the two disaggregation models (models 1 and 2) to

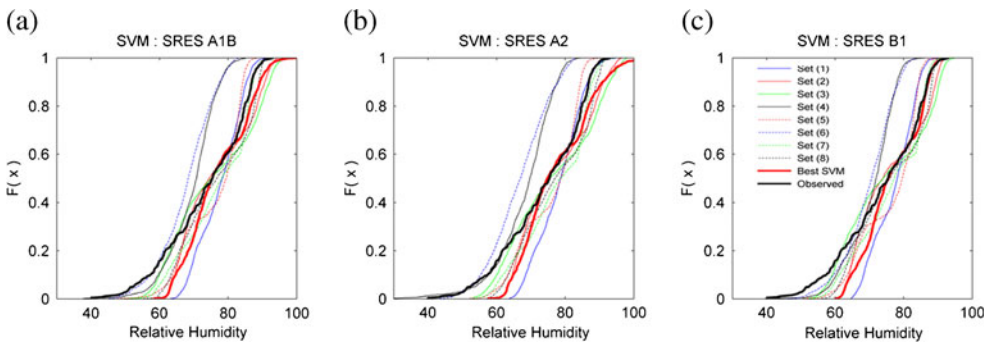

Fig. 14 Uncertainties in downscaled future scenarios of R_H to IPCC SRES scenarios (A1B, A2, B1, and COMMIT) and predictors selected (nine sets of predictors and stratifications). In each of the plots, the *black bold line* represents the CDF obtained using daily observed data for the study region for the period 2001–2100

Table 4 Error statistics computed for daily sequences of relative humidity generated by the two disaggregation models for the validation period

Performance measure	Model 1	Model 2
E_f	0.20	−0.39
MAE	0.21	−0.18
P	0.60	0.57

Maximum values of E_f, MAE, and P indicate better performance

identify the better disaggregation model. The results presented in Table 4 show that the disaggregation model based on *k*-NN technique (model 1) performs better than the disaggregation model based on the triangular distribution method of SWAT (Neitsch et al. 2001) (model 2). The disaggregated values obtained based on model 1 were compared with the observed mean daily values of R_H for 1994–2000 in Figs. 15 and 16. Thus, objective 5 set for the study is addressed.

7 Summary and conclusions

A two-stage methodology was developed to obtain the future projections of daily R_H in a river basin for combinations of IPCC SRES scenarios (A1B, A2, B1, and COMMIT) and the selected predictors (nine sets of predictors) using SVM for downscaling.. In the first stage, SVM model was developed to downscale large-scale atmospheric variables to R_H in a river basin at monthly scale. Subsequently, the monthly sequences of R_H were disaggregated to daily scale using *k*-NN method. The idea behind developing the two-stage methodology is that the monthly sequences of atmospheric variables simulated by the GCM are more reliable than those simulated at daily scale. Further, the SVM requires large computational effort to directly downscale daily sequences of the large-scale atmospheric variables to daily sequences of hydrometeorological variables in a river basin.

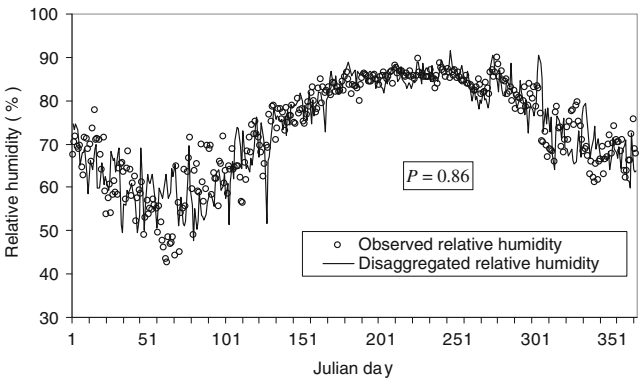

Fig. 15 Observed mean daily relative humidity for the validation period (January 1994 to December 2000) is compared with that obtained using *k*-NN disaggregation model

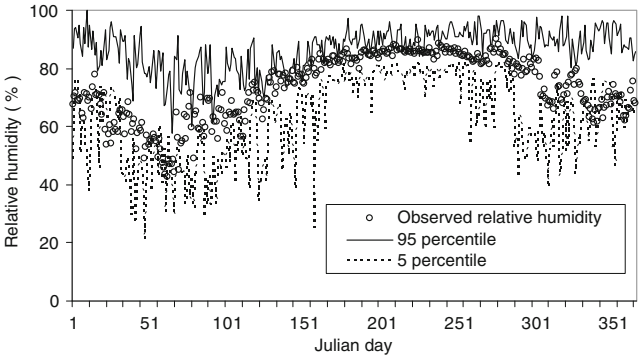

Fig. 16 The distribution of the disaggregated mean daily relative humidity obtained using *k*-NN disaggregation model. The *solid line* denotes the 95 percentile of the disaggregated values, whereas the *dotted line* indicates the 5 percentile of the same, respectively. The *circle* denotes the observed mean daily relative humidity for the validation period (January 1994 to December 2000)

The large-scale atmospheric variables namely temperature and specific humidity at 925 mb, latent heat, and surface temperature, which have a physically meaningful relationship with R_H, were chosen as the predictors for downscaling R_H.

Stratifications based on the seasons and the location (land/sea) were considered to facilitate the development of a separate downscaling model for capturing relationship between the predictors and predictand for each season (wet and dry), location (land and sea), and combination of season and location (wet season and land, dry season and land, wet season and sea, dry season and sea). Scatterplots and the three measures of dependence (product moment correlation, Spearman's rank correlation, and Kendall's tau) were used to evaluate the relationships between predictors in NCEP and GCM data sets and to study the predictor–predictand relationships.

Results indicate that the relationship between the predictor variables and the R_H varies with season and location, and stratifications are necessary for selecting the potential predictors for downscaling. Selecting very few predictors did not improve the model performance, as they were insufficient to represent the thermodynamics and dynamics of circulation in the region. On the other hand, selecting all the predictors had adverse affect on the performance of the developed model, due to the addition of noise. Hence, a judicious combination of the predictors was preferred. The results of SVM downscaling models indicated that the R_H downscaled using the predictors for land-based stratification performed the best.

On the whole, for the best SVM model, the R_H was projected to increase in the future for A1B and A2 scenarios, whereas no trend was discerned for B1 and COMMIT scenarios. The trend in the projected R_H was found to be sensitive to the predictors selected.

The *k*-NN model developed to disaggregate the sequence of monthly mean R_H to daily mean R_H was found to be more efficient than the triangular distribution method of the SWAT model.

Acknowledgments The authors express their gratitude to the editor and the reviewer who have provided constructive and helpful comments. This work is partially supported by the Dept. of Science and Technology, Govt. of India, through AISRF project no. DST/INT/AUS/P-27/2009. The third author acknowledges support from the Ministry of Earth Sciences, Govt. of India, through project no. MoES/ATMOS/PP-IX/09. The support from the Drought Monitoring Cell, Government of Karnataka, is acknowledged. Special thanks are also due to our alumnus Dr. Shivam and Dr. Vidyunmala, Indian Institute of Science, Bangalore, for their valuable inputs.

Appendix 1

The downscaled scenarios are constructed by translating GCM-simulated information from coarser scale to finer watershed scale using spatial downscaling models, based on the assumption that regional climate is conditioned by the climate on a relatively larger scale (e.g., continental). The spatial downscaling techniques can be broadly classified into dynamic downscaling and statistical downscaling. In the dynamic downscaling approach, a Regional Climate Model (RCM) is embedded into GCM. The RCM is essentially a numerical model in which GCMs are used to fix boundary conditions. The major drawback of RCM, which restricts its use in climate impact studies, is its complicated design and high computational cost. Whereas, the statistical downscaling involves deriving empirical relationships that transform large-scale features of the GCM (*LF*) to regional-scale variables (*RSV*)

$$\mathrm{RSV} = g(\mathrm{LF})$$

where RSV represents predictands, LF refers to predictors, and g is a downscaling function which could be deterministic or stochastic.

The classical statistical downscaling techniques include weather classification methods, weather generators, and transfer functions. The simple and commonly used statistical downscaling approaches are based on transfer functions, which model relationships between predictors and predictand using methods such as linear and nonlinear regression, artificial neural networks, canonical correlation, principal component analysis, and SVM. In this paper, the transfer function-based statistical downscaling method is chosen for determining plausible future scenarios of relative humidity.

Least-Square Support Vector Machine

The Least-Square Support Vector Machine (LS-SVM) has been used in this study for downscaling. Details of the same can be found in Suykens (2001) and Tripathi et al. (2006). This subsection presents the underlying principle of the LS-SVM.

Consider a finite training sample of N patterns $\{(\mathrm{x}_i, y_i), i = 1, ..., N\}$, where x_i is the ith pattern in n-dimensional space (i.e., $x_i = [x_{1i}, \ldots, x_{ni}] \in \Re^n$), and it constitutes input to LS-SVM, whereas $Y_i \in \Re$ is the corresponding value of the desired model output. Further, let the learning machine be defined by a set of possible mappings $x \mapsto f(x,w)$, where $f_{(\cdot)}$ is a deterministic function which, for a given input pattern x and adjustable parameters w ($w \in \Re^n$), always gives the same output. The training phase of the learning machine involves adjusting the parameter w. These parameters are estimated by minimizing the cost function Ψ_L (w,e).

$$\psi_{\mathrm{L}}(w,e) = \frac{1}{2} w^{\mathrm{T}} w + \frac{1}{2} C \sum_{i=1}^{N} e_i^2 \tag{5}$$

subject to the equality constraint

$$y_i - \widehat{y}_i = e_i \quad i = 1, ..., N \tag{6}$$

where C is a positive real constant, and $\widehat{y}_i$ is the actual model output. The first term of the cost function represents weight decay or model complexity–penalty function. It is used to regularize the weight sizes and to penalize the large weights. This helps in improving the generalization performance (Hush and Horne 1993). The second term of the cost function represents penalty function.

The solution of the optimization problem is obtained by considering the Lagrangian as

$$L(w,b,e,a) = \frac{1}{2} w^{\mathrm{T}} w + \frac{1}{2} C \sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N} \alpha_i \{\widehat{y}_i + e_i - y_i\} \tag{7}$$

where α_i are Lagrange multipliers, and b is the bias term. The conditions for optimality are given by

$$\begin{cases} \frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \alpha_i \phi(x_i) = 0 \\ \frac{\partial L}{\partial b} = \sum_{i=1}^{N} \alpha_i = 0 \\ \frac{\partial L}{\partial e_i} = \alpha_i - C e_i = 0 \qquad i = 1, ..., N \\ \frac{\partial L}{\partial \alpha_i} = \widehat{y}_i + e_i - y_i = 0 \quad i = 1, ..., N \end{cases} \tag{8}$$

The above conditions of optimality can be expressed as the solution to the following set of linear equations after elimination of w and e_i.

$$\begin{bmatrix} 0 & \overrightarrow{1}^{\mathrm{T}} \\ \overrightarrow{1} & \Omega + C^{-1}I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{9}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}; \overrightarrow{\mathbf{1}} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{N\times 1}; \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix}; \tag{10}$$

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}_{N\times N}$$

In Eq. (9), Ω is obtained from the application of Mercer's theorem.

$$\Omega_{i,j} = K(x_i, x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_j) \quad \forall i,j \tag{11}$$

where $\phi(\cdot)$ represents nonlinear transformation function defined to convert a nonlinear problem in the original lower dimensional input space to linear problem in a higher dimensional feature space.

The resulting LS-SVM model for function estimation is:

$$f(x) = \sum \alpha_i^* K(x_i, x) + b^* \tag{12}$$

where α_i^* and b^* are the solutions to Eq. (7), $K(x_i, x)$ is the inner product kernel function defined in accordance with Mercer's theorem (Courant and Hilbert 1970; Mercer 1909), and b^* is the bias. There are several choices of kernel functions, including linear, polynomial, sigmoid, splines, and radial basis function (RBF). The linear kernel is a special case of RBF (Keerthi and Lin 2003). Further, the sigmoid kernel behaves like RBF for certain parameters (Lin and Lin 2003). In this study, RBF is chosen to map the input data into higher dimensional feature space, which is given by:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i, x_j\|^2}{\sigma}\right) \tag{13}$$

where, σ is the width of the RBF kernel, which can be adjusted to control the expressivity of RBF. The RBF kernels have localized and finite responses across the entire range of predictors.

The advantage of RBF kernel is that it maps the training data non-linearly into a possibly infinite-dimensional space, and thus, it can effectively handle the situations when the relationship between predictors and predictand is nonlinear. Moreover, the RBF is computationally simpler than polynomial kernel, which requires more parameters. It is worth mentioning that developing LS-SVM with RBF kernel involves a judicious selection of RBF kernel width σ and parameter C.

The software used in this study is the "LS-SVMlab: a MATLAB toolbox for Least Squares Support Vector Machines." Details of the software can be found at http://www.esat.kuleuven.ac.be/sista/lssvmlab/tutorial/lssvmlab_paper0.pdf. The running time for the worst case (the maximum of the running times over all nine cases considered in the study) was 15 min, while the same for the best case was 5 min (on a Pentium PC). The average run time over all the cases was 9 min.

References

Alley R, Berntsen T, Bindoff N, Chen Z, Chidthaisong A, Friedlingstein P, Gregory J, Hegerl G, Heimann M, Hewitson B, Hoskins B, Joos F, Jouze LJ, Kattsov V, Lohmann U, Manning M, Matsuno T, Molina M, Nicholls N, Overpeck J, Qin D, Raga G, Ramaswamy V, Ren J, Rusticucci M, Solomon S, Somerville R, Stocker T, Stott P, Stouffer R, Whetton P, Wood R, Wratt D (2007) Climate change 2007: the physical science basis. Summary for Policymakers. IPCC eds.: Intergovernmental Panel on Climate Change. Geneva, Switzerland

Anandhi A (2008) Impact assessment of climate change on hydrometeorology of Indian river basin for IPCC SRES scenarios. Ph.D. thesis, Indian Institute of Science, Bangalore, India

Anandhi A (2011) Uncertainties in downscaled relative humidity for a semi-arid region in India. J Earth Syst Sci 120(3):375–386. doi:JESS-D-10-00145

Anandhi A, Srinivas VV, Nanjundiah RS, Kumar DN (2008) Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. Int J Climatol 28 (3):401–420

Anandhi A, Srinivas VV, Kumar DN, Nanjundiah RS (2009) Role of predictors in downscaling surface temperature to river basin in India for IPCC SRES scenarios using support vector machine. Int J Climatol 29(4):583–603

Bader D (ed) (2004) An appraisal of coupled climate model simulations. Report UCRL-TR-202550. Lawrence Livermore National Laboratory, Livermore, CA

Brown R, Rosenberg N (1997) Sensitivity of crop yield and water use to change in a range of climatic factors and CO2 concentrations: a simulation study applying EPIC to the central United States. Agric For Meteorol 83(3–4):171–203

Cahynová M, Huth R (2010) Circulation vs. climatic changes over the Czech Republic: a comprehensive study based on the COST733 database of atmospheric circulation classifications. Physics and Chemistry of the Earth, Parts A/B/C 35(9–12):422–428

Courant R, Hilbert D (1970) Methods of mathematical physics, vols. I and II. Wiley Interscience, New York

Dai A (2006) Recent climatology, variability, and trends in global surface humidity. J Clim 19(15):3589–3606

Debele B, Srinivasan R, Parlange J (2007) Accuracy evaluation of weather data generation and disaggregation methods at finer timescales. Advances in Water Resources 30(5):1286–1300. doi:10.1016/j.advwatres.2006.11.009

Doblas-Reyes FJ, Hagedorn R, Palmer TN (2006) Developments in dynamical seasonal forecasting relevant to agricultural management. Clim Res 33(1):19–26. doi:10.3354/cr033019

Doty B, Kinter JI (1993) The Grid Analysis and Display System (GrADS): a desktop tool for earth science visualization. Paper presented at the American Geophysical Union 1993 Fall Meeting, San Francisco, CA, 6–10 December

Easterling W, Rosenberg N, McKenney M, Jones C, Dyke P, Williams J (1992) Preparing the erosion productivity impact calculator (EPIC) model to simulate crop response to climate change and the direct effects of CO_2. Special issue: methods for assessing regional agricultural consequences of climate change. Agric For Meteorol 59(1–2):17–34

Enke W, Spekat A (1997) Downscaling climate model outputs into local and regional weather elements by classification and regression. Clim Res 8:195–207

Frei C, Alberghi S, Amici M, Anagnostopoulou C, Bardossy A, Cacciamani C, Goodess C, Haylock M, Hundecha Y, Maheras P, Marchesi S, Morgillo A, Pavan V, Plaut G, Schmidli J, Smith T, Simonnet E, Tolika K, Tomozeiu R, Tosi E, Wilby R (2005) STARDEX-recommendations on the most reliable predictor variables and evaluation of inter-relationships (deliverable D13) (no. EVK2-CT-2001-00115). At: http://www.cru.uea.ac.uk/projects/stardex/deliverables/D13/D13_synthesis_v3.pdf. Accessed 24 Sept 2010

Gaffen DJ, Ross RJ (1999) Climatology and trends of U.S. surface humidity and temperature. J Clim 12(3):811–828

Green H, Kozek A (2003) Modeling weather data by approximate regression quantiles. Journal of Anziam 44(E):C229–C248

Helsel DR, Hirsch RM (2002) Chapter A3. Techniques of water-resources investigations, vol 4. U.S. Geological Survey

Hirai G, Okumura T, Takeuchi S, Tanaka O, Chujo H (2000) Studies on the effect of the relative humidity of the atmosphere on the growth and physiology of rice plants: effects of relative humidity during the light and dark periods on the growth. Plant Production Science 3(2):129–133

Hofer M, Mölg T, Marzeion B, Kaser G (2010) Empirical-statistical downscaling of reanalysis data to high-resolution air temperature and specific humidity above a glacier surface (Cordillera Blanca, Peru). J Geophys Res 115(D12):D12120. doi:10.1029/2009jd012556

Houghton J, Ding Y, Griggs D, Noguer M, van der Linden P, Dai X, Maskell K, Johnson C (eds) (2001) Climate change 2001: the scientific basis. Cambridge University Press, Cambridge

Hush DR, Horne BG (1993) Progress in supervised neural networks: what's new since Lippmann? IEEE Signal Processing Magazine 10:8–39

Huth R (2005) Downscaling of humidity variables: a search for suitable predictors and predictands. Int J Climatol 25:243–250

Jones RH, Westra S, Sharma A (2010) Observed relationships between extreme sub-daily precipitation, surface temperature, and relative humidity. Geophys Res Lett 37(22):L22805. doi:10.1029/2010gl045081

Jung M, Reichstein M, Ciais P, Seneviratne SI, Sheffield J, Goulden ML, Bonan G, Cescatti A, Chen J, de Jeu R, Dolman AJ, Eugster W, Gerten D, Gianelle D, Gobron N, Heinke J, Kimball J, Law BE, Montagnani L, Mu Q, Mueller B, Oleson K, Papale D, Richardson AD, Roupsard O, Running S, Tomelleri E, Viovy N, Weber U, Williams C, Wood E, Zaehle S, Zhang K (2010) Recent decline in the global land evapotranspiration trend due to limited moisture supply. Nature 467(7318):951–954

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 KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77(3):437–471

Keerthi SS, Lin C-J (2003) Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput 15(7):1667–1689

Koch S, Aksakal A, McQueen J (1997) The influence of mesoscale humidity and evapotranspiration fields on a model forecast of a cold-frontal squall line. Mon Weather Rev 125(3):384–409

Kostopoulou E, Giannakopoulos C, Anagnostopoulou C, Tolika K, Maheras P, Vafiadis M, Founda D (2007) Simulating maximum and minimum temperature over Greece: a comparison of three downscaling techniques. Theor Appl Climatol 90(1):65–82. doi:10.1007/s00704-006-0269-x

Kottegoda NT, Rosso R (1998) Statistics, probability, and reliability for civil and environmental engineers. International edition. McGraw-Hill, Columbus

Lall U, Sharma A (1996) A nearest neighbor bootstrap for time series resampling. Water Resourses Research 32(3):679–693

Lin H-T, Lin C-J (2003) A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Department of Computer Science and Information Engineering. National Taiwan University, Taiwan

Maurer EP, Hidalgo HG, Das T, Dettinger MD, Cayan DR (2010) The utility of daily large-scale climate data in the assessment of climate change impacts on daily streamflow in California. Hydrol Earth Syst Sci 14(6):1125–1138. doi:10.5194/hess-14-1125-2010

Mercer J (1909) Functions of positive and negative type and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society, London, A 209:415–446

Monteith J (1981) Evaporation and surface temperature. Q J R Meteorol Soc 107(451):1–27

Neitsch S, Arnold J, Kiniry J, Williams J (2001) Soil and water assessment tool theoretical documentation. Blackland Research Center. Texas Agricultural Experiment Station, Temple

New M, Lister D, Hulme M, Makin I (2002) A high-resolution data set of surface climate over global land areas. Clim Res 21(1):1–25

O'Gorman PA, Muller CJ (2010) How closely do changes in surface and column water vapor follow Clausius-Clapeyron scaling in climate change simulations? Environ Res Lett 5(2):025207

Parthasarathy B, Munot AA, Kothawale DR (1994) All-India monthly and seasonal rainfall series: 1871–1993. Theor Appl Climatol 49(4):217–224. doi:10.1007/bf00867461

Reichler T, Kim J (2008) How well do coupled models simulate today's climate? Bull Am Meteorol Soc 89(3):303–311. doi:10.1175/BAMS-89-3-303

Schmidli J, Frei C, Vidale PL (2006) Downscaling from GCM precipitation: a benchmark for dynamical and statistical downscaling methods. Int J Climatol 26(5):679–689. doi:10.1002/joc.1287

Schoof JT, Pryor SC (2003) Evaluation of the NCEP–NCAR reanalysis in terms of synoptic-scale phenomena: a case study from the Midwestern USA. Int J Climatol 23(14):1725–1741. doi:10.1002/joc.969

Simmons AJ, Willett KM, Jones PD, Thorne PW, Dee DP (2010) Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: inferences from reanalyses and monthly gridded observational data sets. J Geophys Res 115(D1):D01110. doi:10.1029/2009jd012442

Stanghellini C, de Jong T (1995) A model of humidity and its applications in a greenhouse. Agric For Meteorol 76(2):129–148

Suykens JAK (2001) Nonlinear modelling and support vector machines. In: IEEE Instrumentation and measurement technology conference, Budapest, Hungary, 2001. pp 287–294

Teutschbein C, Wetterhall F, Seibert J (2011) Evaluation of different downscaling techniques for hydrological climate—change

impact studies at the catchment scale. Climate Dynamics (in press). 1–19. doi:10.1007/s00382-010-0979-8
Tripathi S, Srinivas VV, Nanjundiah RS (2006) Downscaling of precipitation for climate change scenarios: a support vector machine approach. J Hydrol 330:621–640. doi:10.1016/j.jhydrol.2006.04.030
Vapnik V (1995) The nature of statistical learning theory. Springer, New York
Vapnik V (1998) Statistical learning theory. Wiley, New York
Wang JXL, Gaffen DJ (2001) Late-twentieth-century climatology and trends of surface humidity and temperature in China. J Clim 14 (13):2833–2845
Wetterhall F, Halldin S, Xu C (2005) Statistical precipitation downscaling in central Sweden with the analogue method. J Hydrol 306:136–174
Wetterhall F, Halldin S, Xu CY (2007) Seasonality properties of four statistical-downscaling methods in central Sweden. Theor Appl Climatol 87(1):123–137. doi:10.1007/s00704-005-0223-3
Wetterhall F, Bárdossy A, Chen D, Halldin S, C-y Xu (2009) Statistical downscaling of daily precipitation over Sweden using GCM output. Theor Appl Climatol 96(1):95–103. doi:10.1007/s00704-008-0038-0
Widmann M, Bretherton CS, Salathe EP Jr (2003) Statistical precipitation downscaling over the northwestern United States using numerically simulated precipitation as a predictor. J Clim 16(5):799–816
Wilby RL, Hay LE, Leavesley GH (1999) A comparison of downscaled and raw GCM output: implications for climate change scenarios in the San Juan River basin, Colorado. J Hydrol 225(1–2):67–91
Willett K, Jones PD, Thorne PW, Gillett N (2010) A comparison of large scale changes in surface humidity over land in observations and CMIP3 general circulation models. Environ Res Lett 5(2):025210
Wright JS, Sobel A, Galewsky J (2010) Diagnosis of zonal mean relative humidity changes in a warmer climate. J Clim 23 (17):4556–4569
Wypych A (2010) Twentieth century variability of surface humidity as the climate change indicator in Kraków (Southern Poland). Theor Appl Climatol 101(3):475–482. doi:10.1007/s00704-009-0221-y