period January 1990 to December 2000 (Khobragade 2009). The mean monthly *T*max in the catchment varies from 19°C to 39.5°C and mean annual *Tmax* is 30.6°C. The mean monthly *T*min ranges from 3.4°C to 29.8°C based on decadal (1990–2000) observed value. The observed mean monthly *T*max and *T*min have been shown in Figure 9.3 for various months of year 2000 respectively. The study area receives an average annual precipitation of 597 mm. It has a tropical monsoon climate where most of the precipitation is confined to a few months of the monsoon season. The south–west (summer) monsoon has warm winds blowing from the Indian Ocean causing copious amount of precipitation during June–September months. The observed precipitation has been shown in Figure 9.4 for various months of year 2000. 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 by 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) and future simulations forced by four emission scenarios, namely A1B, A2, B1 and COMMIT.

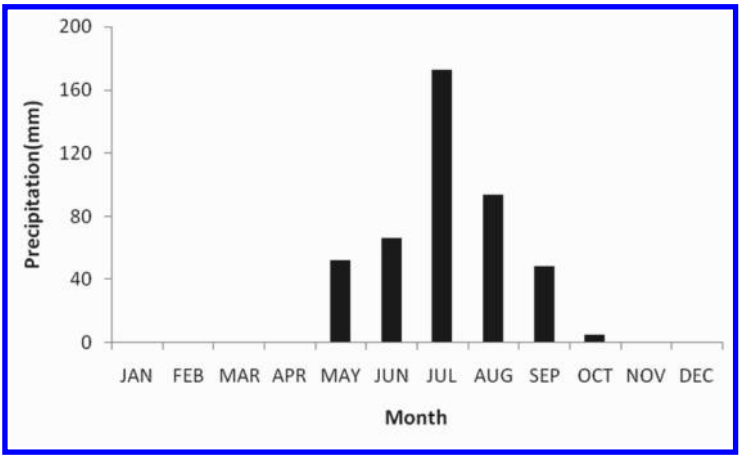

Figure 9.4. Observed precipitation for the study region for year 2000

The nine grid points surrounding the study region as shown in Figure 9.2 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 are 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.

9.4.1 Selection of Predictors

The selection of appropriate predictors is one of the most important steps in a downscaling exercise for downscaling predictands. Predictors have to be selected based both on their relevance to the downscaled predictands and their ability to be accurately represented by the GCMs. The most favorable predictors must be strongly correlated with the predictand, be physically sensible, and have the ability to capture the climate change signal (Goyal et al. 2010). The predictors are chosen by the following criteria: (1) they should be skilful in representing large-scale variability that is simulated by the GCMs and are readily available from archives of GCM output and reanalysis data sets; (2) they should strongly correlated with the surface variables of interest/predictands, i.e., they should be statistically significant contributors to the variability in predictands; and (3) they should represent important physical processes in the context of the enhanced greenhouse effect (Ghosh and Mujumdar 2007; Goyal and Ojha 2012d). In this section, the selection of predictors for Pichola Lake basin has been carried out using (i) scatter plots and cross correlations and (ii) VIP scores obtained from PLS regression. The details of these approaches are given below.

9.4.1.1 Application of Scatter Plots and Cross Correlation

Cross-correlations and scatter plots are in use to select predictors to understand the presence of nonlinearity/linearity trends in dependence structure. Cross-correlations and scatter plots 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, and scatter plots are prepared between the predictor variables in NCEP and GCM datasets. The cross correlations are estimated using three measures of dependence, namely product moment correlation, Spearman's rank correlation and Kendall's tau. Spearman's rank correlation (ρ) is computed using the difference between the ranks of contemporaneous values of predictor and predictand (Di).

Various authors have used large-scale atmospheric variables, viz., air temperature, geo-potential height, zonal (u) and meridional (v) wind velocities, as the predictors for downscaling GCM output to temperature, precipitation and evaporation over an area. For this study, we have used total 9 possible predictor variables, namely, air temperature (at 925,500hPa and 200hPa pressure levels), geo-potential height (at 200hPa and 500hPa pressure levels), zonal (u) and meridional (v) wind velocities (at 925 and 200hPa pressure levels), as the predictors for downscaling GCM output to mean monthly temperature, precipitation and pan evaporation over the lake basin.

The cross-correlations enable verifying the reliability of the simulations of the predictor variables by the GCM, are shown in Tables 9.1, 9.2 and 9.3 for Tmax, Tmin and precipitation, respectively. In general, most of predictor variables are realistically simulated by the GCM where CC was greater than 0.65. It is noted that air temperature at 925hPa (Ta 925) is the most realistically simulated variable with a CC

greater than 0.8, while meridional wind at 200hPa (Va 200) is the least correlated variable between NCEP and GCM datasets (CC = -0.17). It is clear from Tables 9.1, 9.2 and 9.3 that air temperature at 925hPa (Ta 925), air temperature at 500 hPa (Ta500), air temperature at 200 hPa (Ta200), meridional wind at 925hPa (Va 925), zonal wind at 925hPa (Ua925), geo-potential height at 200hPa (Zg200) and geo-potential height at 500hPa (Zg500) are better correlated than meridional wind at 200hPa (Va200) and zonal wind at 200hPa (Ua200). The cross-correlations are computed between the predictor variables in NCEP and GCM datasets (Table 9.4).

Scatter plots are prepared between the predictor variables in NCEP and GCM datasets (Figures 9.5 and 9.6). It is to be noted that these figures represent how well the predictors simulated by NCEP and GCM are correlated. Generally, the correlations are not very high due to the differences in the simulations of GCM (e.g. for different runs) and possible errors in NCEP-reanalysis. In addition, the inherent errors due to re-gridding from GCM scale to NCEP scale also contribute to low correlation.

9.4.1.2 VIP Scores by the PLS Regression

The VIP (Variable Importance in the Projection) scores obtained by the PLS regression has been paid an increasing attention as an importance measure of each explanatory variable or predictor. The variable selection procedure under PLS is proposed with an application to downscaling technique for identifying influencing variables to understand the impact of climate change. The VIP scores which are obtained by PLS regression, can be used to select most influential variables or predictors, X. The VIP score can be estimated for *j-th* X-variable by

$$VIP_j = \sqrt{\frac{p}{\sum_{i=1}^{k} R_d(Y,t_i)} \sum_{i=1}^{k} R_d(Y,t_i) w_{ij}^2} \qquad \text{(Eq. 9.2)}$$

where R_d is defined as the mean of the squares of the correlation coefficients (R) between the variables and the component.

$$Rd(X,c) = \frac{1}{p} \sum_{i=1}^{k} R^2(x_j,c) \qquad \text{(Eq. 9.3)}$$

Table 9.1. Cross-correlation computed between probable predictors in NCEP data and observed Tmax (predictands), P, S and K represent Product moment correlation, Spearman's rank correlation and Kendall's tau, respectively

	Va200	Va925	Ta200	Ta500	Ta925	Zg200	Zg500	Ua925	Ua200
P	0.44	0.67	0.81	0.71	0.88	0.54	0.51	0.53	0.34
S	0.21	0.43	0.67	0.57	0.71	0.40	0.38	0.38	0.29
K	0.37	0.61	0.71	0.59	0.88	0.61	0.54	0.56	0.45

Table 9.2. Cross-correlation computed between probable predictors in NCEP data and observed Tmin (predictands), P, S and K represent Product moment correlation, Spearman's rank correlation and Kendall's tau, respectively

	Va200	Va925	Ta200	Ta500	Ta925	Zg200	Zg500	Ua925	Ua200
P	0.34	0.62	0.81	0.64	0.78	0.61	0.60	0.72	0.31
S	0.48	0.43	0.63	0.44	0.60	0.59	0.42	0.51	0.42
K	0.28	0.61	0.78	0.58	0.80	0.67	0.54	0.73	0.39

Table 9.3. Cross-correlation computed between probable predictors in NCEP data and observed Precipitation (predictands), P, S and K represent Product moment correlation, Spearman's rank correlation and Kendall's tau, respectively.

	Va200	Va925	Ta200	Ta500	Ta925	Zg200	Zg500	Ua925	Ua200
P	0.11	0.59	0.31	0.46	0.20	0.41	0.61	0.39	0.18
S	0.18	0.45	0.39	0.61	0.23	0.39	0.45	0.32	0.11
K	0.17	0.63	0.48	0.47	0.34	0.51	0.48	0.45	0.12

Table 9.4. 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

	Va200	Va925	Ta200	Ta500	Ta925	Zg200	Zg500	Ua925	Ua200
P	-0.18	0.67	0.66	0.81	0.83	0.81	0.60	0.79	0.23
S	-0.14	0.43	0.46	0.64	0.68	0.64	0.39	0.56	0.57
K	-0.20	0.61	0.68	0.85	0.87	0.85	0.59	0.76	0.73

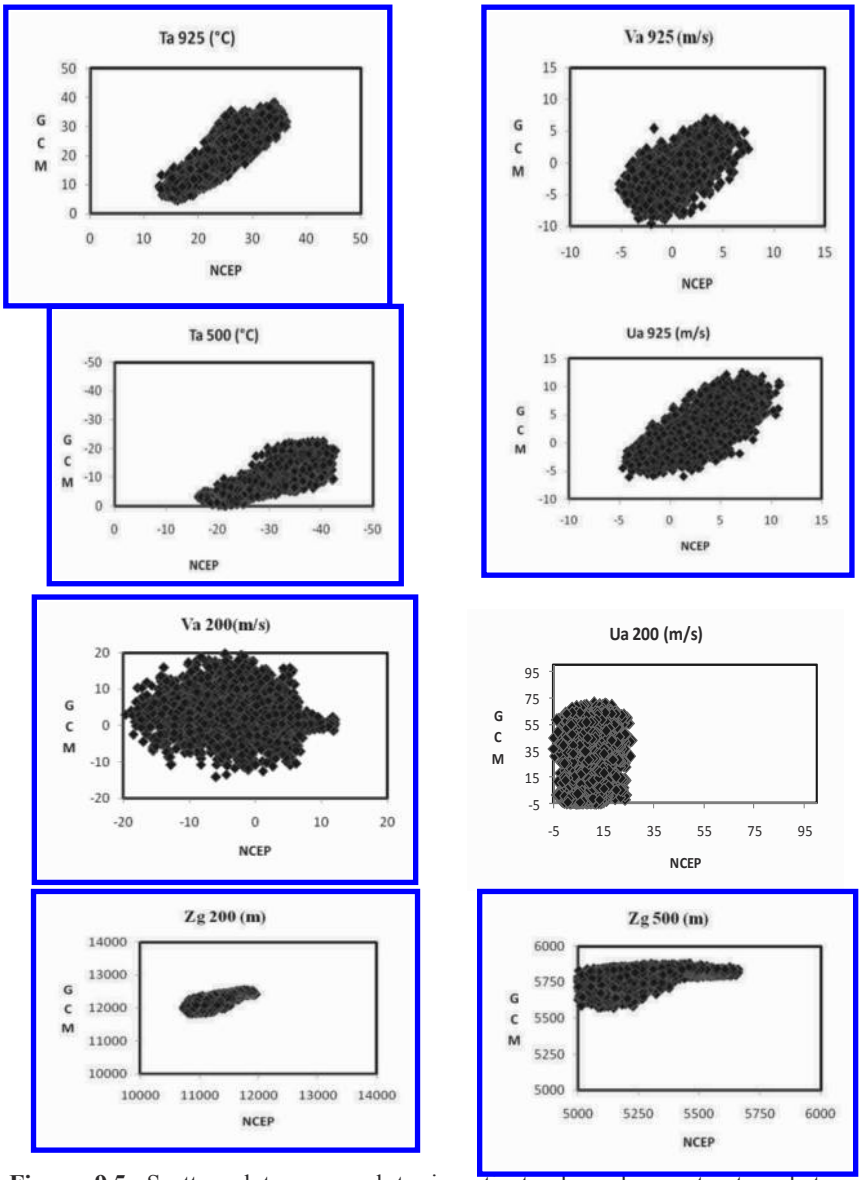

Figure 9.5. Scatter plots prepared to investigate dependence structure between probable predictor variables in NCEP and GCM datasets

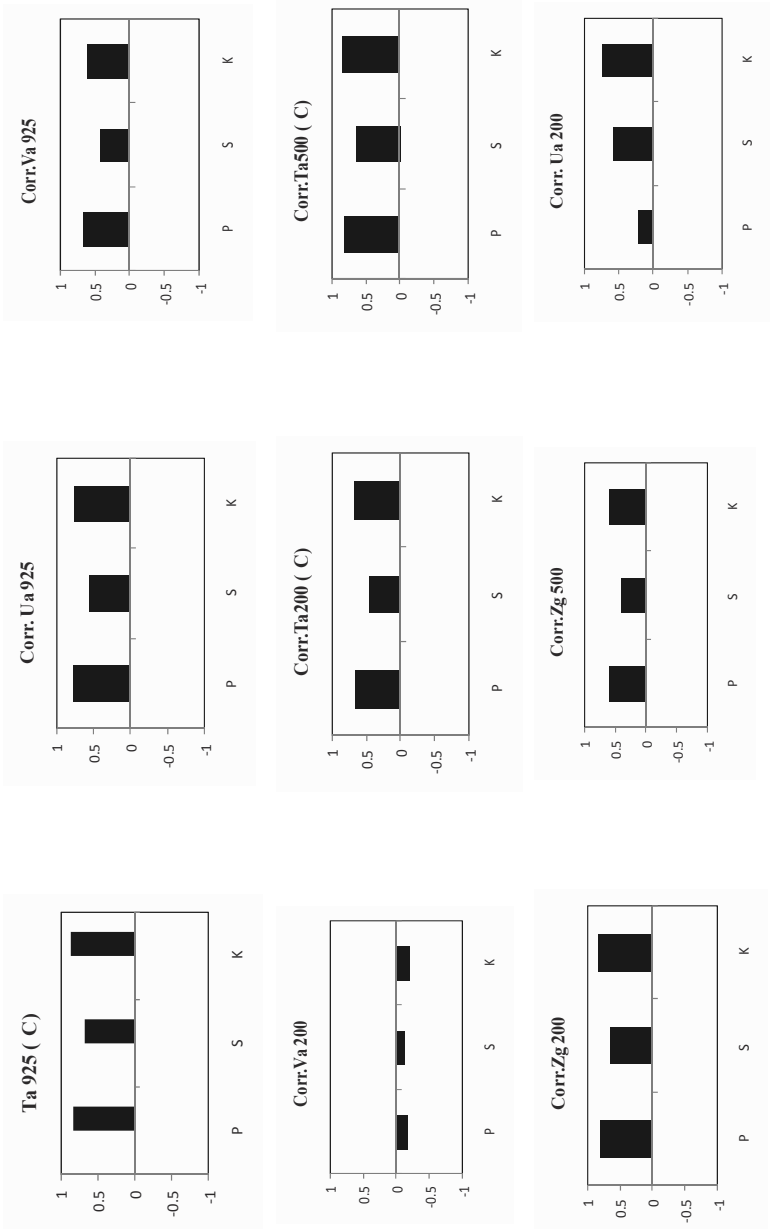

Figure 9.6. Bar plots for cross-correlation computed between probable predictors in NCEP and GCM datasets. P, S and K represent product moment correlation, Spearman's rank correlation and Kendall's tau respectively

Usually the predictor variable whose VIP score is greater than 0.8 and above is considered as an important variable. It can be seen from Figures 9.7, 9.8 and 9.9 that seven predictor variables namely air temperature at 925hPa, 500hPa and 200hPa; zonal wind (925hPa); meridional wind (925hPa); geo-potential height 500hPa and 200hPa have their VIP scores greater than 0.8. Correlation matrices of predictors also yielded the similar results. It is noted that different predictors control different local variables, and mean temperature is most sensitive to surface and near surface atmospheric factors.

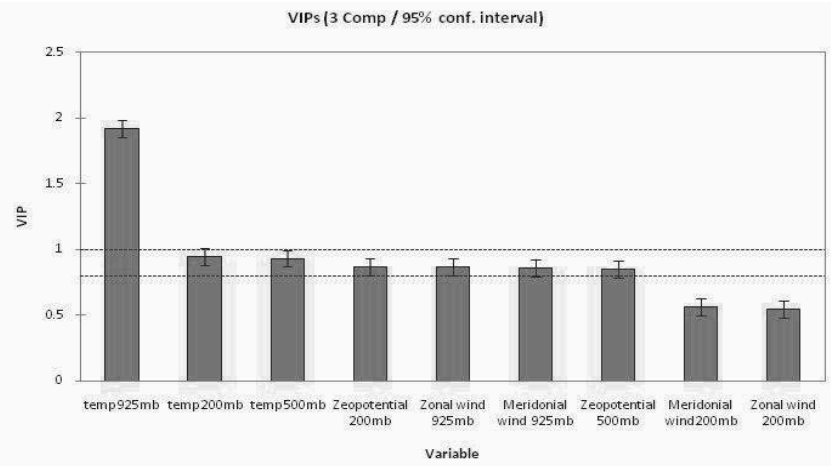

Figure 9.7. VIP of the predictand variable (Tmax) of the two-component PLSR model

9.4.2 Correcting Bias by a Multiplicative Shift

Many GCMs either overestimate or underestimate maximum and minimum temperature as well as precipitation. The correction scheme brings the distributions close to the observed pattern. A simple multiplicative shift is used to correct the bias of the mean monthly GCM simulated variable as follows:

$$X_i^{'} = X_i \frac{\overline{X}_{obs}}{\overline{X}_{GCM}} \qquad \text{(Eq. 9.4)}$$

where $X_i^{'}$, X_i refers to raw and corrected GCM simulated variable, and $\overline{X}_{GCM}$ and $\overline{X}_{obs}$ are long term mean monthly variable from the GCM and the observations for given month (Ines and Hansen, 2006)

9.5 Evaluation of Linear Regression Methods

9.5.1 Model Development

In this section, various linear regression approaches are used to downscale the mean monthly precipitation for the Pichola lake region. 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 to 2000 because it is of sufficient duration to establish a reliable climatology, yet not too long, nor too contemporary to include a strong global change signal. The procedure typically involves subtraction of mean and division by standard deviation of the predictor variable for a predefined baseline period for both NCEP/NCAR and GCM output. A feature vector (standardized predictor) is formed for each month of the record using the data of standardized NCEP predictor variables. However, another way to implement the regression model is that principal components should be extracted first since 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). Then, the use of principal component (PCs) as input to a downscaling model helps in making the model more stable and at the same time reduces its computational burden. Here, regression approaches with and without principal components have been used in this analysis.

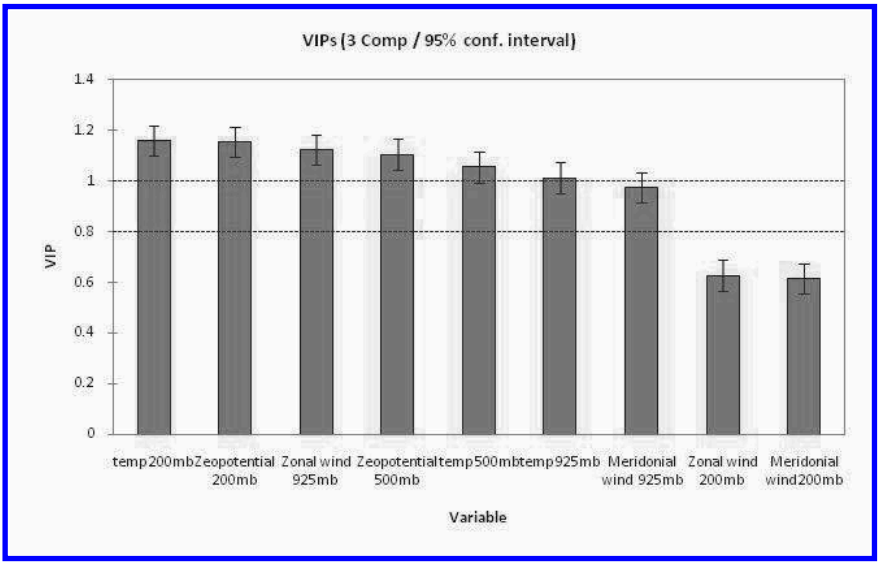

Figure 9.8. VIP of the predictand variable (Tmin) of the two-component PLSR model

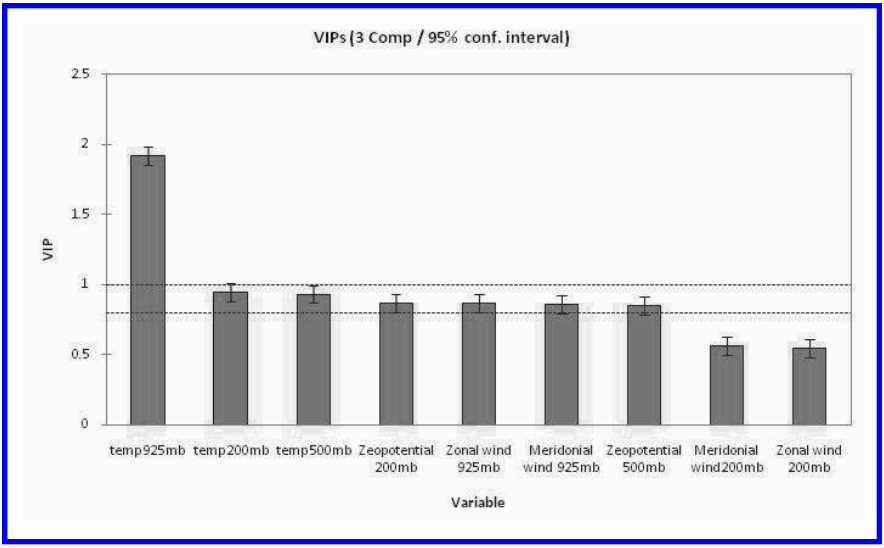

Figure 9.9. VIP of the predictand variable (precipitation) of the three-component PLSR model

To develop downscaling models using regression approaches (see Table 9.5), the feature vectors which are prepared from 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. In case of using PCA, it is observed that the four leading principal components (PCs) of the PCA method explained about 97% of the information content (or variability) of the original predictors. Hence, PCs are extracted to form feature vectors from the standardized data of potential predictors. The 11-year mean monthly observed precipitation data series were broken up into a calibration period and a validation period. The models were calibrated on the calibration period of 1990 to 1995 and validation involved the period of 1996 to 2000. Seven predictor variables, namely air temperature (925hPa, 500hPa and 200hPa); zonal wind (925hPa); meridional wind (925hPa); geo-potential height (500hPa and 200hPa) at 9 NCEP grid points with a dimensionality of 63, are used as the standardized data of potential predictors. These feature vectors are provided as input to the various regressions downscaling model.

Table 9.5. Different regression models used for obtaining projections of precipitation

Approach	Stepwise		Forward		Backward		Direct	
	Without PCs	Without PCs	Without PCs	Without PCs	Without PCs	Without PCs	Without PCs	Without PCs
Model	M1	MP1	M2	MP2	M3	MP3	M4	MP4

9.5.2 Training and Validation Results

Results of the different regression models (viz.*M1* to *M4* and *MP1* to *MP4*) as discussed in Table 9.5 are tabulated in Table 9.6. For predictand precipitation, the coefficient of correlation (CC) was in the range of 0.60–0.95; RMSE was in the range of 27.71–58.33; N-S Index was in the range of 0.24–0.90 and MAE was in the range of 0.23–0.72 for regression based models for the training and validation set. It can be observed from Table 9.6 that the performance of direct regression models with and without principal components for mean monthly precipitation are clearly superior to that of forward-, backward- and stepwise-regression-based models in the training data set while the performance of stepwise- and forward-regression-based models for predictand are clearly superior to that of backward- and direct-regression-based models in the validation data set. Results of forward and stepwise regression are quite similar. However, models developed using principal components yielded slightly better results. It can be inferred that model *MP4* using direct regression performed best for predictand precipitation. Now, multiplicative shift is used to correct the bias of GCM of model MP4. The corrected model MP4 performed better than uncorrected in terms of various performance meausres (CC, RMSE and N-S Index), as shown in Table 9.7. It can be inferred that the performance of direct regression models bias corrected (viz. MP4(corrected) performed well.

Table 9.6. Various performance statistics of models using various regression approaches for precipitation

Model	CC		RMSE		N-S Index	
	Training	Validation	Training	Validation	Training	Validation
M1	0.90	0.79	39.37	45.80	0.81	0.53
M2	0.95	0.60	27.77	58.33	0.90	0.24
M3	0.94	0.65	32.03	55.14	0.87	0.32
M4	0.91	0.79	39.33	45.80	0.81	0.53
MP1	0.90	0.80	39.18	44.01	0.80	0.51
MP2	0.94	0.61	27.71	55.34	0.91	0.25
MP3	0.95	0.66	31.03	55.64	0.88	0.35
MP4	0.93	0.81	38.65	44.04	0.82	0.57

Table 9.7. Various performance statistics of model using bias correction for precipitation

Model	CC		RMSE		N-S Index	
	Training	Validation	Training	Validation	Training	Validation
MP4 (corrected)	0.94	0.82	37.71	41.44	0.86	0.62

A comparison of mean monthly observed precipitation with precipitation simulated using forward regression models *MP4 (corrected)* has been shown from Figure 9.10 for the calibration and validation period. Regression coefficients (Aij) for predictor (*precipitation*) corresponding to model *M4* has been shown in Table 9.8 where i ranges from 1 to 7 indicating Ta 925, Ua 925, Va 925, Ta 500, Ta 200, Zg