Logistic Model as a Statistical Downscaling Approach for Forecasting a Wet or Dry Day in the Bagmati River Basin

Rajendra Man Shrestha[*,§], Srijan Lal Shrestha and Azaya Bikram Sthapit

Central Department of Statistics, Tribhuvan University, Kirtipur, Nepal

Abstract: A binary logistic model is developed for probabilistic prediction of a wet or dry day based upon daily rainfall data from 1981 to 2008 taken from 25 stations of Bagmati River basin. The predictor variables included in the model are daily relative humidity, air surface temperature, sea level pressure, v-wind which are expressed as principal components of 9 grids of the National Centers for Environmental Protection (NCEP)/National Center for Atmospheric Research (NCAR) Reanalysis data with resolution of $2.5^0 \times 2.5^0$. Principal component analysis is used to reduce the dimension of the predictors in the presence of spatial correlations between grids and thus reduce their multicollinearity effect. The result depicts that the model has 86.4 percent predictive capability in the analysis period (1981-2000) and 86.1 in the validation period (2001-2008) along with support of receiver operating characteristic (ROC) analysis. The results demonstrate that the first two principal components of relative humidity are the key predictor variables with respective odds ratios (ORs) of 4.18 and 3.61, respectively. The other statistically significant predictors are the second principal component of v-wind with OR 1.43, the second and first principal components of air surface temperature with ORs 1.38 and 0.76, respectively and the first principal component of sea level pressure with OR 0.44. Goodness-of-fit test, ROC analysis and other main diagnostic tests showed that the fitted logistic model is characterized by good fits for analysis as well as validation period.

Keywords: Binary logistic model, climate change, principal components, rainfall, statistical downscaling, weather variables.

1. INTRODUCTION

Climate is a very important natural process that affects human life and the environment. The prediction about weather and climate pattern, especially rainfall is required not only by the agricultural sector but also for hydropower project [1]. In nature, the rainfall pattern depends on a day that may be wet or dry and if it is a wet day, then its pattern may be extreme. The extreme rainfall when happens may cause serious damage with great socioeconomic losses by heavy floods or by prolong droughts [2]. This situation drives us to have a sound methodology and technique to understand such phenomena correctly as far as possible. The past researches show that the models can simulate climate such that it can be used to predict the rainfall occurrence in particular, after investigation, for a given day in an area. General Circulation Model (GCM) is one of the recent climate models to observe the impact and to predict the climate change. However, GCM outputs are not suitable for direct use to assess the climate change impact at local level because of their oversimplification in terms of coarse resolution input information, equations and others [3]. Furthermore, GCM uses information on orography, land surface or other at coarse resolution. But climate forcing and circulations that influence local climate generally occur at much finer scale than that accounted by GCM. Therefore, one of the common methods used to solve this problem is the statistical downscaling technique such that there is a statistical relationship from observations between large scale variables and a local variable at basin.

From past studies, it is found that there are several works implemented to use different downscaling techniques. For example, Wigena (2006) has elaborated statistical downscaling model to predict the rainfall in Indramayu and one of his analysis was to determine the best domain output by using projection pursuit regression [4]. Furthermore, Cavazos and Hewitson (2002) studied the performance of GCM and NCEP/NCAR output to find the potential combination of response variables by using artificial neural network (ANN) [5]. Similarly, the most commonly used statistical downscaling techniques are regression methods [6]. In the context of Nepal, a few climate change impact studies have been conducted in Nepalese region [7, 8]. For instance, Mishra *et al*. (2011) used quantile-based bias correction method for climate projection downscaling and impact assessment on precipitation over upper Bagmati River basin in Nepal [3]. Babel *et al*. (2013) used SDSM version 4.2 as a statistical downscaling technique to study about climate change and water resources in the upper Bagmati River basin, Nepal [9]. Likewise, Parajuli *et al*. (2014) have studied about impact of climate change on River flow and hydropower production in Kulekahni hydropower project of Nepal by using SDSM [10]. Till the present research works, there are no research reports/journals articles as such available to investigate probabilistic forecasting that a day is a wet day or dry day which used a suitable statistical

*Address correspondence to this author at the Central Department of Statistics, Tribhuvan University, Kirtipur, Nepal; Tel: 977-01-4331710; E-mail: rajendramanshrestha65@gmail.com

[§]On leave from Padma Kanya Multiple Campus, Kathmandu, Nepal for PhD research.

1874-2823/15 2015 Bentham Open

downscaling model in Nepal. Thus, in order to fill this knowledge gap at the target area, the present study is conducted for further investigation of the rainfall pattern under influence of all possible climatic predictor variables simulated by GCMs.

In this study, the objective is set to develop a predictive model by use of logistic regression method on the basis of reanalysis data available from NCEP/NCAR project. This predictive model is used to investigate whether a given day is a wet / dry with associated probability. It is also expected that, with available daily GCM outputs with future emission scenarios, this model will be applicable for future projection of rainfall occurrence or wet day under the impact of climate change in the Bagmati River basin. While reviewing the papers pertaining to the rainfall pattern with use of logistic model as statistical downscaling technique, it is found that Nadja (2005) used GLM with logit link (logistic model) to simulate daily rainfall at Heathrow, Birmingham and Manchester airports, United Kingdom. The results were that all of the models projected a decrease in mean daily rainfall in summer and an increase in winter at Heathrow [11]. In addition, Prasad *et al.* (2010) used a logistic regression approach for monthly rainfall forecasts in meteorological subdivision of India based on DEMTER retrospective forecasts. The model showed good performance in capturing extreme rainfall years and appeared to perform better than the direct model forecasts of total precipitation in such years [12]. A study used quantile regression as statistical downscaling technique to estimate extreme monthly rainfall at station Bangkir Indonesia. The results showed that at 95th percentile, the pattern of forecasted rainfall in January to December 2008 was similar to actual rainfall with correlation 0.98 and the forecasted rainfall (843 mm) in February 2008 was considered as the extreme rainfall month which confirms well to the highest actual rainfall (727 mm) with probability 0.99 [13].

2. MATERIALS AND METHODOLOGY

This section deals with materials and methods adopted for the study.

2.1. Study Area

The Bagmati River basin, located within the middle mountain of Nepal extends from 26^0 45’-27^0 49’ N and 85^0 02’-$85^0$27’ E and has a catchment area of 3750 square kilometer (km) in Nepal. The Bagmati River originates from the Shivapuri hills of the Mahabharata range in the Kathmandu Valley and drains out of Nepal across the India state-Bihar. It reaches the River Ganges after passing through the inner Mahabharata range and the plain of Terai. Babel *et al.* (2013) has mentioned that the elevation of the Bagmati River basin ranges from about less than 80 m in Terai, its southern part to 2900 meter (m) in the Mahabharata range, its northern part. Its length is about 51 km in Nepal. Its main tributaries are Manohara, Bishnumati, Kulekhani, Kokhajor, Marin, Chandi, Jhanjh and Manusmara. The Kathmandu valley comprises of 15% of the basin area in Nepal. Main source of water in the Bagmati River basin is rain and natural springs [9].

2.2. Climate of the Study Area

There are four well-defined seasons in Nepal classified as winter (December to February), pre-monsoon (March to May), monsoon or summer (June to September) and post-monsoon (October to November). The climatic condition of the Bagmati River basin is quite changing due to the intrinsic topography. Temperature generally decreases with elevation and becomes low in winter and high in summer. More specifically, the climate changes from cold temperate in higher mountains *via* warm temperate at mid-elevation levels to subtropical in the southern low land. It seems the whole Bagmati River basin is divided according to its climatic zone. Cool temperate humid zone lies between 2000 and 3000 m which cover only about 5% of the basin with mean annual temperature varying between 10^0C to 15^0C. The warm temperate humid zone lies between 1000 and 2000 m which cover about 60% part of the basin with mean annual temperature varying between 15^0C to 20^0C. Lastly, the sub-tropical humid zone lies below 1000 m which covers southern part of the basin with the Siwaliks and Terai and mean annual temperature ranging between 20^0C to 30^0C. The mean relative humidity of the basin varies between 70% and 86% and annual rainfall is about 1800 mm with 80% of the total rain occurring in the monsoon season [9]. Rainfall occurrence in the basin is mainly due to the south east monsoon, generally starting from June and ending at September. In this course, the humid monsoon air stream blows from the Bay of Bengal and rises till it meets the Himalaya. Then ultimately, rainfall occurs heavily on some section of the southern Himalayan slopes. It also occurs heavily along the Chure range. As mentioned in a report of Department of Hydrology and Meteorology in Nepal, the area close to the Indian boarder receives about 1500 mm rain in a year. It rises up to 2000 mm at the foot hills of the Chure but it diminishes at the northern part of the Chure. It is also experienced that rainfall reduces due to the rain shadow effect. Furthermore, the rainfall pattern also changes by the orographic effect in this region [3]. For the study, whole of the Bagmati River basin is considered.

2.3. Data

There are three kinds of stations, namely Precipitation, Climatology and Agro-meteorology with elevation ranging from 131 m at Karmaiya, Sarlahi district to 2163 m at Nagarkot, Kathmandu district established at different districts to cover the Bagmati River basin. There are about 30 such stations started earliest from September 1966 at Thankot, Kathmandu to latest started in April 2002 at Nangkhel, Bhaktapur. However, the present study considered only 25 stations due to incomplete time-series data of daily rainfall. The daily rainfall data of them are obtained from the Department of Hydrology and Meteorology, Kathmandu, Nepal for the time-period of January 1981 - December 2008. In order to build a meaningful transfer function in statistical downscaling technique with use of logistic regression method, the rainfall data is aggregated to match better with the large-scale observations obtained from NCEP/NCAR Reanalysis data and GCMs outputs [14].

Table 1. Table for area distribution of stations.

Station	Area (Sq. k.m.)	Weight	Index No	Mean	SD	Station	Area (Sq. k.m.)	Weight	Index No	Mean	SD
1	132.0325	0.0366	1015	5.00	12.25	14	156.6880	0.0435	1073	2.37	8.32
2	171.6377	0.0476	1022	5.06	12.71	15	21.5802	0.0060	1074	3.26	10.78
3	32.7590	0.0091	1029	3.32	8.51	16	25.4970	0.0071	1077	2.29	8.73
4	28.2322	0.0078	1030	4.37	10.72	17	29.4200	0.0082	1079	1.73	7.22
5	35.8642	0.0100	1035	5.45	12.41	18	29.2124	0.0081	1080	1.21	5.72
6	1.8051	0.0005	1038	4.56	11.93	19	25.5804	0.0071	1081	1.68	7.21
7	46.3296	0.0129	1039	4.06	10.35	20	29.1294	0.0081	1082	1.23	6.14
8	22.7268	0.0063	1043	5.06	12.15	21	22.9222	0.0064	1083	1.16	5.58
9	32.0111	0.0089	1049	4.07	10.52	22	1291.4700	0.3583	1117	7.13	20.35
10	34.1230	0.0095	1052	4.16	10.66	23	17.9395	0.0050	1119	3.49	12.11
11	31.6456	0.0088	1059	5.00	11.27	24	282.6790	0.0784	1120	4.07	15.62
12	207.0428	0.0574	1060	4.04	10.83	25	861.5560	0.2390	1121	3.06	12.06
13	34.5486	0.0096	1071	4.38	11.05	Total Area	3604.4400	1	Min	1.16	5.58

Source: DHM, Nepal; Standard Deviation (SD); Period: 1981-2008..

Table **1** shows the descriptive statistics of all the 25 rainfall stations. These statistics are computed from the time-series during the period 1981-2008. This table shows the spatial variation of the daily rainfall across 25 stations within and closer to the basin. Minimum daily mean rainfall is 1.16 mm with standard deviation (SD) of 5.58 mm and maximum daily rainfall is 7.13 mm with SD of 20.35 mm. Therefore, aggregation of the 25 stations daily rainfall data by simple un-weighted average method seems inappropriate. Thus, weighted averaging technique is adopted. To compute weighted average daily rainfall, proportion of area of each station is taken as the weight with respect to the area of the whole basin. This area is computed by Thiessen Polygon. Table **1** shows the area in the second and eight columns. Its proportion as weight is presented in third and ninth columns for each station. By this method, area weighted rainfall is generated for each day of time-series of 1981-2008. The following figure (Fig. **1**) shows the trend of the mean area weighted rainfall (in mm) during 1981-2008 period.

Fig. (**1**) shows gradual increase in mean weighted daily area rainfall during the period 1981-2008 AD. This may be due to several atmospheric variables, for example sea level pressure, temperature, relative humidity, etc. Here, the study aims to develop and use the logistic regression model to forecast the rainfall pattern by use of a binary variable that represents the rainfall occurrence in a particular day. It is assumed that this variable is influenced by several atmospheric predictor variables. These variables are obtained from various general circulation models but their source is limited to the NCEP/NCAR reanalysis data.

2.3.1. Predictors

In climate change impact studies, the usual practice is to use potential predictor variables with their realizations simulated by various climate models like General Circulation Models (GCMs) that project global climatic variables under different emission scenarios with coarse resolutions [15]. Thus, their direct use is not suitable for assessment of impact of climate change at local level. The large-scale data of predictor variables are also available as National Centers for Environmental Prediction (NCEP) reanalysis products [16]. NCEP reanalysis data are available as daily or monthly mean for period 1948 to 2013, with spatial resolution of $2.5^0\times 2.5^0$. Appropriate selection of the climatic variables is guided by the most important skills in a downscaling process [5-6]. Moreover, the choice of predictors could vary from place to place with respect to nature of GCMs outputs and a predictand chosen. However, the following three criteria are considered in the selection process [17]. They are: (1) Predictors are variables of relevance and can be realistically modeled by GCM or Reanalysis data; (2) the transfer function is valid also under altered climatic conditions. But this assumption in principle cannot be proven in advance. The observational record should cover a wide range of variations in the past, ideally with content of all expected future realizations of the predictors; (3) the predictors have the physical relationship with a predictand chosen. Under these criteria, some predictor variables are selected and they are Sea Level Pressure (SLP) at 850 hPa, Geopotential Height (GPH) at 850 hPa, Air Surface Temperature (AST) in Kelvin, U-component of wind (U-W) at 850 hPa, V-component of wind (V-W) at 850 hPa, and Precipitable Water (PW) at 850 hPa. As the present study focuses on only to build a predictive model, only NCEP/NCAR reanalysis data are taken into consideration for the study. Before building a logistic regression model, descriptive analysis of spatial-temporal distribution pattern of all the seven potential predictor variables were analyzed one by one with help of graphs which plotted their 9 gridded mean values against year.

In Fig. (**2**), first upper three line graphs present that there is no significant difference on average GPH across three longitudinal positions 82.5^0E, 85.0^0E and 87.5^0E at the same latitudinal position 25^0N. Similarly, we see that there is no significant difference in average GPH across three longitudinal positions 82.5^0E, 85.0^0E and 87.5^0E at the same

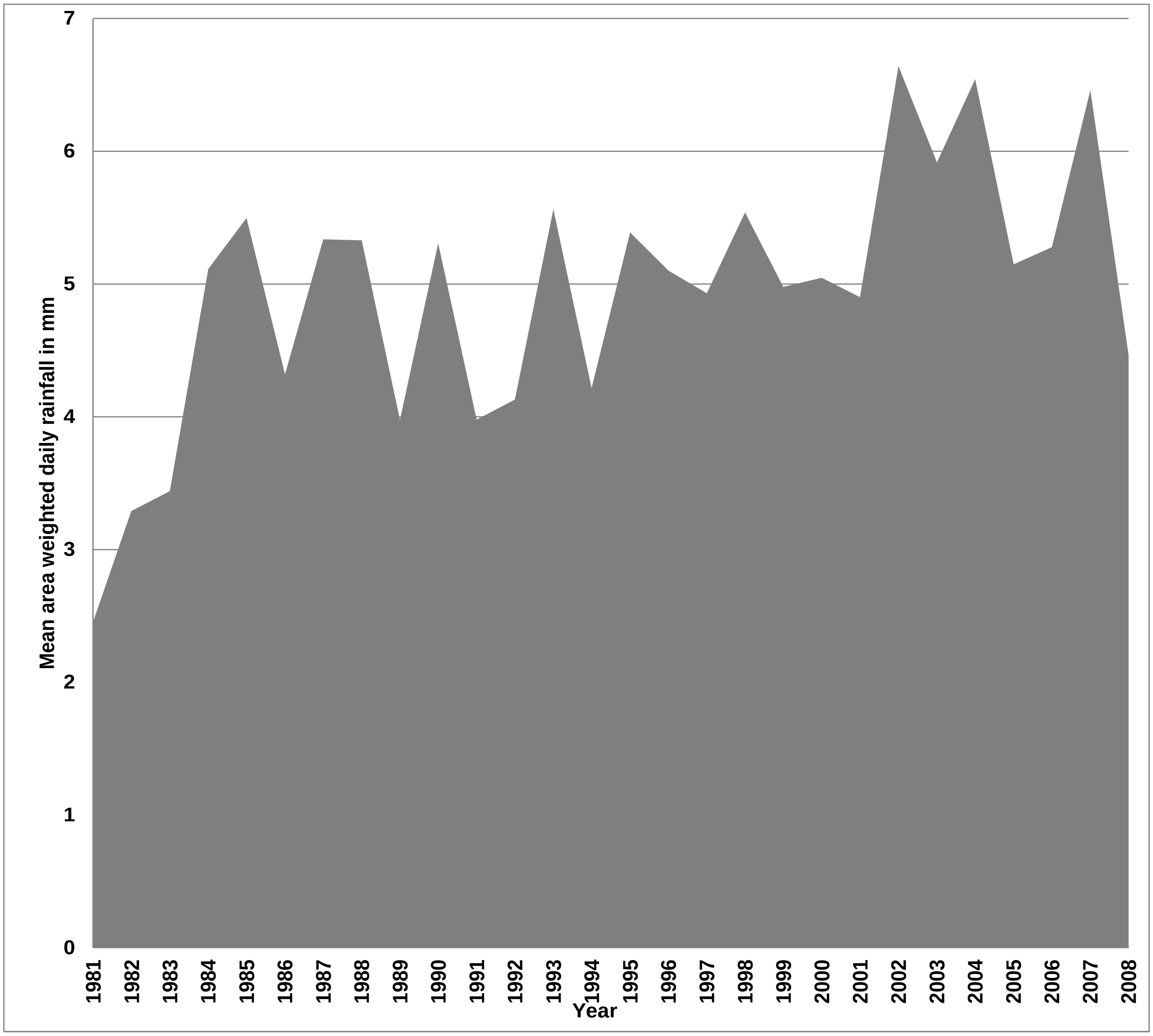

Fig. (1). Average daily area weighted rainfall during 1981-2008.

latitudinal positions 27.5^{0}N and 30^{0}N, respectively. But there seems a larger gap between the first set of three lines and the second and third set of three lines. It means that there is a significant difference in average GPH between the positions at 82.5^{0}E and, at 85.0^{0}E and 87.5^{0}E at three different latitudinal positions 25^{0}N, 27.5^{0}N and 30^{0}N. However, this variation is not so much apparent between latitudinal positions 27.5^{0}N and 30^{0}N. Thus, it can be concluded that on average GPH increases latitudinal-wise from south to north within northern hemisphere with increasing variations. In addition, there is a gradual upward trend across all the nine lines during 1981-2008 periods with distinct oscillations. Thus, it shows that on average there is both spatial and temporal variation in the GPH. Although there is a positive correlation between GPH and rainfall at basin level found in the past studies, it may or may not be possible to infer that the rainfall increases from south to north within northern hemisphere because GPH is a large-scale output with coarse in nature at the basin scale. However, such relation may appear in the result of a model formulation. Similarly, in Fig. (**3**), relative humidity shows somewhat similar behavior as shown by GPH in Fig. (**2**) but their oscillation pattern seems less in amplitude. In addition, there is a very slow upward trend across all nine lines for the period of 1981-2008 year with distinct oscillations.

In Figs. (**4**, **5**), both Sea Level Pressure and Precipitable Water show similar behavior as shown by GPH in Fig. (**2**). But, their oscillation seems less in amplitude too. In addition, there is a very slow upward trend across all nine lines during 1981-2008 periods with distinct oscillations. In Fig. (**6**), U-W shows less spatial variation for the first few years but the variation seems gradually increasing onwards. In Fig. (**7**), V-W shows smaller variation for the last few years. It is clear that two U-W and V-W have different erratic behavior in their spatial distribution. However, in Fig. (**8**), AST shows smooth variation without any erratic oscillation. Therefore, all the seven predictor variables may have some different effects on the rainfall pattern in the study region.

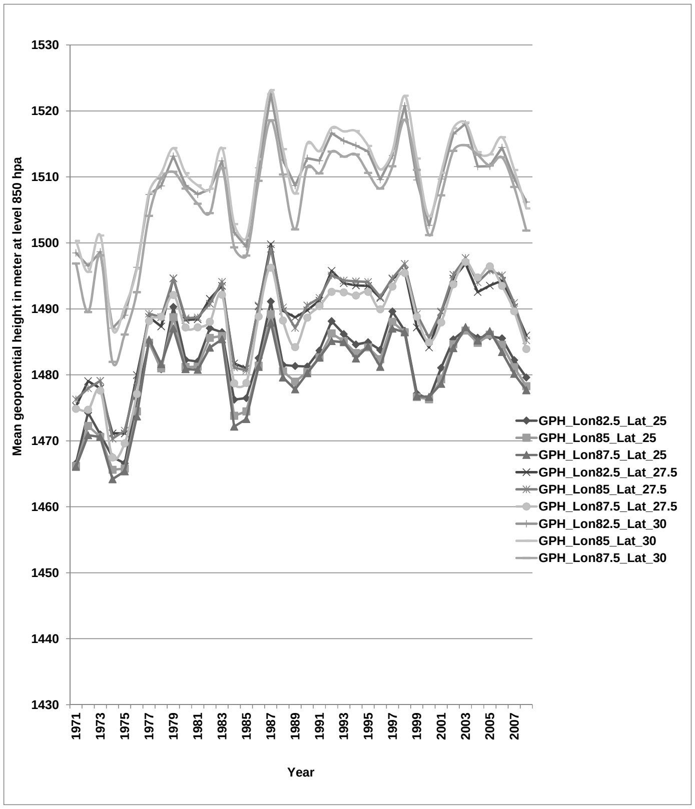

Fig. (2). Spatial and temporal distribution of GPH at 850 hPa.

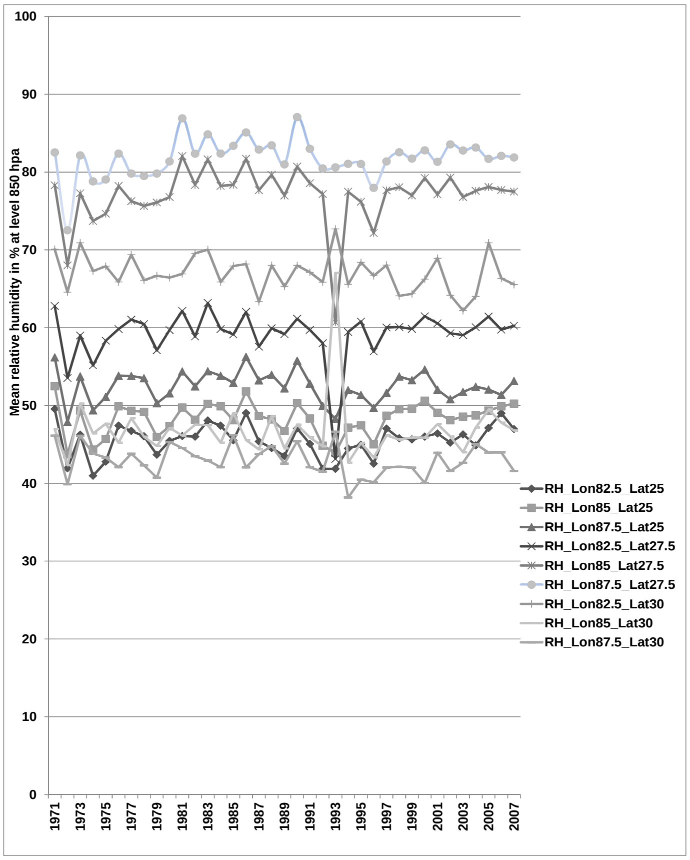

Fig. (3). Spatial and temporal distribution of RH at 850 hPa.

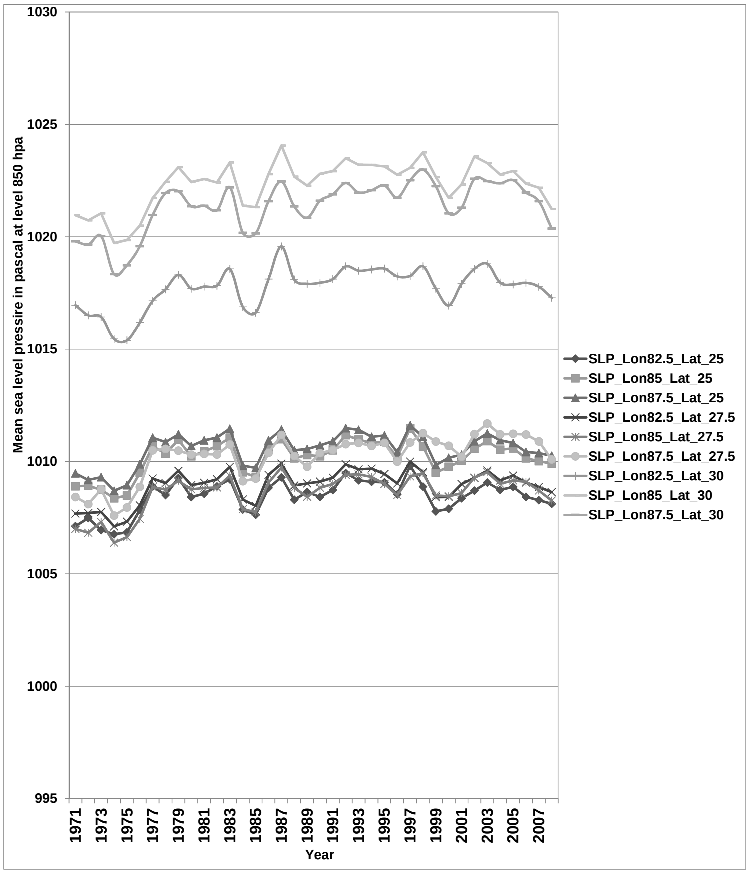

Fig. (4). Spatial and temporal distribution of SLP at 850 hPa.

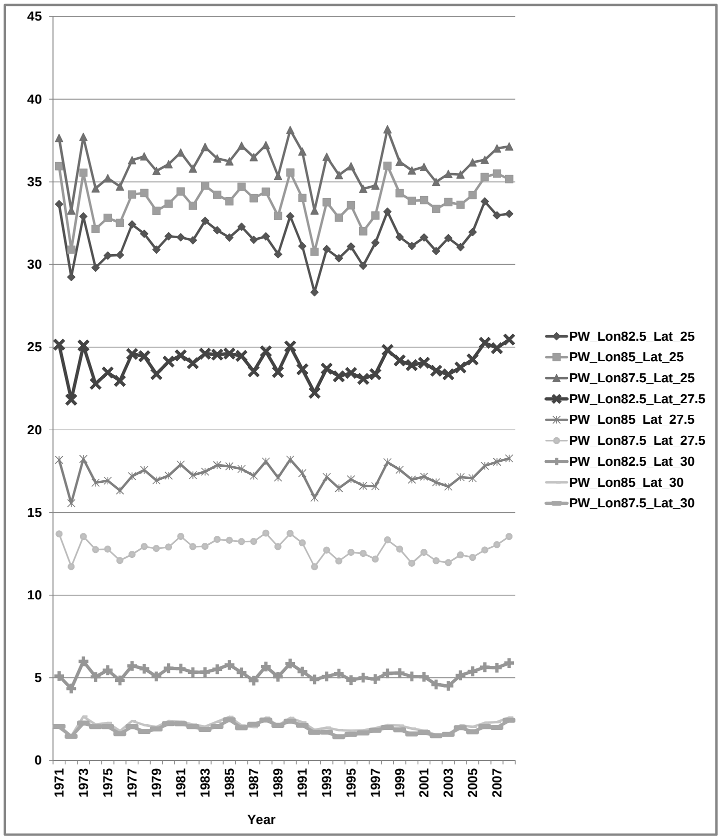

Fig. (5). Spatial and temporal distribution of PW at 850 hPa.

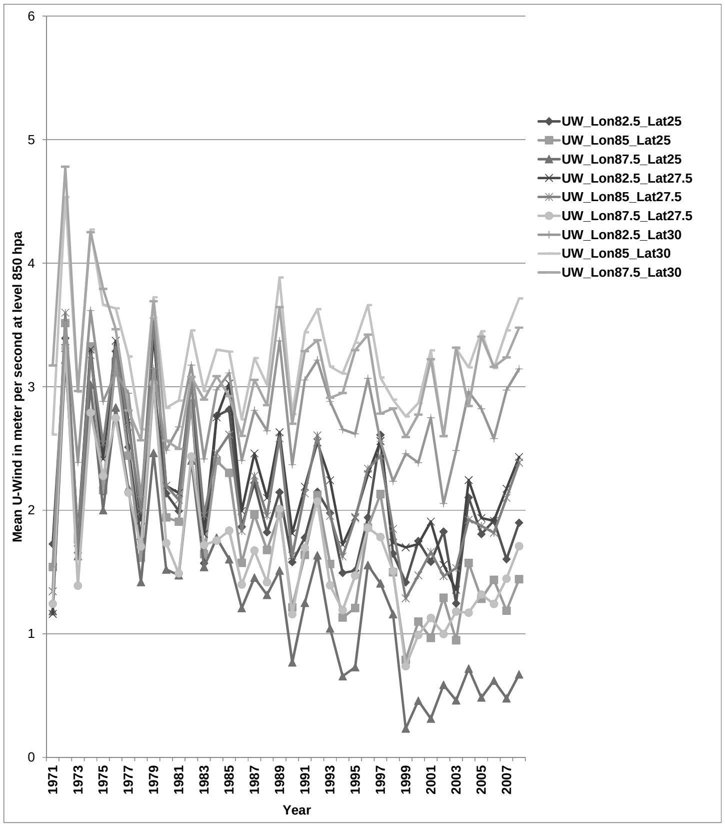

Fig. (6). Spatial and temporal distribution of U-Wind at 850 hPa.

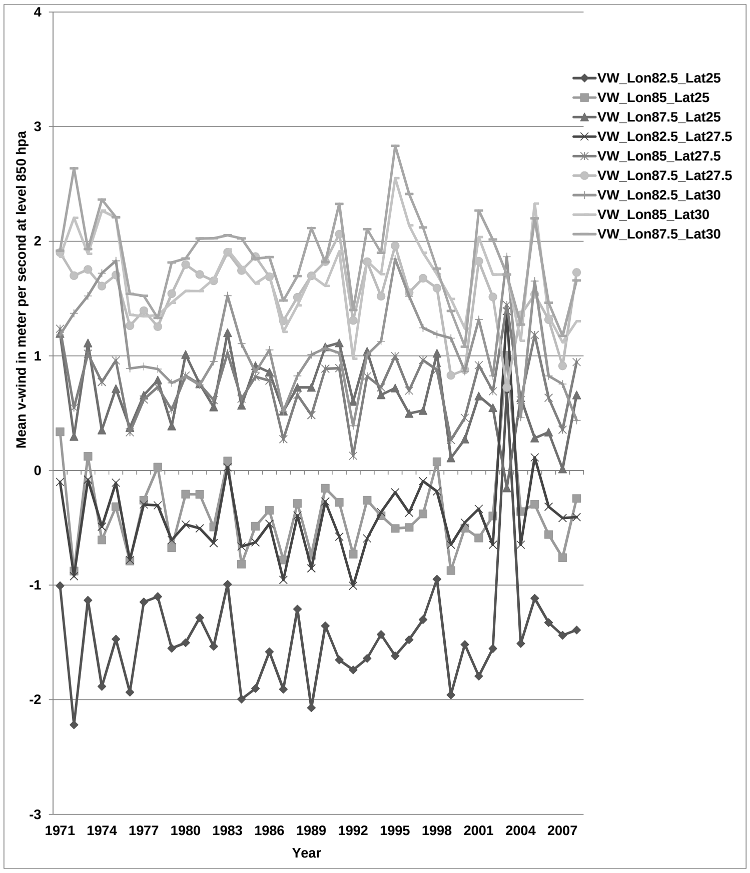

Fig. (7). Spatial and temporal distribution of V-Wind at 850 hPa.

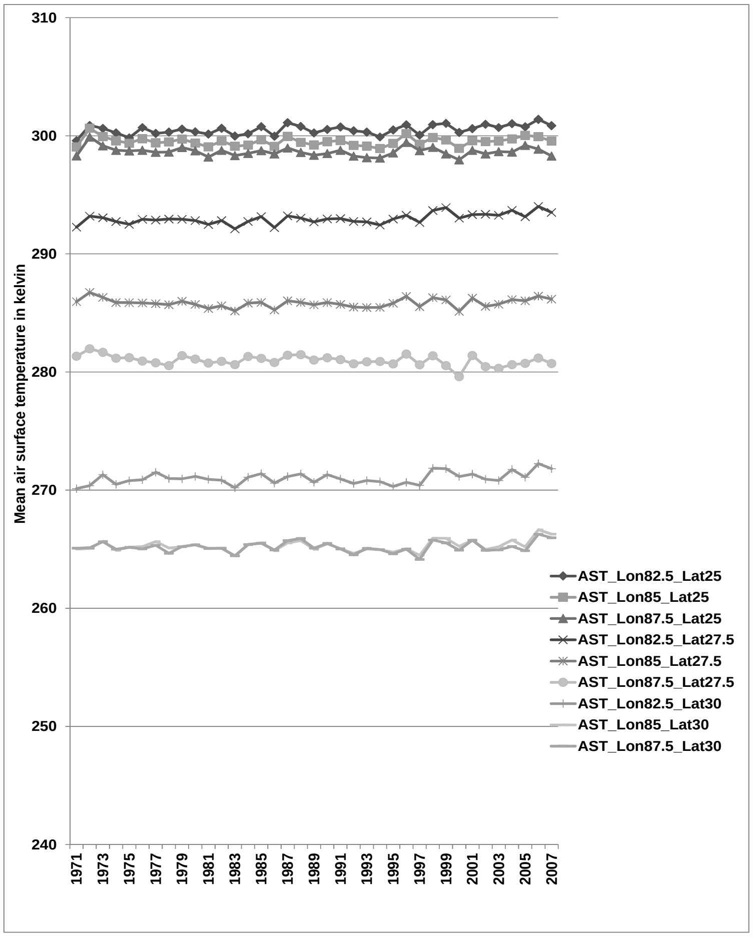

Fig. (8). Spatial and temporal distribution of AST at 850 hPa.

2.3.2. Correlation Between Daily Area Weighted Rainfall and Each of 9 Gridded Predictors

Pearson correlation analysis was performed to examine whether there is a significant correlation between daily area weighted rainfall and a predictor in 9 grid observations. The results (Table **2**) reveal correlations between daily area weighted rainfall (in mm) and each of seven predictor variables for the analysis data (sample size=7305). Correlation coefficient seems changing from one grid to other for every predictor. The strength and direction of the correlation are not fixed across 9 grids. Observed correlations do vary for different predictors across 9 grids and are statistically significant too. However, the correlations are relatively higher for Precipitable Water, Relative Humidity, Geopotential Height and Sea Level Pressure in all of the 9 grids but relatively lower for the remaining predictors and weakest for U-wind in all grids. Spatially, there are small variations in correlation between grids for most of the predictors except for Air Surface Temperature and V-Wind.

Table 2. Correlation table.

Predictor	Lon82.5_ Lat_25 (G1)	Lon85_ Lat_25 (G2)	Lon87.5_ Lat_25 (G3)	Lon82.5_ Lat_27.5 (G4)	Lon85_Lat_ 27.5 (G5)	Lon87.5_ Lat_27.5 (G6)	Lon82.5_ Lat_30 (G7)	Lon85_ Lat_30 (G8)	Lon87.5_ Lat_30 (G9)
Geopotential Height (mm)	-.420**	-.419**	-.410**	-.442**	-.431**	-.415**	-.435**	-.413**	-.389**
Relative humidity (%)	.394**	.416**	.414**	.372**	.362**	.376**	.391**	.461**	.469**
Sea Level Pressure (Pa or Mbar)	-.400**	-.408**	-.406**	-.415**	-.417**	-.417**	-.440**	-.444**	-.428**
Precipitable Water (mm)	.484**	.495**	.498**	.493**	.517**	.528**	.467**	.493**	.512**
Air Surface Temperature (^{0}K)	.147**	.169**	.214**	.305**	.376**	.419**	.441**	.449**	.448**
U-Wind (ms^{-1})	.054**	-.060**	-.106**	.144**	.017	-.061**	.119**	-.027*	-.131**
V-Wind (ms^{-1})	.175**	.265**	.352**	.303**	.381**	.416**	.119**	.193**	.237**

*Correlation is significant at the 0.05 level (2-tailed); **Correlation is significant at the 0.01 level (2-tailed); Units of measurements are provided in brackets.

2.3.3. Principal Component Analysis (PCA)

Principal component analysis (PCA) is performed for each of predictors GPH, RH, SLP, PW, AST, U-W and V-W based upon analysis period 1981-2008 spanning 7305 days. A number of components are determined by the eigenvalue 1 or more as a threshold value under the varimax rotation method. The results (Table **3**) are as follows.

- GPH has 1 component represented by Geopotential Height_1 (GPH1) with all 9 loadings higher than 0.94 and explains94.98 % of the total variation.
- RH has 2 extracted components represented by Relative Humidity_1 (RH1) which has higher loadings for the first 6 gridded areas (RH1_G1 - RH1_G6) of loadings (> 0.65) and Relative Humidity_2 (RH2) which has higher loadings (> 0.70) for the last 3 gridded areas (RH2_G7-RH2_G9), both together explaining 85.24 % of the total variation.
- SLP has 1 component, represented by Sea Level Pressure_1 (SLP1) with all 9 gridded items of loadings higher than0.93 and explains 93.53 % of the total variation.
- PW has also 1 component represented by Precipitable Water_1 (PW1) with all 9 gridded items of loadings higher than 0.95 and explains 94.08 % of the total variation.
- AST has 2 components represented by Air Surface Temperature_1 (AST1) with higher loadings (>0.80) for the last 5 gridded areas (AST1_G5-AST1_G9) and Air Surface Temperature_2 (AST2) with higher loadings (> 0.74) for the first 4 gridded areas (AST2_G1-AST2_G4), both together explaining 96.95 % of the total variation.
- U-W has 2 components represented by U-WIND_1 (UW1) with higher loadings (>0.69) for the first 6 gridded areas (UW1_G1-UW1_G6) and U-WIND_2 (UW2) with higher loadings (>0.85) for the last 3 gridded areas (UW2_G7-UW2_G9), both together explaining 90.15 % of the total variation.
- V-W has 3 components represented by V-WIND_1 (VW1) with higher loadings (>0.80) to the last 3 gridded areas (VW1_G7-VW1_G9), V-WIND_2 (VW2) with higher loadings (>0.68) to 3 gridded areas (VW2_G3, VW2_G5-VW2_G6) and V-WIND_3 (VW3) with higher loadings (>0.71) to 3 gridded areas (VW3_G1-VW3_G2 and VW3_G4), all together explaining 95.34 % of the total variation.
- In PCA each component for any predictor has some distinct spatial characteristics. Loadings in extracted components obtained from analysis period are used for validation period for all the predictors. Results of PCA with loadings are shown in Table **3**.

2.4. Methods

Statistical downscaling model is based on logistic regression model which shows a functional relationship between a binary response and a pool of predictors which accounts seven atmospheric variables as defined above. The daily time-series is available for 1981-2008 is split into two parts, first part for the purpose of analysis spanning 20 years (1981-2000) and second part for validation of the fitted model spanning 8 years (2001-2008) [18]. The logistic regression model is applied by employing following two important criteria: a) for a given pool of predictors, the objective is to estimate the probability of rainfall occurrence, and b) to assess the magnitude of the effects of atmospheric variable on the rainfall occurrence by odds ratio (OR). Here, a response variable, Y_i is defined as a day with rainfall

Table 3. Spatial grid loadings of predictors.

Predictor	Lon82.5_ Lat_25 (G1)	Lon85_ Lat_25 (G2)	Lon87.5_ Lat_25 (G3)	Lon82.5_ Lat_27.5 (G4)	Lon85_ Lat_27.5 (G5)	Lon87.5_ Lat_27.5 (G6)	Lon82.5_ Lat_30 (G7)	Lon85_ Lat_30 (G8)	Lon87.5_ Lat_30 (G9)
GPH1	0.961	0.973	0.975	0.991	0.997	0.989	0.977	0.964	0.944
RH1	0.863	0.898	0.853	0.883	0.857	0.656	0.482	0.313	0.284
RH2	0.347	0.350	0.338	0.291	0.311	0.508	0.730	0.936	0.918
SLP1	0.969	0.975	0.977	0.985	0.987	0.973	0.953	0.952	0.933
PW1	0.971	0.974	0.961	0.985	0.994	0.979	0.955	0.957	0.952
AST1	0.291	0.337	0.421	0.643	0.804	0.883	0.923	0.944	0.943
AST2	0.934	0.937	0.854	0.741	0.573	0.438	0.361	0.322	0.314
UW1	0.952	0.964	0.896	0.793	0.791	0.696	0.330	0.242	0.146
UW2	0.114	0.193	0.251	0.442	0.581	0.621	0.856	0.964	0.925
VW1	0.027	-0.008	-0.024	0.372	0.441	0.402	0.808	0.965	0.896
VW2	0.232	0.674	0.906	0.321	0.688	0.897	-0.123	0.151	0.361
VW3	0.944	0.712	0.332	0.847	0.543	0.125	0.498	0.156	-0.116

GPH1 = 0.961(GPH1_G1) + 0.973 (GPH1_G2) + 0.975 (GPH1_G3) + 0.991(GPH1_G4) + 0.997(GPH1_G5) + 0.989(GPH1_G6) + 0.977(GPH1_G7) + 0.964(GPH1_G8) + 0.944(GPH1_G9) and so on.

occurrence if the area weighted rainfall is more than 1 mm per day in the study area. Consequently, Y_i assumes two possible values 1 and 0 with probability of rainfall occurrence (p_i) (for wet day), and 1- p_i, probability of no rainfall (for dry day), respectively. Then, odds of success is expressed as $\left(\frac{p_i}{1-p_i}\right)$ which is the ratio of the probability of success (rainfall occurrence) to the probability of failure (rainfall non-occurrence). The logistic regression model [19] is expressed:

$$Log\left(\frac{p_i}{1-p_i}\right)=\beta_0+\sum_{i=1}^{k}\beta_i x_i \qquad (1)$$

p_i can be computed by:

$$p_i=\frac{e^{\beta_0+\sum_{i=1}^{k}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{k}\beta_i x_i}}=\frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{k}\beta_i x_i\right)}} \qquad (2)$$

where β_0 and β_i (coefficient attached with the ith regressor) are the unknown model parameters to be estimated from observed data and x_i is ith regressor of the model.

Odds ratio: Odds ratio (OR) plays an important role in logistic regression. For multiple predictors, OR associated with the regressor x_j can be interpreted as the increase (or decrease) in probability of success associated with unit increase in the predictor assuming values of the remaining predictors are constant. The computational formula of OR associated with x_i is:

$$\frac{Odds\left(x_i+1\right)}{Odds\left(x_i\right)}=e^{\beta_i} \qquad (3)$$

OR>1, if increase in the value the predictor variables increases the chance or risk of occurrence of success (rainfall). Conversely, if OR<1, if increase in the value the predictor variables decreases the chance or risk of occurrence of the success. Finally, if OR=0 then values of the predictor variable have equal influence on success or failure i.e. a day is with or without rainfall.

2.4.1. Model Fit and Diagnostics

For model fit, initially stepwise method (forward and backward likelihood ratio tests) is used for the selection of statistically significant predictors and goodness of fit was assessed by Hosmer-Lemeshow test and fitted model is accepted if p value is not less than 0.1 and also by omnibus and Negalkerke R^2. The fitted model was checked for multicollinearity and some variables were dropped in the presence of high correlation between predictors which produced high variance inflation factors (VIF). Deviance residual analysis is used in detecting the outlier(s) (observations not well explained by the model) against linear predictor and the residuals falling within ±3 were not regarded as outliers. In order to assess the capability of accurate classification of a model, Receiver Operating Characteristic (ROC) is used. Its common measure is area under curve (AUC) with values ranging between 0 and 1. The bigger value suggests a better overall performance of the model. The ROC curve is obtained by plotting sensitivity of all values against its specificity.

3. RESULTS

In the formulation of logistic regression model, the forward and backward model selection methods produced six different logistic models. But the sixth model produced by them were the same. The predictors included in the model were Geopotential Height_1, Relative Humidity_1, Relative Humidity_2, Sea Level Pressure_1, Precipitable Water_1, Air Surface Temperature_1 and V-Wind_2 with all significant coefficients (p<.001). However, a goodness-of-fit test shown by the Hosmer-Lemeshow test (χ^2 =32.681, df =8 and p<.001) and the omnibus test (Deviance =-2LL ~ χ^2 = 5085.339 with df=7 and p<.001) both have a significant p-value with Nagelkerke R Square 0.678. It meant that the model had a good fit under the omnibus test or Deviance test criteria but it is not supported by the Hosmer-Lemeshow test. Results of these regressions are not shown for brevity. When a multicollinearity test was performed with the predictand as a continuous variable, the results showed multicollinearity problem with high VIFs associated with several predictors. Thereafter, two predictors namely GPH1 and PW1 were excluded as predictor variables (since they produced higher VIFs comparatively). When the logistic regression model was rerun, the Hosmer-Lemeshow test showed acceptable goodness of fit results with insignificant p value just greater than 0.1. The analysis result is shown in Table **4**.

The fitted model showed the following:

- Omnibus test with Deviance statistic is statistically significant with p value <0.001.
- Hosmer-Lemeshow test is statistically insignificant with p value > 0.10.
- Wald statistics for all the predictor variables are significant with p value < 0.001 which meant that RH1, RH2, SLP1, AST1, AST2 and VW2 have significant effects on whether a day is wet or dry. Furthermore, the 95 percent confidence interval for each predictor also does not include 1.
- All VIFs have values less than 5 which rules out the presence of multicollinearity.
- The receiver operating characteristic (ROC) curve (Fig. **9**) has area under curve of 0.931. ROC curve above 0.80 area under curve indicates that the model has a good predictive ability.

3.1. Residual Analysis

Fig. (**10**) shows a graph of standardized deviation residuals (SDR) against predicted value of linear predictor. There are five points above 3 SDR. On examining the influence of these points as outliers show no significant influence on the standard error of the coefficients of the predictors. Therefore, these points are retained in the fitted model. The diagnosis of overdispersion is an important concept in the analysis of discrete data. Many a time data admit more variability than expected under the assumed distribution resulting in what is known as everdispersion. If overdispersion is present in a dataset, the estimated standard

Table 4. Estimated model coefficients.

Predictors	B	S.E.	Wald	df	Sig.	OR	95% C.I. for OR		Collinearity Statistics	
							Lower	Upper	Tolerance	VIF[a]
Relative Humidity_1	1.431	0.069	434.18	1	.000	4.184	3.657	4.787	0.44	2.275
Relative Humidity_2	1.285	0.071	325.79	1	.000	3.614	3.143	4.155	0.273	3.665
Sea Level Pressure_1	-0.823	0.084	95.81	1	.000	0.439	0.373	0.518	0.211	4.737
Air Surface Temperature_1	-0.274	0.078	12.45	1	.000	0.76	0.653	0.885	0.216	4.621
Air Surface Temperature_2	0.326	0.066	24.52	1	.000	1.385	1.217	1.576	0.397	2.518
V-WIND_2	0.358	0.056	40.37	1	.000	1.431	1.281	1.598	0.439	2.28
Constant	-0.86	0.044	381.32	1	.000	0.423				
Hosmer-Lemeshow test : χ^2 =13.153, df = 8 and p = 0.107 (> 0.10)										
Null Deviance: 9821.236, df = 7304										
Residual Deviance: 4760.249, df = 7298										
Deviance: 5060.987, df = 6 and p =.000 (<.001)										

Note: [a]VIF is a result of multiple regression of daily weighted area rainfall in mm with the above listed predictor in Table **3**.

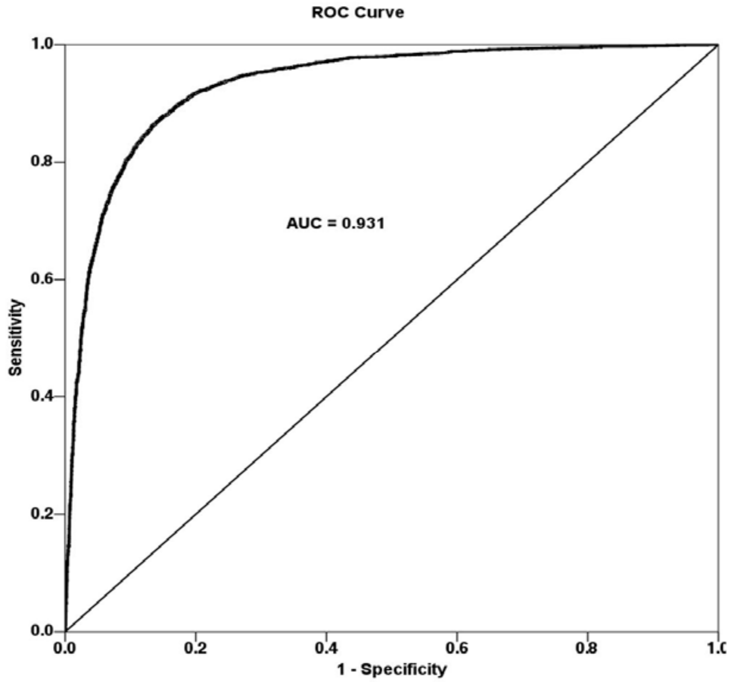

Fig. (9). ROC curve.

errors and test statistics of the overall goodness-of-fit will be distorted. However, the model shows deviance divided by its degrees of freedom equal to 0.652 (<1.0) (and for Pearson Chi-square, it is 1.062). Hence, there is absence of overdispersion in the model. This statistic supports that the model has a good property of fit to the data. When checking linear relationship between log-odds and linear predictor with the logistic model, it is found that the coefficient (-0.024) with standard error (0.012) of the square of linear predictor is insignificant at 1 percent level. This shows that there is no violation of assumption that there is a linear relation between log-odds and linear predictor in the model.

3.2. Validation of the Model

The fitted model is validated considering the estimates obtained from the analysis data (1981-2000) and forecasting

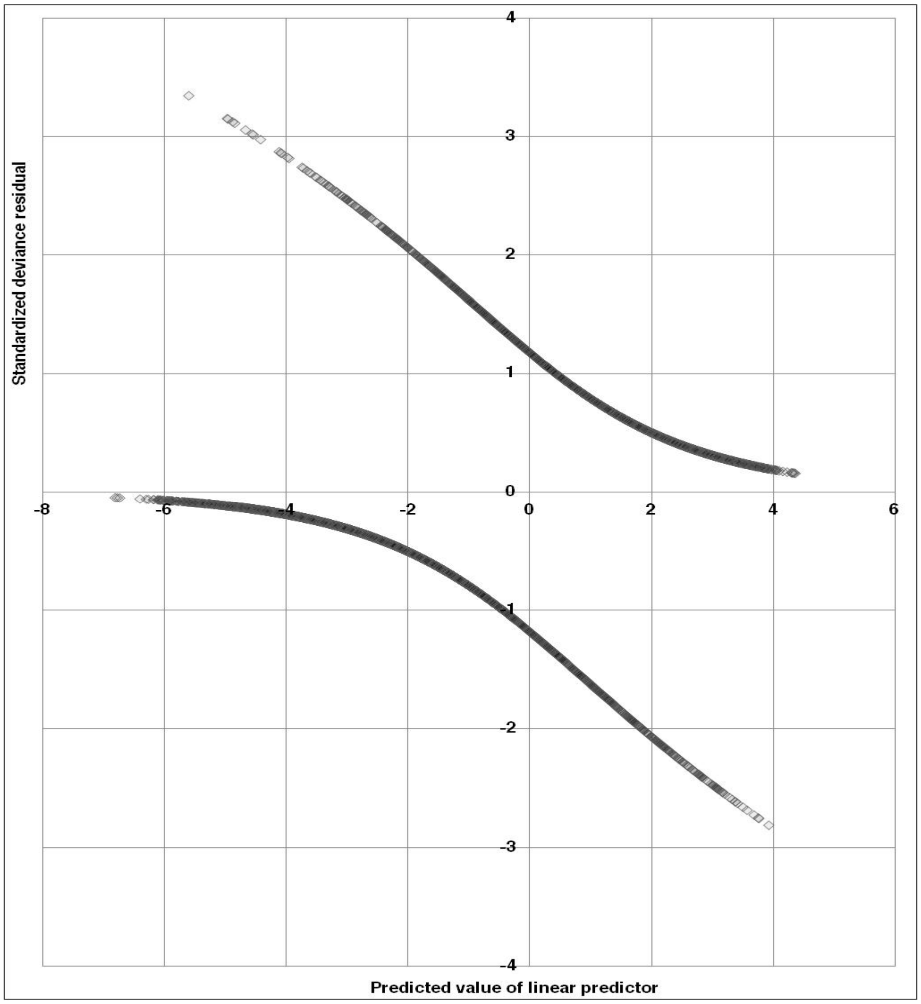

Fig. (10). Standardized deviance residual plot.

the dependent variable for the observed data of predictor variables for validation period (2001-2008). The classification table (Table **5**) reveals that the percentage correct prediction for analysis period is 86.4 (> 0.70) percent whereas for validation period, it is 86.1 percent and demonstrates that the fitted model has high percentage predictive power. Therefore, all the diagnostic tests and validation test confirm that the developed logistic model is indeed a good model with high predictive ability for prediction of a wet or dry day in the study area, the Bagmati River basin in Nepal.

4. DISCUSSION AND CONCLUSION

The results on OR (Table **4**) shows that the first and second principal components associated with relative humidity (RH1 and RH2) are key predictors of the fitted model with ORs equal to 4.184 and 3.614, respectively. The first principal component has higher loadings to the first 6 grids whereas the second component has higher loadings to the last 3 grids. Other predictors in the fitted model with descending values of OR are the V-WIND _2 (OR 1.431), Air Surface Temperature_2 (OR 1.385), Air surface temperature_1 (OR 0.76) and Sea Level Pressure_1 (OR 0.439). Sea Level Pressure_1 has the smallest OR and less than 1, indicating less detective power for a wet day but has more power for detecting a dry day.

If we examine the effect of air surface temperature on rainfall, we find that increase in temperature increases the chance of rainfall as shown by the first principal component associated with temperature. Conversely, the second principal component shows that increase in temperature decreases the chance of rainfall with OR less than 1. This is due to the fact that the two components have higher loadings to different sets of grids or spatial locations. Air Surface Temperature_2 mostly belongs to latitude of 25^0 N along with 82.5 ^{0}-87.5^0 E longitude whereas Air Surface Temperature_1 belongs to latitude of 27.5^0 -30^0 N along with 82.5 ^{0}-87.5^0 E longitude (Table **3**). The former predictor (AST2) shows higher loadings to lower part (southern) of Nepal, or Terai region, and the latter predictor (AST1) shows higher loadings to upper/middle part (northern) part of Nepal, or Hill or Himalayan region. According to the report of Department of Hydrology and Meteorology in Nepal, the area close to the Indian boarder receives about 1500 mm rain in a year. This fact verifies that the rainfall happens more in Terai than in Himalayas. Khalil *et al* (1993) had developed relationship between Precipitation and Temperature over the 80-year period from 1905 to 1984 at nearly 1000 stations in United States. He found that over most of the United States,

Table 5. Observed versus predicted classification table.

Frequency of Observed Days		Frequency of Predicted Days					
		Analysis Period			Validation Period		
		Day		Percent Correct	Day		Percent Correct
		Dry	Wet		Dry	Wet	
Day	Dry day	3939	458	89.6	1498	150	90.9
	Wet day	538	2370	81.5	257	1017	79.8
Overall percent correct				86.4			86.1

summer precipitation and temperature were negatively correlated with indication of warm summers tended to be dryer in the central and southern Great Plains and a significant positive correlation between them over the area south of the Great Lakes covering the eastern portion of the Corn Belt in winter [20]. This supports the fact that the relation between the precipitation and temperature may change with season or place. A study by Trenberth *et al.* (2005) found negative correlation between precipitation and surface temperature over land during summer and positive correlation at high latitudes in winter. He also added that ocean conditions drive the atmosphere with higher surface air temperature positively associated with precipitation [21].

Examination of the effect of sea level pressure on rainfall revealed that the first principal component associated with sea level pressure (SLP1) shows negative relation with rainfall with OR less than 1 (0.439). The work of Gerapetritis H, (1999) reported that there was positive correlation between relative humidity and precipitation and negative correlation with sea level pressure on using the FRH/FRHT data [22]. Roy *et al.* (2012) reported in his logistic regression model that there was positive effect of afternoon relative humidity, negative effect of maximum temperature and positive effect of minimum temperature with the precipitation event [23]. Filho *et al.* (2014) found that the rainfall pattern was highly positively associated with relative humidity, maximum temperature and V-Wind component in the logistic regression model used in the northern Brazil [24]. Kutiel *et al.* (2001) shows that relationship between rainfall in Turkey and the regional sea level pressure is large in winter and non-existing in summer. Pressure patterns associated with dry conditions, showed usually positive departure, whereas, pressure associated with wet conditions showed negative SLP departures. Examination of the effect of V-Wind on rainfall showed positive relation with rainfall with OR 1.431 [25]. Also, Prasad *et al.* (2010) demonstrated that the rainfall had positive correlation with V-Wind when the logistic regression model was used to relate them on the basis of realizations obtained for whole India [12]. Babel *et al.*, (2013) developed a statistical downscaling model and found that there was positive association between precipitation and mean sea level pressure and relative humidity at 850 hPa on the basis of monthly NCEP data of during the period 1961-2001 [9]. The analysis of these results depict that there are similar scenario of rainfall pattern with aforementioned and selected predictors. But the outputs may differ in sign and magnitude for some predictor like air surface temperature depending the space or time or seasonality.

Logistic regression model taken as a predictive statistical downscaling model on basis of NCEP/NCAR outputs can be used to forecast a day as wet day or dry day with definite probability projection. The relative humidity with two components is a key variable in identifying a day as wet day or dry day during year. Air surface temperature with its different components state the rainfall pattern differently by projecting a day as wet day or dry day in the lower part (Terai region) and upper part (mountain or hill region). The logistic model does not consider the seasonality effect, which may change the rainfall pattern under the influence of such selected predictor variables. The outcome of this study suggests in developing a model with extreme rainfall and a number of wet days with consideration of seasonality.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors are grateful and thank University Grants Commission, Sanothimi, Bhaktapur for providing PhD scholar grants for conducting the research work. The authors acknowledge NCEP / NCAR for making Reanalysis data available in net for free downloading and Department of Hydrology and Meteorology, Kathmandu for providing rainfall data.

REFERENCES

[1] Wigena AH, Djuraidah A. Quantile regression in statistical downscaling to estimate extreme monthly rainfall. Sci J Appl Math Stats 2014; 2(3): 66-70.

[2] Hastenrath S. Exploring the Climate Problems of Brazil's Nordeste: a review. Climate Change 2011; 112(2): 243-51.

[3] Mishra BK, Herath S. Climate projections downscaling and impact assessment on precipitation over upper Bagmati River basin, Nepal. 3rd International Conference on Addressing Climate Change for Sustainable Development through Up-Scaling Renewable Energy Technologies; Kathmandu, RETRUD-11Conference Proceedings, Nepal, 2011.

[4] Wigena AH. Modeling of statistical downscaling using projection pursuit regression for forecasting monthly rainfall. Indonesia: Bogor Agricultural University 2006.

[5] Cavazos T, Hewitson BC. Performance of NCEP-NCAR reanalysis variables in statistical downscaling of daily precipitation. Climate Res 2005; 28: 95-107.

[6] Hewitson BC, Crane RG. Climate downscaling: techniques and application. Climate Res 1996; 7: 85-95.

[7] Chaulagain NP. Impacts of climate change on water resources of Nepal: The physical and socioeconomic dimensions. Germany: University of Flensburg 2006.

[8] Chaudhary P, Aryal KP. Global warming in Nepal: challenges and politics imperatives. J Forest Livelihood 2009; 8(1): 4-13.

[9] Babel MS, Bhusal SP, Wahid SM, Agarwal A. Climate change and water resources in the Bagmati River basin, Nepal. Theoret Appl Climatol 2013; doi: 10.1007/s00704-013-0919-4

[10] Shrestha S, Khatiwada M, Babel MS, Parajuli K. Impact of climate change on river flow and hydropower production in Kulekhani hydropower project of Nepal. Environ Process 2014; 1: 231-50.

[11] Leith N. Using generalised linear models to stimulate daily rainfall under scenarios of climate change. Technical Report. London, UK: University College London 2005.

[12] Prasad K, Dash SK, Mohanty UC. A logistic regression approach for monthly rainfall forecasts in meteorological subdivisions of India based on DEMETER retrospective forecasts. Int J Climatol 2009; 30: 1577-88.

[13] Li X, Sailor D. Application of tree structured regression for regional precipitation prediction using general circulation model output. Climate Res 2000; 16: 17-30.

[14] Simon T, Hense A, Su B, Jiang T, Simmer C, Ohlwein C. Pattern-based statistical downscaling of East Asian Summer Monsoon precipitation. Tellus A 2013; 65: 197-49. http://dx.doi.org/10.3402/tellusa.v65i0.19749.

[15] Kalnay E, Kanamitsu M, Kistler R, *et al*. The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 1996; 77: 437-71.

[16] Tripathi S, Srinivas VV, Najundiah RS. Downscaling of precipitation for climate change scenarios: a support vector machine approach. J Hydrol 2006; 330: 621-40.

[17] Benestad RE, Hanssen-Bauer I, Chen D. Empirical- Statistical Downscaling. Singapore: World Science Publishers 2008; p. 215.

[18] Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied Linear Statistical Models. New Delhi, India: McGraw-Hill Education India Edition Pty Ltd 2013.

[19] Agresti A. Categorical Data Analysis. New Jersey, USA: John Wiley & Sons Inc 2002.

[20] Zhao W, Khalil MAK. The relationship between precipitation and temperature over the contiguous United States. J Climate 1993; 6:1232-6.

[21] Trenberth KE, Shea DJ. Relationships between precipitation and surface temperature. Geophys Res Let 2005; 32: L14703. doi: 10.1029/2005GL022760.

[22] Gerapetritis H. A probability of precipitation equation for Columbia, South Carolina derived from logistic regression. East Reg Technic Attachment 1999; 99(1): 5.

[23] Imon A, Roy M, Bhattacharjee S. Prediction of rainfall using logistic regression. Pak J Stats Operat Res 2012; 8(3): 655-67.

[24] Filho WLFC, Lucio PS, Spyrides MHC. Precipitation extremes analysis over the Brazillian Northeast *via* logistic regression. Scient Res 2014; 4: 53-9.

[25] Kuitel H, Hirsch-Eshkol TR, Türkeş M. Sea level pressure patterns associated with dry or wet monthly conditions in Turkey. Theoretic Appl Climatol 2001; 69: 39-67.

Received: March 17, 2015 Revised: June 3, 2015 Accepted: June 8, 2015