Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Hydrology 319 (2006) 357–382
Journal of Hydrology
www.elsevier.com/locate/jhydrol

Uncertainty analysis of statistical downscaling methods

Mohammad Sajjad Khan*, Paulin Coulibaly, Yonas Dibike

Department of Civil Engineering/School of Geography and Geology, McMaster University, Hamilton, Ont., Canada L8S 4L7

Received 25 June 2004; revised 24 May 2005; accepted 28 June 2005

Abstract

Three downscaling models namely *Statistical Down-Scaling Model* (SDSM), *Long Ashton Research Station Weather Generator* (LARS-WG) model and *Artificial Neural Network* (ANN) model have been compared in terms various uncertainty assessments exhibited in their downscaled results of daily precipitation, daily maximum and minimum temperatures. In case of daily maximum and minimum temperature, uncertainty is assessed by comparing monthly mean and variance of downscaled and observed daily maximum and minimum temperature at each month of the year at 95% confidence level. In addition, uncertainties of the monthly means and variances of downscaled daily temperature have been calculated using 95% confidence intervals, which are compared with the observed uncertainties of means and variances. In daily precipitation downscaling, in addition to comparing means and variances, uncertainties have been assessed by comparing monthly mean dry and wet spell lengths and their confidence intervals, cumulative frequency distributions (cdfs) of monthly mean of daily precipitation, and the distributions of monthly wet and dry days for observed and downscaled daily precipitation. The study has been carried out using 40 years of observed and downscaled daily precipitation, daily maximum and minimum temperature data using NCEP (National Center for Environmental Prediction) reanalysis predictors starting from 1961 to 2000. The uncertainty assessment results indicate that the SDSM is the most capable of reproducing various statistical characteristics of observed data in its downscaled results with 95% confidence level, the ANN is the least capable in this respect, and the LARS-WG is in between SDSM and ANN.

Keywords: SDSM; LARS-WG; ANN; Downscaling; Uncertainty analysis

1. Introduction

Statistical downscaling methods such as multiple linear regression, nonlinear regression (e.g. artificial neural networks) and stochastic weather generators are easier and less costly to implement as compared to dynamical downscaling technique which requires limited-area models (LAMs) or regional climate models (RCMs). Thus, statistical downscaling methods are the most largely used in anticipated hydrologic impact studies under climate change scenarios. However, no study has specifically focused on assessing uncertainty in downscaling results due to different statistical downscaling methods. The goal of this study is to compare three statistical downscaling models namely *Statistical Down-Scaling Model* (SDSM), *Long Ashton Research Station Weather*

* Corresponding author. Fax: +1 905 529 9688.
E-mail address: khanms3@mcmaster.ca (M.S. Khan).

0022-1694/$ - see front matter © 2005 Elsevier Ltd All rights reserved.
doi:10.1016/j.jhydrol.2005.06.035

Generator (LARS-WG) and *Artificial Neural Network* (ANN) by quantifying uncertainties in their downscaled results using various uncertainty measures.

The uncertainties in downscaling result from (a) the concept on which the downscaling models are based, and (b) from the data used. In this study, different uncertainty assessment methods are used along with robust statistical testing of the first and second moments of the observed and downscaled data. The most commonly used statistical methods for assessing model uncertainty include analyzing the statistical properties of the model errors (Kaleries et al., 2001; Yar and Chatfield, 1990; Coleman and Steele, 1989; Wahl, 2004; Loukas et al., 2002; Montanari and Brath, 2004), sometimes along with the confidence intervals for the estimates of means and variances of model results (Caruso, 1999; Ang and Tang, 1975). While the distribution of data is non-Gaussian, analysis is further extended beyond comparison of means and variances (Wilby et al., 1998). As such, in the current study, in case of uncertainty assessment in daily precipitation, in addition to comparing means and variances, distributions of monthly mean daily precipitation, monthly dry and wet days as well as mean dry and wet-spell statistics along with their confidence intervals have been considered. Here, the uncertainty assessment includes both the downscaling model results and the corresponding observations. The analysis aims to provide a comprehensive comparison of the SDSM, LARS and ANN downscaling models in term of uncertainties in their downscaled results using NCEP reanalysis datasets. The paper is organized as follows: a brief description of the study area and data is provided in Section 2, followed by a description of three downscaling experiments in Section 3. In Section 4, the uncertainty analysis methods are presented. In Section 5, the three downscaling models have been compared based on the results of uncertainty assessments. Finally, in Section 6, a comprehensive summary and conclusions are provided.

2. Study area and data

The study area selected for this research is a sub-basin called Chute-du-diable with an area of 9700 km^2, located in the Saguenay-Lac-Saint Jean watershed in northern Quebec in Canada (Fig. 1). Two meteorological stations inside that sub-basin namely Chute-du-Diable (CDD) and Chute-des-Passes (CDP) are used for the downscaling experiments. For each station, forty years (1961–2000) daily precipitation, daily maximum and minimum temperature records have been used as predictands. Observed large-scale NCEP (national centre for environmental prediction) reanalysis atmospheric variables (Table 1) for the same time period (1961–2000) have been used as predictors (Kistler, et al, 2001).

3. Downscaling experiments

The first downscaling model is a multiple regression based method and is referred to as *Statistical Down-Scaling Model* (SDSM) (Wilby, et al., 2002). During downscaling with the SDSM, a multiple linear regression model is developed between a few selected large-scale predictor variables and local scale predictands such as temperature and precipitation. The parameters of the regression equation are estimated using the efficient dual simplex algorithm. Large-scale relevant predictors (Table 1) are selected using correlation analysis, partial correlation analysis and scatter plots, and also considering physical sensitivity between selected predictors and predictand for the site in question. Precipitation is modeled as a conditional process in which local precipitation amounts are correlated with the occurrence of wet-days, which in turn correlated with regional-scale atmospheric predictors. A wet day is defined as a day with nonzero precipitation amount of 0.3 mm or more. As the distribution of daily precipitation is skewed, a fourth root transformation is applied to the original series to convert it to a normal distribution, and then used it in regression analysis. Temperatures are modeled as unconditional process in SDSM, in which a direct link is assumed between the large-scale predictors and local scale predictand. No transformation is applied to daily temperature data as daily temperature data are mostly normally distributed. The model is structured as monthly model for both daily precipitation and temperature downscaling, in which case, twelve regression equations are derived for twelve months

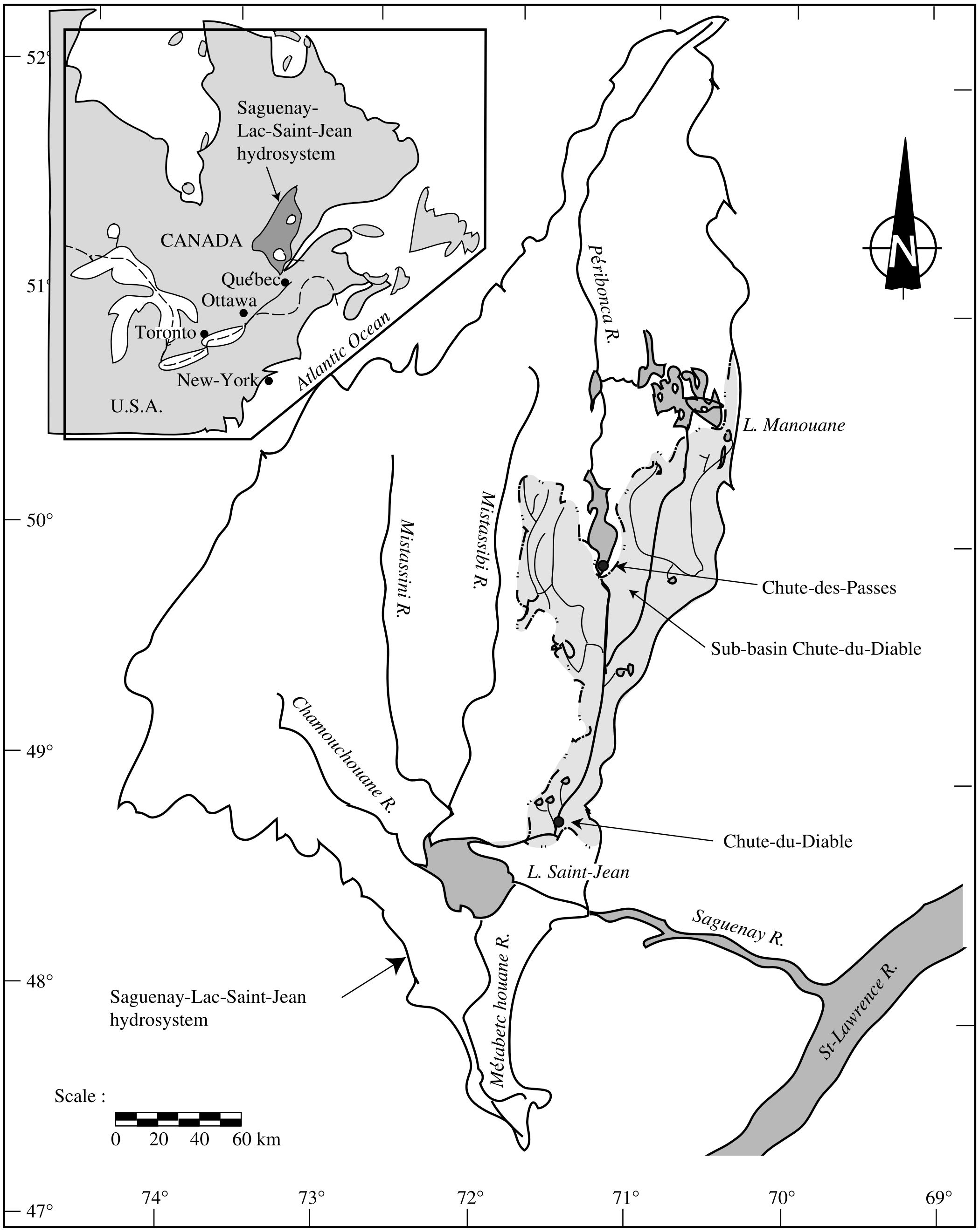

Fig. 1. Location map of the Saguenay–Lac–Saint–Jean watershed and the Chute-du-Diable sub-basin.

using different regression parameters for each month equation. The model is calibrated and validated separately for daily precipitation, daily maximum and minimum temperatures using thirty years (1961–1990) predictors and predictand for calibration and ten years (1991–2000) predictors and predictand for validation. During calibration, mean and variance of downscaled daily precipitation and temperature are

Table 1
Large-scale predictor variables selected for predicting the meteorological variables with SDSM and ANN methods

Predictor variable	Description	Precipitation				Max. Temp.				Min. Temp.			
		CDD		CDP		CDD		CDP		CDD		CDP	
		SDSM	ANN	SDSM	ANN	SDSM	ANN	SDSM	ANN	SDSM	ANN	SDSM	ANN
Temp	Mean temperature	×	×	×	×	×	×	×	×	×	×	×	×
mslp	Mean sea level pressure						×		×		×		×
p__u	Zonal velocity component near surface						×		×				
p5_u	Zonal velocity component at 500 hPa height		×		×								
p8_u	zonal velocity component at 850 hPa height		×		×								
p__v	meridional velocity component near surface	×	×	×	×	×	×	×	×	×	×		×
p8_v	Meridional velocity component at 850 hPa height	×	×	×	×								
p__z	Vorticity						×		×	×	×	×	×
p_zh	Divergence near surface										×		×
p5zh	Divergence at 500 hPa height		×		×								
p8zh	Divergence at 850 hPa height	×		×							×		×
p500	500 hPa geopotential height					×		×		×		×	
p850	850 hPa geopotential height						×		×		×	×	×
s500	Specific humidity at 500 hPa height	×	×	×	×	×		×				×	
s850	Specific humidity at 850 hPa height					×	×	×	×	×	×	×	×
Shum	Near surface specific humidity		×		×	×	×	×	×	×		×	

Note: All the predictors, with the exception of wind direction, have been normalized with respect to the 1961–1990 mean and standard deviation.

adjusted by bias correction and variance inflation factor to force model replicate the observed data. Bias correction compensates for any tendency to over- or under-estimate the mean of downscaled variables. Variance inflation changes the variance of downscaled daily weather variables by adding or reducing the amount of 'white noise' applied to regression model estimates of the local process to better accord with observations. Use of this stochastic (random) component also enables the SDSM regression model to produce multiple ensembles of downscaled weather variables. One hundred ensembles of downscaled daily precipitation, t_{max} and t_{min} have been generated but only the first ensemble was used in uncertainty analysis to maintain consistency with the ANN model, which has generated only one ensemble. In uncertainty analysis, we used total forty years downscaled data, which includes thirty years calibration results and ten years validation results. The reason for choosing 40 years data length is that the larger data length is capable of representing the true climatic condition for the site in question including less frequent climate events.

The second downscaling model is a stochastic weather generator called *Long Ashton Research Station Weather Generator* (LARS-WG) (Semenov and Barrow, 1997, 2002). In LARS-WG, during precipitation downscaling observed daily local station precipitation of each month are analyzed using a number of years historical data to obtain statistical characteristics such as number of dry days, wet days and mean daily precipitation in each month of a year. This information is used to develop semi-empirical distributions for the lengths of wet and dry day series and daily precipitation amount. From these semi-empirical distributions random values for wet and dry day series are generated for each month, and for a wet day, the precipitation value is generated from the semi-empirical precipitation distribution. So, in LARS-WG, precipitation modeling is also a two-step process like the SDSM model conditioned on wet and dry-days. Unlike SDSM, Temperature is modeled as conditional process in LARS-WG, conditioned on dry and wet status of the days. Mean and standard deviation of daily maximum and minimum temperature of each month for a number of historical years are calculated for wet and dry days and considered those as statistical parameters for temperature downscaling. The annual cycles of monthly means and standard deviations are approximated by Fourier series, and the residuals are approximated by a normal distribution. In this way the statistical parameters of observed daily precipitation and daily maximum and minimum temperatures are derived, which are used to generate number of ensembles replicating observed data or can be modified by using changes observed in large-scale-model generated precipitation and temperatures for different periods of time in climate change impact study. So, in LARS-WG downscaling unlike SDSM, large-scale atmospheric variables are not directly used in the model, rather, based on the relative monthly changes in mean daily precipitation amount, daily wet and dry series duration, mean daily temperature and temperature variability (standard deviation) between current and future periods predicted by a GCM, local station climate variables are adjusted proportionately to represent climate change. In this study, during LARS-WG downscaling 40 years (1961–2000) observed weather data (daily precipitation, daily maximum and minimum temperature) are analyzed to determine their statistical parameters. Those statistical characteristics of the observed data have been used to generate synthetic data for 300 years during validation of the model. The statistical characteristics of the observed and synthetic weather data were analyzed to determine if there were any statistically significant differences using *t*-test, *F*-test and Chi-squared (χ^2) test. After having satisfactory test results, the parameter files derived from observed weather data during the model calibration process were used to generate a number of ensembles of synthetic weather data for the time period of 1961–2000. Again, for the consistency with the ANN model, only the first ensemble is used for deriving uncertainty attributes of the downscaled results.

The third downscaling model is artificial neural network (ANN), developed by Coulibaly et al., 2005. This model is a non-linear regression type in which a relationship is developed between a few selected large-scale atmospheric predictors and basin scale meteorological predictands. In developing that relationship a *time lagged feed forward network* is used in which inputs are supplied through tap delay line and the network is trained using a variation of backpropagation algorithm (Principle et al., 2000). A slightly different approach is used in selecting predictors for the case of neural network downscaling. First the networks are trained with all (the twenty two) predictor variables as input to the networks. Then a sensitivity analysis is done to determine the most relevant predictors, which should be selected for further retraining. Sensitivity analysis provides a measure of the relative importance among the predictors (inputs of the neural network) by calculating how the model output varies in response to variation of an input. The network learning is disabled during this operation such that the network weights are not affected. The basic idea of sensitivity analysis is that the inputs to the neural network are shifted slightly and the corresponding change in the output is reported. Each input is varied between its mean $\pm$ standard deviation while all other inputs are fixed at their respective means. The network output is then computed for a specified number of inputs above and below the mean. This process is repeated for each input. The sensitivity is calculated by dividing the standard deviation of the output by the standard deviation of the input, which was varied to create the output. This way the most sensitive predictors are selected as relevant predictor variables (Table 1) for each neural network separately for precipitation, maximum and minimum temperature downscaling. The neural network is then retrained with the few selected predictor variables independently for precipitation, t_{max} and t_{min} till acceptable validation performance is achieved. The results show that, even though the set of variables selected as most relevant to ANN downscaling is not identical to SDSM downscaling predictors, some large scale predictor variables such as s500 (specific humidity at 500 hPa height), p_v (meridional wind velocity component at different levels), p500 (500 hPa geopotential height), temp (mean surface temperature) and shum (near surface specific humidity) are identified as relevant in most of the cases. Several training experiments are conducted with different combinations of input time lags and number of neurons in the hidden layer till the optimum network is identified. For the case of downscaling of precipitation with ANN, a time lag of six (days) and 20 neurons in the hidden layer gave the best performing network. Note that only one hidden layer and one output node are used in the ANN model. Hyperbolic tangent activation function is used at both the hidden and output layers of the neural networks. In the case of temperature downscaling, ANN with a time lag of three (days) and 12 neurons in the hidden layer have performed the best, this suggests that the predictand–predictors relationship is less complex in the case of temperature downscaling. Unlike SDSM, precipitation is downscaled with ANN as unconditional process by establishing direct link between large-scale predictors and local scale predictand (precipation). Moreover, the ANN model structure is considered deterministic restricting to simulate only one time series of downscaled daily precipitation, t_{max} and t_{min}.

4. Uncertainty assessment in downscaled results

This section describes techniques of uncertainty assessment used in this study in analyzing uncertainties of downscaled daily temperature and precipitation data. In case of daily temperature data, because of their nearly normal distribution, the uncertainty has been assessed with comparison of means and variances of downscaled temperature data with observed ones. In that comparison, deviations (referred to as model errors) between downscaled and observed monthly means and variances of daily t_{max} and t_{min} have been evaluated at 95% confidence level. Moreover, 95% confidence intervals in the estimates of means and variances of downscaled temperature in each month have been compared with observed confidence intervals to assess whether the downscaling models can reproduce uncertainty as found in the observed data. In uncertainty assessment of downscaled daily precipitation data, comparison of means and variances should not be enough because of non-normality of the distribution of daily precipitation amount and also because of mixed distribution of wet and dry days in a daily precipitation series. Therefore, in assessment of uncertainty in downscaled daily precipitation, in addition to comparing means and variances, monthly mean dry-spell and wet-spell statistics and their confidence intervals, distribution of monthly mean of daily precipitation, and distributions of monthly wet and dry days have been compared. In comparing those statistics either parametric or non-parametric approach can be employed. An exploratory data analysis, discussed in the following section, can help one decide which approach should be used.

4.1. Exploratory data analysis

The classical methods of statistical inference depend heavily on the assumption that the data are outlier-free and nearly normal, and that the data are serially uncorrelated if they are collected with regular time interval. If data are not nearly normal and not outlier-free, the results of the classical methods of statistical inference may be misleading. In such situation, *robust* or *nonparametric* methods may be used. As such, a graphical exploratory data analysis is carried out to answer the following questions:

- Do the data come from a nearly normal distribution?
- Do the data contain outliers?
- If the data were collected over time, is there any evidence of serial correlation (correlation between successive values of the data)?

A good picture of the shape of the distribution generating the data and the presence of outliers can be obtained by looking at the following collection of four plots: a *density plot*, a *histogram*, a *boxplot*, and a *normal qq-plot*. Density plots are essentially smooth versions of histograms, which provide smooth estimates of population frequency, or probability density curves. Box plots are used for conveying location and variation information in data; a box plot identifies the middle 50% of the data, the median, and the extreme points. A normal *qq*-plot (or quantile–quantile plot) consists of a plot of the ordered values of the data versus the corresponding quantiles of a standard normal distribution (a normal distribution with mean zero and variance one). If the *qq*-plot is fairly linear, the data are reasonably Gaussian; otherwise, they are not. Among those four plots, the histogram and density plots give the best picture of the distribution shape, while the *box* and normal *qq-plots* give the clearest display of outliers.

As data are collected over time, the data may exhibit serial correlation particularly for daily $t_{\max}$ and $t_{\min}$. This can be checked with time series and autocorrelation function (ACF) plots. In time series plots, by looking at a plot of data against time, one can check obvious time series features, such as trends and cycles. One can check the presence of less obvious serial correlation by looking at a plot of autocorrelation function for the data. The autocorrelation function is a measure of the correlation between $X_t$and X_{t+k} for a given *lag k*, and can be calculated as

$$\mathrm{ACF}(k) = \frac{\mathrm{Cov}(X_t, X_{t+k})}{\mathrm{Var}(X_t)} = \frac{\frac{1}{n-k}\sum_{t=1}^{n-k}(X_t-\bar{X})(X_{t+k}-\bar{X})}{\frac{1}{n-1}\sum_{t=1}^{n}(X_t-\bar{X})^2} \tag{1}$$

The results of the exploratory data analysis are described in Section 5.1. Based on those results the following non-parametric tests have been used in this study.

4.2. Non-parametric test for the difference of two population means

One of the best nonparametric methods for constructing a hypothesis test p-value for $\mu_1-\mu_2$ (difference of two population means), is the Wilcoxon rank sum method (Conover, 1980; Lehmann, 1975). This non-parametric test is also used in this study to test the difference of the means of observed and downscaled precipitation. In terms of hypothesis testing, p-value has the following interpretation: the p-value is the level of significance for which observed test statistic lies on the boundary between acceptance and rejection of the null hypothesis. At any significance level greater than the p-value, one rejects the null hypothesis, and at any significance level less than the p-value one accepts the null hypothesis. For example, if p-value is 0.03, one rejects the null hypothesis at a significance level of 0.05, and accepts the null hypothesis at a significance level of 0.01. A detailed description of the theory of Wilcoxon rank sum test can be found in Conover (1980) and Lehmann (1975). As described in Conover (1980), in case of hypothesis testing, one needs to combine both samples into a single ordered sample and then assign ranks to the sample values from the smallest to the largest, without regard to which population each value came from. Then the test statistic can be the sum of the ranks assigned to those values from one of the populations. If the sum is too small (or too large), there is some indication that the values from that population tend to be smaller (or larger, as the case may be) than the value from the other population. Hence, the null hypothesis of no differences between populations may be rejected if the ranks associated with one sample tend to be larger than those of the other sample.

4.3. Non-parametric test for the equality of two population variances

While data are continuous but not normally distributed, Levene's test (Levene, 1960) is usually used to test whether the two sample population variances are equal or not. The computational method employed here for Levene's Test is a modification of Levene's original procedure by Brown and Forsythe (1974). This method considers the distances of the observations from their sample median rather than their sample mean. Using the sample median rather than the sample mean makes the test more robust when the underlying data followed a skewed distribution. The hypothesis for the Levene's test can be defined as:

$$H_0: \quad \sigma_1 = \sigma_2 = ... = \sigma_k$$

$$H_a: \quad \sigma_i \neq \sigma_j \quad \text{for at least one pair } (i,j).$$

In performing Levene's test, a variable X with sample size N is divided into k subgroups, where N_i is the sample size of the ith subgroup, and the Levene test statistic is defined as:

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i(\bar{Z}_i - \bar{Z})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_i)^2} \tag{2}$$

where Z_{ij} is defined as:

$$Z_{ij} = |X_{ij} - \bar{X}_i| \tag{3}$$

where $\bar{X}_i$ is the median of the ith subgroup, $\bar{Z}_i$ are the group means of the Z_{ij} and $\bar{Z}$is the overall mean of the Z_{ij}. The Levene's test rejects the hypothesis that the variances are equal if

$$W > F_{(\alpha,k-1,N-k)}$$

where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F distribution with $k-1$ and $N-k$ degrees of freedom at a significance level of α. The statistical software Minitab (2003) has been used in performing the Levene's test.

4.4. Non-parametric confidence intervals in the estimates of means and variances

Confidence intervals in the estimates of means and variances provide information about uncertainty in the estimates of mean and variances. In this study, the most commonly used non-parametric technique, bootstrapping has been used for finding the confidence intervals of means and variances. The idea of bootstrapping is to resample a large number of new data sets with replacement from the original data set. Starting with a sample size of n, the algorithm for doing so is as follows:

1. Draw a new sample of size n *with replacement* from the original sample.
2. Calculate the mean or variance of the new sample call it m_1.
3. Repeat steps 1 and 2, 1000 times, calling the ith new sample mean or variance m_i.
4. Plot the distribution of these 1000 sample means or variances.
5. Construct the 95% confidence interval for the mean or variances by finding the 2.5th and 97.5th percentiles of this constructed distribution.

Using the S-PLUS function (S-PLUS 6, 2001), the bootstrap confidence intervals for the estimated means and variances have been calculated for daily precipitation, daily maximum and minimum temperatures for each month. For example, in estimating the confidence interval for the mean of observed daily precipitation in January for the time period of 1961–2000, an observed sample of size $n=1240$ (40 years daily January data gives $n=40\times31=1240$ data points) has been used. Then 1000 new samples, each of the same size as the observed data, are drawn with replacement from the observed data. The 1000 re-samples are drawn because this is the recommended minimum for estimating percentiles, required for estimating confidence interval for the means and variances (S-PLUS 6, 2001). The mean is first calculated using the observed data and then recalculated using each of the new samples, yielding a bootstrap distribution of the statistics of mean. From this distribution, the bias-corrected and accelerated (BCa) percentiles are estimated. The BCa percentile is more accurate than the empirical percentile. The empirical percentiles are simply the percentiles of the empirical distribution of the replicates while the BCa method transforms the specified probability values to determine which percentile of the empirical distribution most accurately estimate the percentiles of interest, and then applying corrections for bias and standard error, BCa confidence interval is estimated as follows (DiCiccio and Efron, 1996):

$$2\cdot5\text{th percentile} = \Phi\left(z_0 + \frac{z_0 + z^{0.025}}{1 - a(z_0 + az^{0.025}}\right) \tag{4}$$

$$97\cdot25\text{th percentile} = \Phi\left(z_0 + \frac{z_0 + z^{0.975}}{1 - a(z_0 + az^{0.975}}\right) \tag{5}$$

where z^{α} is the α quantile of standard normal distribution, z_0 and a, namely bias-correction and acceleration, are two parameters to be estimated, by (2.8) and (6.6) in DiCiccio and Efron (1996). The bias is defined as the difference between the values of the statistics of interest, which are calculated from the original data as well as using the bootstrap replicates, and acceleration refers to the fact that the standard error of the statistics of interest is not uniform over differing values of the statistic.

4.5. Non-parametric goodness-of-fit test

Kolmogorov–Smirnov non-parametric goodness-of-fit test has been used to compare cumulative distribution function (cdf) of downscaled and observed monthly mean daily precipitation as well as for comparing distribution of monthly dry and wet days series. The test can be described as follows. Suppose, $F_1(x)$ and $F_2(x)$ are two cdfs of two sample data of a variable x. The null hypothesis and the alternative hypothesis concerning their cdfs are:

$$H_0 : F_1(x) = F_2(x) \quad \text{for all } x$$

$$H_A : F_1(x) \neq F_2(x) \quad \text{for at least one value of } x$$

and the test statistic, T is defined as

$$T = \sup_x |F_1(x) - F_2(x)| \tag{6}$$

which is the maximum vertical distance between the distributions $F_1(x)$ and $F_2(x)$. If the test statistic is greater than some critical value, the null hypothesis is rejected.

5. Results and discussion

5.1. Exploratory data analysis

The exploratory data analysis plots of observed (1961–2000) daily precipitation, daily maximum and minimum temperatures at the Chute-du-Diable (CDD) station for the month of January are shown in Fig. 2. The histogram and density plots of daily precipitation in Fig. 2, reveal a distinctly skewed distribution, skewed toward the left. The distribution is not normal, and probably not even ‘nearly’ normal. The data may also possess outliers, which are illustrated by the *box* and *qq*-plots in Fig. 2. The box plot does not show any distinct shape, and the *qq*-plot is not straight either. But in the case of daily maximum and minimum temperatures as shown in Fig. 2, the shape of the histogram and density plots indicate normality of the data, and the *box* and *qq-plots* show absence of distinct outliers in data.

Another exploratory data analysis plot consists of the plots of time series and autocorrelation functions for the observed (1961–2000) daily precipitation, daily $t_{\max}$ and $t_{\min}$ at the station CDD for the month of January are shown in Fig. 3. Those plots investigate autocorrelation or serial dependency among data points. The time series and ACF plots in Fig. 3 of daily precipitation (showed only for 300 days) reveal no possible significant serial correlation among data points at 95% confidence level because all ACF values lie within the 95% confidence bands (horizontal dashed lines) for lags greater than 0. The ACF plots of daily $t_{\max}$ and $t_{\min}$ in Fig. 3, suggest some serial

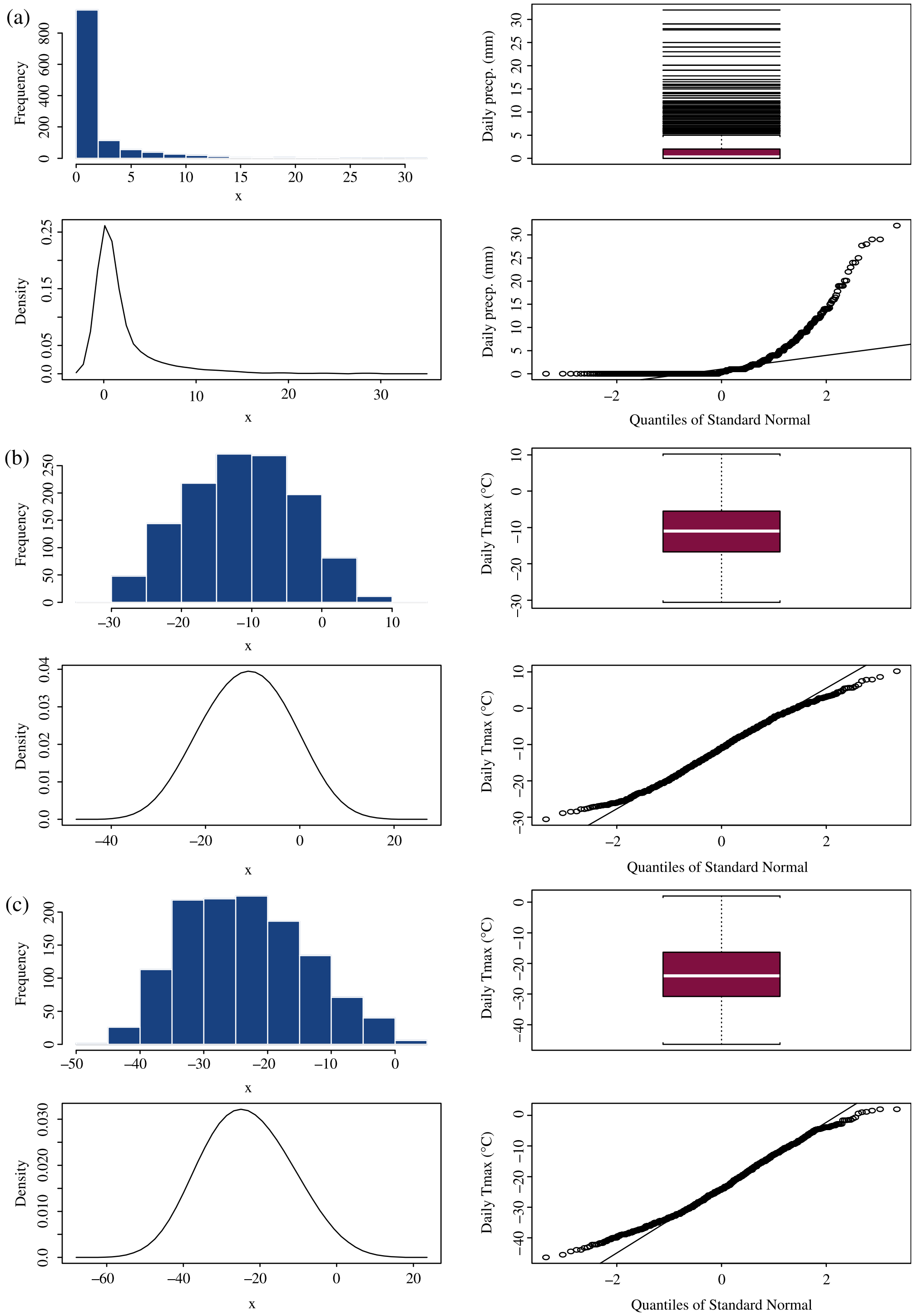

Fig. 2. Exploratory data analysis plots of observed (a) daily precipitation (top four plots); (b) daily tmax (middle four plots) and, (c) daily $t_{\min}$ (bottom four plots) for the month of January at the Chute-du-Diable (CDD) station.

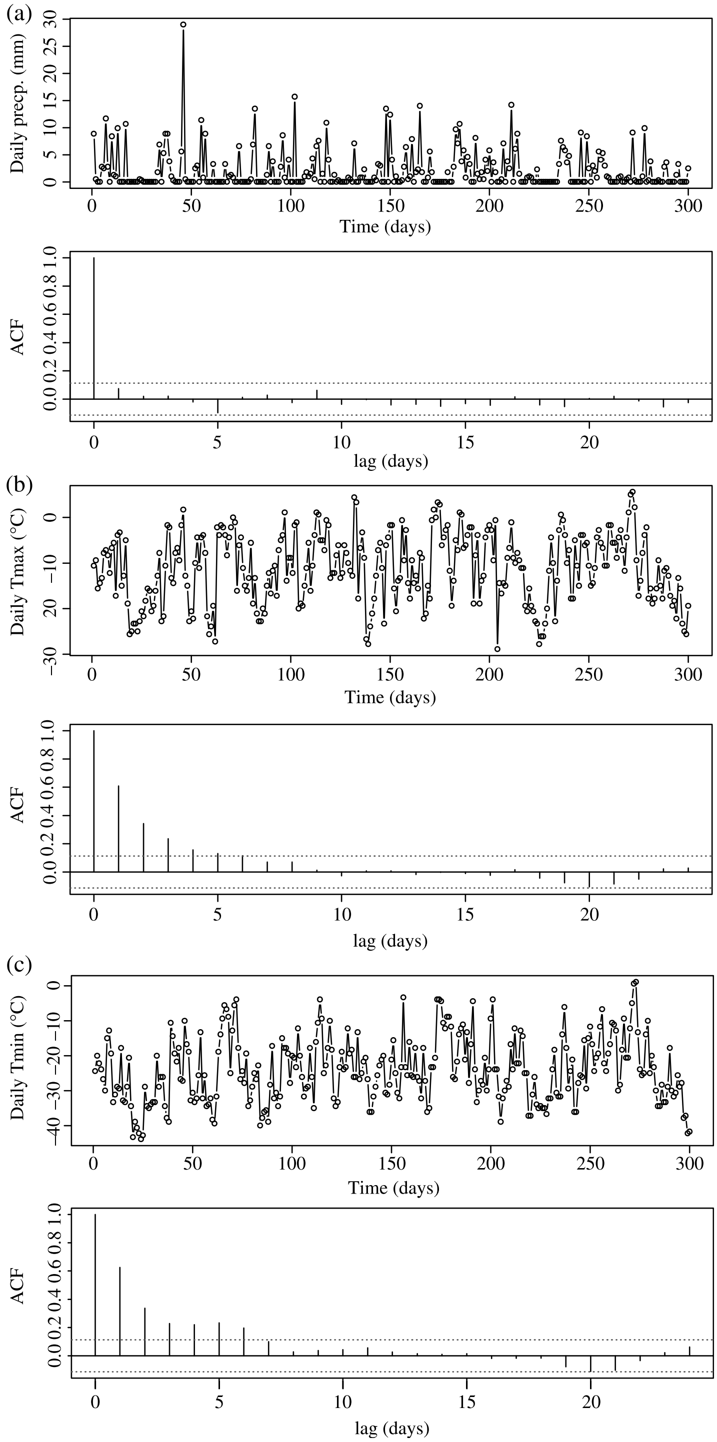

Fig. 3. Time series and ACF plots of observed (a) daily precipitation (top two plots); (b) daily tmax (middle two plots), and (c) daily tmin (bottom two plots) for the month of January at the station Chute-du-Diable (CDD).

correlations at the first few lags, which diminishes at higher lags.

Based on these graphical analyses results in Figs. 2 and 3, it can be assumed that the daily precipitation data are not normal and may contain some outliers, and do not have autocorrelation among data points. On the other hand, daily temperature data seems to be nearly normal, outlier free but correlated. It can therefore be concluded that the daily precipitation as well as temperature data do not hold all the assumptions of parametric data analysis. Similarly, the exploratory data analysis of downscaled and observed daily precipitation, daily maximum and minimum temperatures for all other months, which are not shown here supports the same conclusion. As such, in uncertainty assessment of downscaled daily precipitation, daily maximum and minimum temperatures, non-parametric approaches described in Sections 4.2–4.5 have been used. The following subsections discuss the results of those assessments. Note that both the calibration and validation data sets are combined in order to obtain a larger data set of 40 years length, which is more representative of the true climatic condition for the site in question including the less frequent climate events.

5.2. Error evaluation in the estimates of means

The absolute downscaling model errors (absolute values of the observed minus downscaled data) in the estimates of mean daily precipitation, daily maximum and minimum temperatures for each month have been shown in Fig. 4 for the stations CDD and CDP. Those errors are tested at 95% confidence level using non-parametric Wilcoxon rank-sum test for daily precipitation as described in Section 4.2. In daily precipitation downscaling at the station CDD, the ANN model errors are the least in most of the months (Fig. 4). However, at 5% significance level, both the SDSM and ANN models errors are found insignificant for all months because all estimated p-values are above 0.05 (see Table 2). The LARS model errors are found significant in the month of June, July and September while for other months the LARS model errors are found insignificant at 5% significance level. At the station CDP, all three models errors are insignificant at 5% significance level because all p-values of Wilcoxon rank-sum test are found above 0.05.

The model errors in downscaled daily maximum temperatures are shown in Fig. 4 (two plots in the middle row), indicate that the SDSM model errors are the least for all months for both the stations CDD and CDP. The reported p-values of the Wilcoxon rank sum tests for the differences of means of observed and downscaled data for all months are found high above 0.05 for SDSM model (see Table 3). This indicates that the SDSM model errors in all months are insignificant at 95% confidence level. The ANN and LARS models errors in daily t_{max} downscaling are not insignificant for all months. The p-values in Table 3 clearly indicate that in most of the months, the ANN and LARS models errors are significant for both the stations CDD and CDP. Similarly, the model errors and the statistical significance test results for the daily t_{min} are shown in Fig. 4 (bottom two plots) and Table 4, respectively, which also concludes that in daily t_{min} downscaling, the SDSM model produces the least error in all months at 95% confidence level while the ANN and LARS models errors are significant in most of the months.

5.3. Error evaluation in estimates of variances

A comparative plots of the variances of observed and downscaled daily precipitation, daily t_{max} and t_{min} for each month are shown in Fig. 5 for both the stations CDD and CDP. The equality of variances between observed and downscaled daily precipitation, daily t_{max} and t_{min} has been statistically tested in each month at 95% confidence level using the Levene's test. The corresponding test results for daily precipitation, daily t_{max} and t_{min} are shown in Tables 5–7, respectively. In case of daily precipitation variance, the graphical comparison in Fig. 5 shows that the ANN model does not represent the variability closer to the observed data, rather they show less variability, while the SDSM and LARS models variability is closer to the observed data. The variance test results in Table 5 show that, in the case of ANN model, for all months except January, February, March, May, October and December at the station CDD, and January, February, March and May at the station CDP, the p-values are all found below 0.05. This supports that the variances of the observed daily precipitation

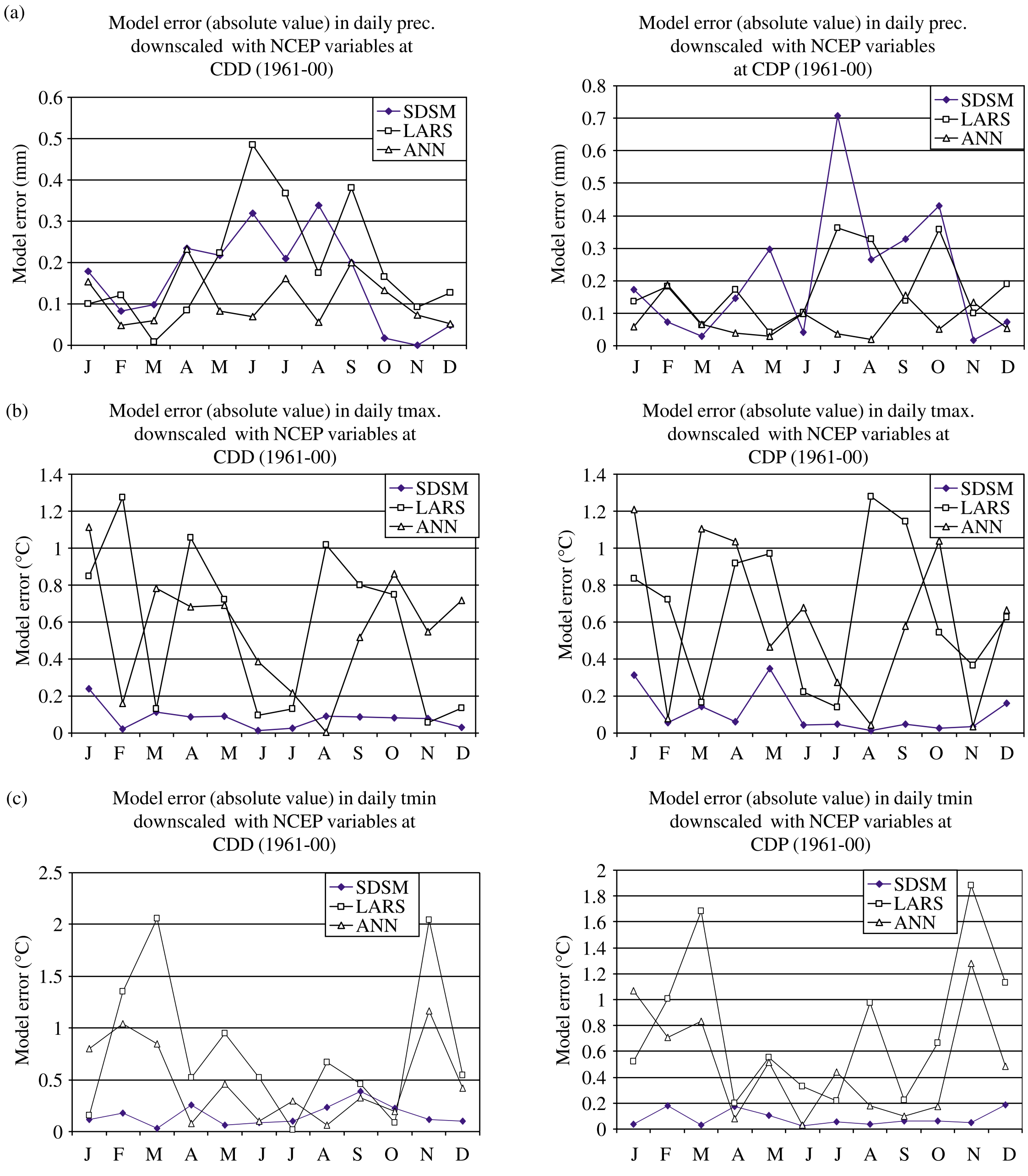

Fig. 4. Model errors (absolute values) in downscaled daily (a) precipitation, (b) t_{max}, and (c) t_{min}.

data and the ANN model simulated downscaled data are not same at 5% significance level in most of the months. On the other hand, the variance test results between the observed and the SDSM and LARS models simulation results show that the test *p*-values are all above 0.05 in all months (Table 5). This concludes that the observed and the SDSM and LARS models simulated daily precipitation variability can be considered statistically equal in all months with 95% confidence level.

Table 2
Test results (p-values) of the Wilcoxon rank sum test for the difference of means of observed and downscaled daily precipitation at 95% confidence level

	Station CDD			Station CDP		
	SDSM, p-value	LARS, p-value	ANN, p-value	SDSM, p-value	LARS, p-value	ANN, p-value
J	0.137	0.889	0.353	0.051	0.550	0.055
F	0.195	0.367	0.704	0.125	0.512	0.293
M	0.50	0.792	0.052	0.597	0.103	0.078
A	0.560	0.106	0.061	0.793	0.892	0.085
M	0.303	0.985	0.340	0.707	0.088	0.79
J	0.687	0.005	0.304	0.600	0.436	0.434
J	0.842	0.010	0.131	0.708	0.104	0.788
A	0.452	0.879	0.256	0.135	0.554	0.512
S	0.513	0.012	0.158	0.608	0.600	0.994
O	0.384	0.170	0.211	0.283	0.331	0.661
N	0.708	0.110	0.064	0.226	0.177	0.391
D	0.088	0.668	0.454	0.123	0.149	0.922

In the case of daily t_{max} downscaling, the graphical comparison of daily variances in each month between the observed and downscaled data in Fig. 5 indicates that the ANN model variability is not close enough with the observed variability but the SDSM and LARS models variability are closer to the observed variability. The equality of variance test results (see Table 6) at 95% confidence level indicate that the ANN model variability can be considered equal with the observed variability only in the months of February, March, April, October and December at the station CDD while at the station CDP they can be considered equal in the months of February and April only. The SDSM model variability can be considered equal with the observed data in all months at the stations CDD and CDP. For the LARS model, the variance test of daily tmax for each month at the station CDD indicates the same variability as of the observed data in almost all months except in the month of March. For the station CDP in the case of LARS model, the variability between the observed and simulated daily maximum temperature cannot be considered equal in the months of March, May, September and November.

Table 3
Test results (p-values) of the Wilcoxon rank sum test for the difference of means of observed and downscaled daily t_{max} at 95% confidence level

	Station CDD			Station CDP		
	SDSM, p-value	LARS, p-value	ANN, p-value	SDSM, p-value	LARS, p-value	ANN, p-value
J	0.833	0.002	0.019	0.983	0.021	0.018
F	0.497	0.001	0.819	0.835	0.037	0.431
M	0.466	0.000	0.007	0.796	0.000	0.001
A	0.947	0.006	0.009	0.999	0.005	0.003
M	0.950	0.000	0.156	0.886	0.001	0.524
J	0.395	0.903	0.056	0.592	0.120	0.005
J	0.512	0.344	0.266	0.495	0.146	0.052
A	0.796	0.000	0.874	0.673	0.000	0.793
S	0.726	0.006	0.001	0.629	0.000	0.000
O	0.558	0.003	0.002	0.865	0.044	0.000
N	0.968	0.008	0.006	0.348	0.004	0.351
D	0.744	0.438	0.009	0.866	0.015	0.154

Table 4
Test results (p-values) of the Wilcoxon rank sum test for the difference of means of observed and downscaled daily t_{min} at 95% confidence level

	Station CDD			Station CDP		
	SDSM, p-value	LARS, p-value	ANN, p-value	SDSM, p-value	LARS, p-value	ANN, p-value
J	0.883	0.511	0.027	0.583	0.376	0.007
F	0.613	0.000	0.017	0.389	0.000	0.063
M	0.831	0.000	0.031	0.682	0.000	0.041
A	0.613	0.767	0.837	0.167	0.069	0.098
M	0.582	0.000	0.021	0.955	0.000	0.000
J	0.755	0.005	0.811	0.586	0.080	0.669
J	0.515	0.878	0.011	0.655	0.681	0.003
A	0.824	0.000	0.100	0.697	0.000	0.053
S	0.104	0.160	0.405	0.488	0.415	0.579
O	0.675	0.021	0.142	0.305	0.005	0.122
N	0.441	0.000	0.000	0.333	0.000	0.000
D	0.982	0.743	0.095	0.889	0.009	0.176

In the case of daily tmin downscaling, the graphical comparison of daily variances in each month in Fig. 5 shows that in some months the ANN model variability is not close enough to the observed variability, while the other two models variability seems to be closer to the observed data. The test results of the equality of variances (Table 7) reveal that, at the station CDD, the ANN model variances are statistically equal with the observed variability only in the months of April, May, September and December. At the station CDP, they are equal only in the months of October and November for the ANN model. For the SDSM model, the variance test results show that the variability between the observed and downscaled

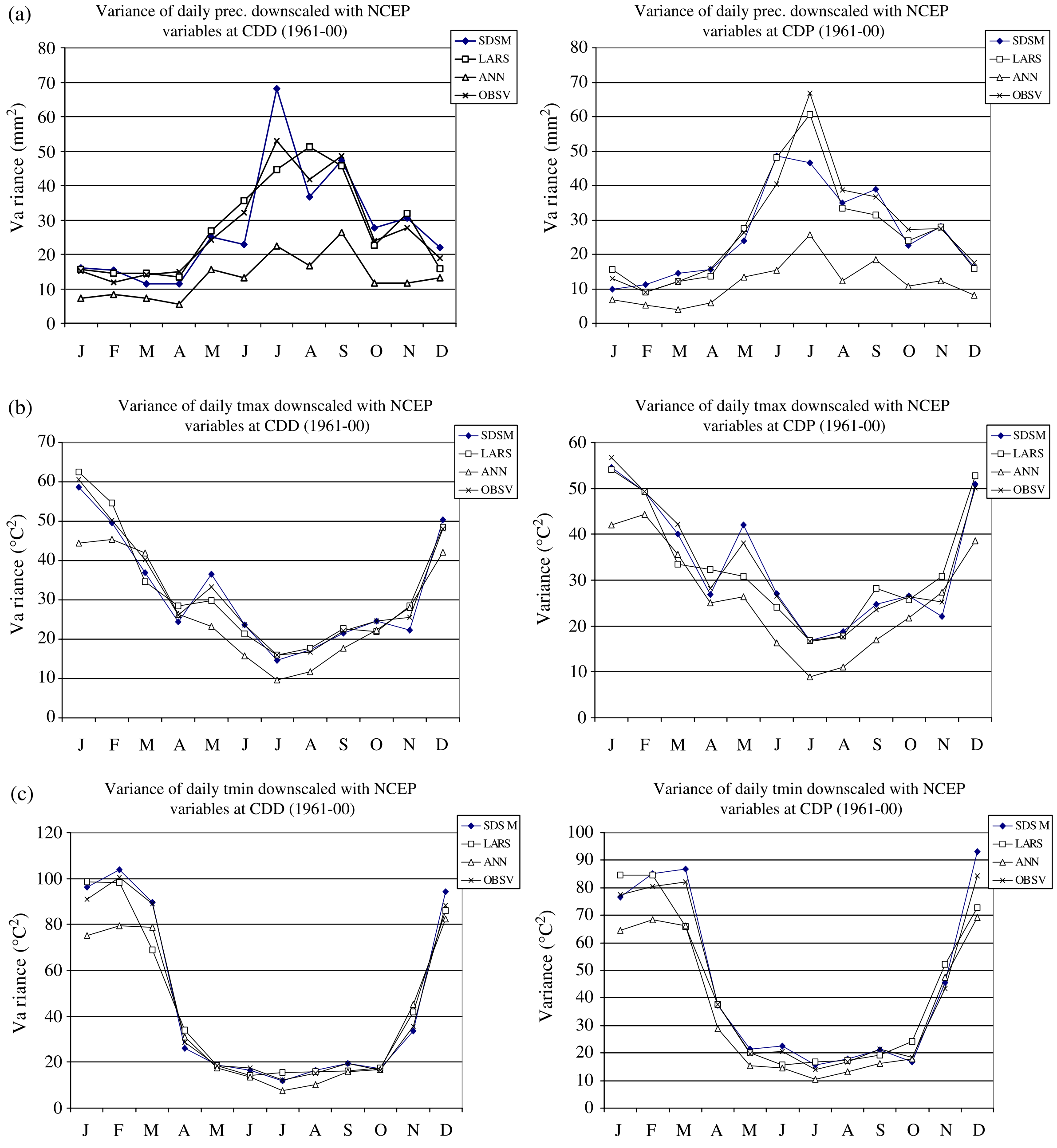

Fig. 5. Variances of the downscaled daily (a) precipitation, (b) t_{max} and (c) t_{min}.

data can be considered equal in all months at the stations CDD and CDP at 95% confidence level. For the LARS model in case of daily tmin downscaling, in the months of January, February, May, August, November and December, the variability is found equal at the station CDD and, at the station CDP that equality is found in the months of January, February, April, May, July, August and November.

Table 5
Test results (p-values) of the Levene's test for the equality of variances of the observed and downscaled daily precipitation at 95% confidence level

	Station CDD			Station CDP		
	SDSM, p-value	LARS, p-value	ANN, p-value	SDSM, p-value	LARS, p-value	ANN, p-value
J	0.390	0.880	0.068	0.256	0.343	0.107
F	0.595	0.424	0.638	0.591	0.147	0.332
M	0.656	0.887	0.650	0.844	0.646	0.153
A	0.127	0.614	0.051	0.370	0.268	0.001
M	0.227	0.224	0.215	0.140	0.847	0.084
J	0.182	0.060	0.033	0.876	0.705	0.025
J	0.530	0.147	0.002	0.501	0.326	0.000
A	0.626	0.992	0.031	0.397	0.224	0.000
S	0.562	0.299	0.050	0.312	0.442	0.009
O	0.851	0.347	0.122	0.051	0.10	0.002
N	0.525	0.730	0.080	0.617	0.847	0.001
D	0.547	0.584	0.574	0.845	0.327	0.019

5.4. Confidence intervals in the estimates of means

The uncertainty in the estimates of means of the observed and downscaled daily precipitation, daily maximum and minimum temperatures has been quantified by estimating confidence intervals about means. Non-parametric bootstrap approach (described in Section 4.4) has been used for estimating 95% confidence intervals about their means in each month.

Table 6
Test results (p-values) of the Levene's test for the equality of variances of the observed and downscaled daily $t_{\max}$ at 95% confidence level

	Station CDD			Station CDP		
	SDSM, p-value	LARS, p-value	ANN, p-value	SDSM, p-value	LARS, p-value	ANN, p-value
J	0.066	0.111	0.000	0.207	0.191	0.000
F	0.927	0.475	0.182	0.778	0.764	0.314
M	0.085	0.010	0.263	0.304	0.000	0.018
A	0.910	0.121	0.838	0.750	0.075	0.081
M	0.356	0.926	0.000	0.567	0.001	0.000
J	0.457	0.261	0.000	0.730	0.088	0.000
J	0.064	0.224	0.000	0.877	0.669	0.000
A	0.431	0.546	0.000	0.764	0.322	0.000
S	0.981	0.732	0.001	0.560	0.002	0.000
O	0.732	0.511	0.249	0.929	0.322	0.015
N	0.221	0.475	0.003	0.107	0.000	0.032
D	0.414	0.948	0.060	0.615	0.949	0.000

Table 7
Test results (p-values) of the Levene's test for the equality of variances of the observed and downscaled daily $t_{\min}$ at 95% confidence level

	Station CDD			Station CDP		
	SDSM	LARS	ANN	SDSM	LARS	ANN
	p-value	p-value	p-value	p-value	p-value	p-value
J	0.791	0.855	0.004	0.669	0.464	0.006
F	0.671	0.207	0.000	0.994	0.749	0.005
M	0.166	0.000	0.018	0.933	0.000	0.000
A	0.969	0.000	0.274	0.735	0.779	0.000
M	0.294	0.308	0.092	0.085	0.246	0.000
J	0.324	0.000	0.000	0.394	0.000	0.000
J	0.272	0.040	0.000	0.596	0.140	0.000
A	0.630	0.078	0.000	0.944	0.112	0.000
S	0.781	0.000	0.001	0.660	0.035	0.000
O	0.538	0.000	0.950	0.764	0.000	0.916
N	0.246	0.101	0.001	0.526	0.100	0.263
D	0.470	0.176	0.231	0.865	0.000	0.000

The plots of those confidence intervals are shown in Fig. 6 for both the stations CDD and CDP (left three plots are for the station CDD and the right three plots are for the station CDP). The results are discussed below.

In case of daily precipitation downscaling, the graphical comparison of uncertainty in the estimates of means of the observed and downscaled daily precipitation indicates that at the station CDD, the ANN model exhibited the least uncertainty in all months. However, that uncertainty is not representative of the observed uncertainty of daily precipitation in each month, because the observed uncertainty of daily precipitation at each month is found higher than the ANN model uncertainty. The other two downscaling models (the SDSM and LARS) exhibited uncertainty in downscaling daily precipitation almost at the same level as exhibited by the observed daily precipitation. The similar results are obtained in daily precipitation downscaling in each month for the station CDP as shown in Fig. 6.

In case of daily $t_{\max}$ downscaling (middle two plots of Fig. 6), again, the ANN model uncertainty in the estimates of means of daily $t_{\max}$ is not found representative of the observed uncertainty in most of the months compared to the other two downscaling models. At the station CDD, the uncertainty in the ANN model downscaled daily tmax is found almost

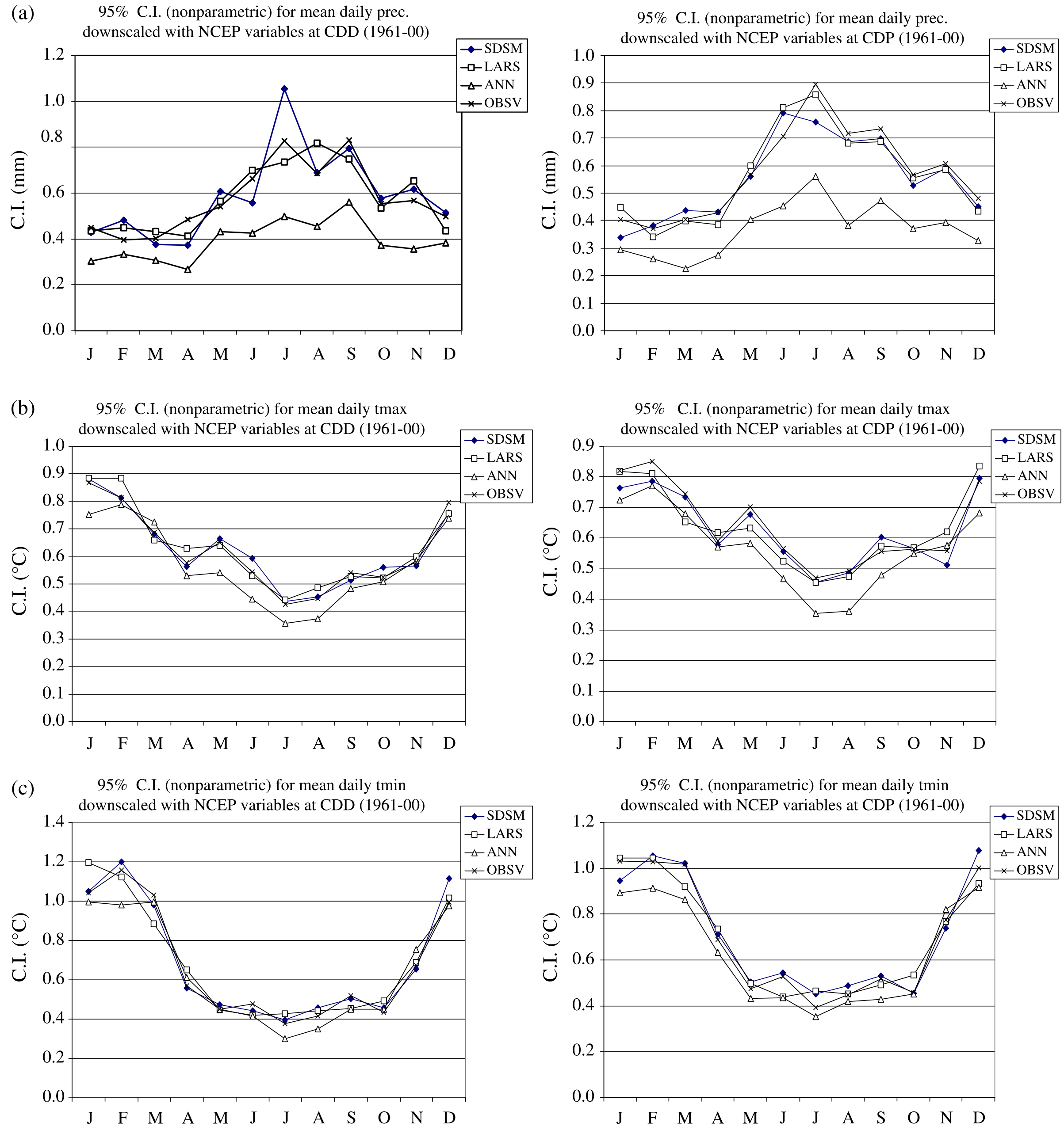

Fig. 6. Ninety-five percent confidence intervals for the estimates of mean daily (a) precipitation, (b) $t_{\max}$, and (c) $t_{\min}$.

equal to the observed uncertainty only in the months of February, March, October and November. On the other hand, the SDSM simulated uncertainty at the station CDD is found almost equal to the observed uncertainty in almost all months except the month of June. The LARS model uncertainty in downscaling daily $t_{\max}$ at the station CDD is found almost equal to the observed uncertainty in most of the months except the months of February, April and August. At the station CDP, the similar results are obtained; the ANN model represented the observed uncertainty the least, the SDSM is the best and the LARS is in between.

The uncertainty in the estimates of means of downscaled and observed mean daily tmin for each month are also shown in Fig. 6 for the stations CDD and CDP (two bottom plots). At the station CDD, in case of the ANN model, the uncertainty in downscaled daily t_{min} cannot be considered representative of the observed uncertainty in the months of February, July and August. For the SDSM model, the uncertainty in downscaled mean daily t_{min} may not be considered equal with the observed uncertainty only in the month of December; for the LARS model those unequal months are January, March and April. In downscaling daily t_{min} at the station CDP, according to graphical comparison between uncertainty of the observed and downscaled information, it can be said that the ANN model had exhibited uncertainties representative of the observed uncertainty of daily t_{min} only in the months of July, August, October and November. In case of SDSM, the uncertainty is found representative of the observed uncertainty in almost all months except January, July and December. For the LARS model, the uncertainty is found representative of the observed uncertainty in seven months except the months of March, June, July, October and December. Therefore, it can be said, in general, the ANN model simulated uncertainty is the least representative of the observed uncertainty, the SDSM is the best and the LARS is in between.

5.5. Confidence intervals in estimates of variance

The uncertainty in the estimates of variances has been quantified by calculating 95% confidence intervals of variances of the observed and downscaled daily precipitation, daily t_{max} and t_{min} in each month at the stations CDD and CDP following the procedure discussed in Section 4.4. The graphical comparison of uncertainty is shown in Fig. 7. In case of daily precipitation downscaling (top two plots of Fig. 7), at the station CDD, the ANN model exhibited the least uncertainty for 11 months but only three months' (February, March and December) uncertainty is closer to the observed uncertainty. For the SDSM model, at the station CDD, the simulated uncertainty is found closer to the observed uncertainty in most of the months except January, June, October and December. In the case of LARS simulation, at the station CDD, the uncertainty in the estimates of variances of downscaled daily precipitation is found closer to the observed uncertainty in most of the months except June and July. The similar trend is found at the station CDP, that is, the ANN is the least representative of the observed uncertainty in the estimates of variances, the LARS is the most and the SDSM is in between.

In the case of uncertainty in the estimates of variances of daily tmax downscaling as shown in Fig. 7 (middle two plots), at the station CDD, the ANN model showed the least uncertainty in most of the months of the year but that uncertainty is found closer to the observed uncertainty only in the months of March, April and November. For the SDSM model, the downscaled daily t_{max} at the station CDD showed closer uncertainty to the observed data in most of the months except January, May, June and November. In case of LARS model downscaling of daily tmax at the station CDD, in the most of the months the uncertainty in the estimates of variances is found closer to the observed data except January, February and December. In investigating uncertainty in daily t_{max} at the station CDP, it is found that the ANN model uncertainty is closer to the observed uncertainty only in the months of March and November. For the SDSM model downscaling of daily t_{max} at the station CDP, the uncertainty is found closer to the observed uncertainty in most of the months except May and June. In case of LARS model downscaling of daily t_{max} at the station CDP, the uncertainty is found closer to the observed in 8 months except January, April, September and December.

The uncertainty in the estimates of variances of daily t_{min} at each month has been shown in Fig. 7 (two bottom plots). The graphical comparison of observed and downscaled confidence intervals of variances of daily t_{min} for each month, indicates that at the station CDD, the uncertainty in downscaled daily t_{min} is found closer to the observed in seven months except January, February, March, November and December. In case of SDSM downscaling of daily t_{min} at the station CDD, the uncertainty is found closer to the observed in 8 months except January, February, April, and December. For the LARS model downscaling of daily t_{min} at the station CDD, the uncertainty of the downscaled information is found closer to the observed in 11 months except the month of

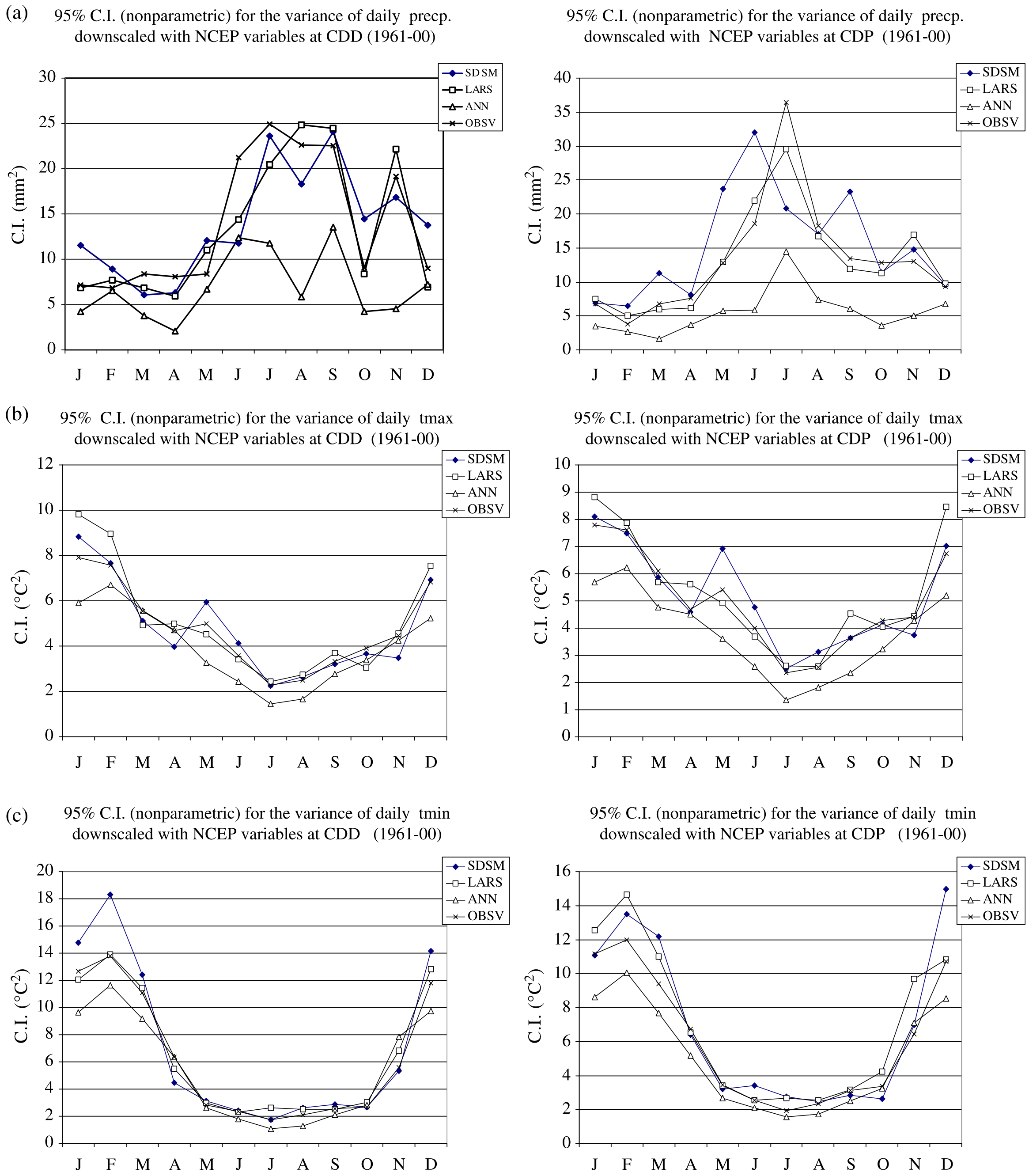

Fig. 7. Ninety-five percent confidence intervals for the estimates of variance of mean daily (a) precipitation, (b) $t_{\max}$, and (c) $t_{\min}$.

November. At the station CDP, in case of ANN model downscaling of daily $t_{\min}$, the uncertainty is found closer to the observed data in 7 months except January, February, March, April and December. In the case of SDSM model downscaling, the uncertainty is found closer to the observed in 9 months except the months of February, March and December. For the LARS model downscaling of daily $t_{\min}$, the uncertainty is found closer to the observed uncertainty in seven months except January, February, March, November and December. In general, the SDSM and LARS models downscaled data exhibit uncertainty closer to the observed data in most of the months in terms of confidence intervals of the estimates of variances. However, the ANN model cannot exhibit the uncertainty at the same level of the observed data in most of the months in terms of confidence intervals in the estimates of variances.

5.6. Distribution of monthly mean of daily precipitation

The distribution of monthly mean of daily precipitation has been constructed using 40 years (1961–2000) data. For each month in a year, daily precipitation mean has been estimated, which gives one data point for a particular month in a year. For 40 years, a series of forty data points of mean daily precipitation for a particular month has been constructed for observed as well as downscaled daily precipitation data. Then, using that 40 years data, the cumulative distributions of frequency (CDF) of observed and downscaled monthly mean of daily precipitation have been estimated, and compared using Kolmogorov–Smirnov nonparametric goodness-of-fit test to evaluate whether the two sample distributions come from the same population or not. The test results (p-values) provided in Table 8, indicate that for the station CDD during SDSM downscaling, in all months p-values are found above 0.05 (95% confidence level), while for the LARS model except June and September in all other months p-values are found above 0.05, and for the ANN model except October in all other months p-values are found above 0.05—suggesting that all the three downscaling models reproduce quite well the distribution of monthly mean of daily precipitation at 95% confidence level. Similarly, the analysis of the downscaled results at the station CDP (Table 8), shows that the SDSM model p-values are all above 0.05 except July, while the LARS model p-values are all above 0.05 except December, and the ANN model p-values are all above 0.05 in all months. This analysis indicates again that all downscaling models can reproduce the distribution of monthly mean of daily precipitation at 95% confidence level.

Table 8
Kolmogorov-Smirnov goodness-of-fit test results (p-values) for comparing distributions of monthly mean of daily precipitation constructed using forty years (1961–2000) observed and downscaled daily precipitation data

	Station CDD			Station CDP		
	SDSM, p-value	LARS, p-value	ANN, p-value	SDSM, p-value	LARS, p-value	ANN, p-value
J	0.406	0.578	0.578	0.765	0.578	0.765
F	0.765	0.404	0.918	0.765	0.404	0.265
M	0.578	0.404	0.097	0.765	0.99	0.165
A	0.404	0.918	0.165	0.578	0.265	0.265
M	0.404	0.99	0.404	0.054	0.404	0.765
J	0.765	0.02	0.404	0.765	0.918	0.578
J	0.765	0.265	0.165	0.028	0.265	0.578
A	0.165	0.918	0.578	0.165	0.165	0.265
S	0.765	0.028	0.578	0.765	0.765	0.265
O	0.404	0.765	0.028	0.097	0.165	0.097
N	0.765	0.765	0.165	0.165	0.578	0.054
D	0.578	0.404	0.918	0.054	0.028	0.265

5.7. Dry and wet spell length statistics

Dry and wet spell lengths for a particular month can be defined as maximum number of consecutive dry and wet days in that month, respectively. For instance, if in a particular month of a given year, maximum four consecutive days are found dry and six consecutive days are found wet, then we consider that the dry-spell and wet-spell lengths for that particular month are 4 and 6, respectively. Dry and wet spell lengths are of particular interest for hydrologic modeling, and are thus considered as additional criteria for assessing the downscaling model performance. Using 40 years (1961–2000) data, wet-spell and dry-spell lengths in each month have been estimated for observed and downscaled daily precipitation data. The arithmetic average of the 40 data points of dry and wet spells in a particular month provides mean statistics of dry and wet-spell lengths for that particular month. Comparative plots of that statistics between observed and downscaled daily precipitation are shown in Fig. 8 for both the stations CDD and CDP. At the station CDD it is found that the ANN model consistently underestimated dry-spell length for all months while the other two models (LARS and SDSM) estimated dry-spell length closer to the observed data. The similar trend is found for the station CDP for

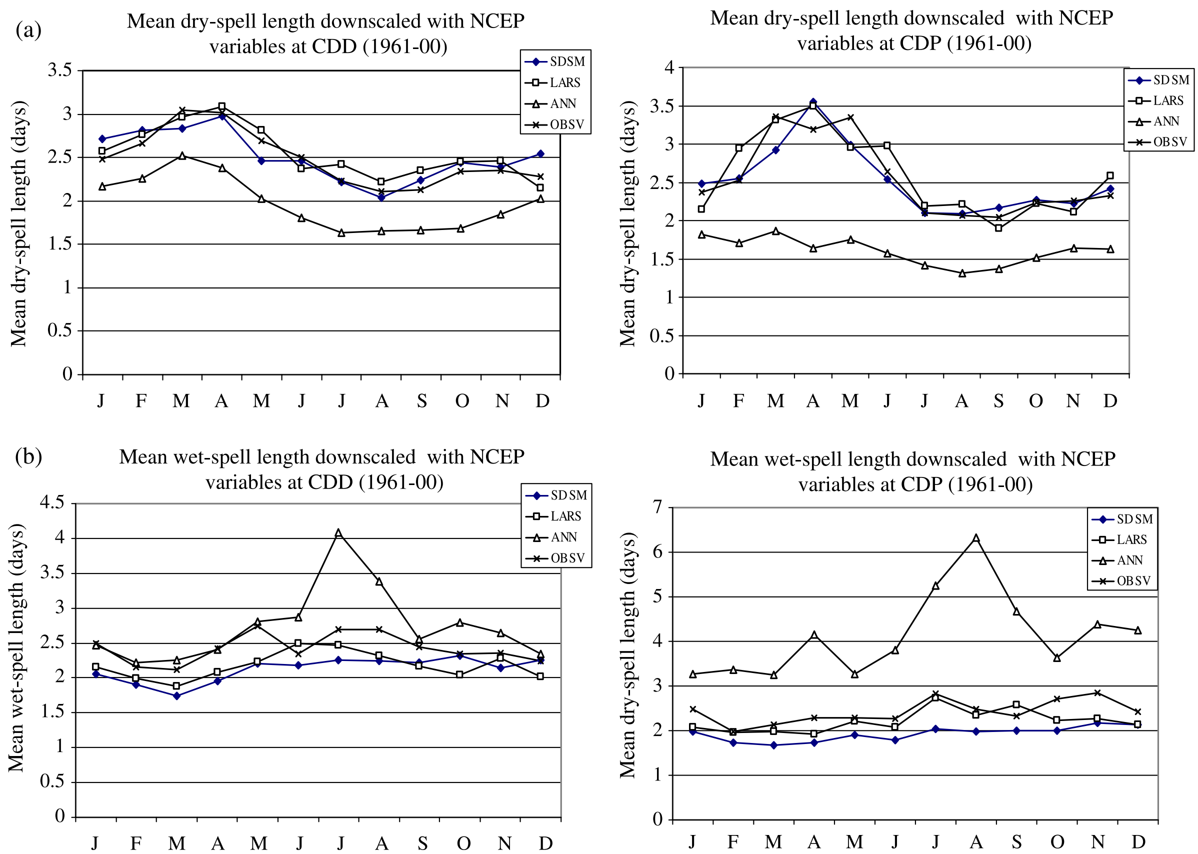

Fig. 8. Comparison of monthly mean dry-spell and wet-spell lengths between observed and downscaled daily precipitation.

dry-spell length comparison. In wet-spell length comparison, at the station CDD, from January to May, the ANN model results are found closer to the observed data, while in the rest of the months the ANN model overestimated wet-spell lengths. The other two models simulated wet-spell lengths are found very close to the observed values at the station CDD. At the station CDP, the ANN model over-estimated wet-spell lengths for all months of the years while the other two models wet-spell lengths are found closer to the observed data. This analysis indicates that in terms of mean wet and dry-spell lengths comparison, the SDSM and the LARS models performance can be considered at the same level while the ANN model is further apart in replicating observed spell statistics. This poor performance of the ANN model may be due to the fact that the ANN model tends to generate small trace precipitation even in actual dry periods. To further analyze the significance of the estimated dry and wet spell lengths, uncertainty analysis of the estimates of mean dry and wet spell lengths is performed by calculating the 95% non-parametric bootstrap confidence intervals of mean dry and wet-spell lengths using the forty years data. The uncertainty of the observed daily precipitation spell lengths has been compared with the uncertainty of the downscaled daily precipitation spell lengths. The graphical plots of those uncertainty estimates are shown in Fig. 9. By comparing those plots for the stations CDD and CDP it appears that the SDSM and the LARS models can in general closely replicate observed uncertainty of mean (dry and wet) spell lengths in almost all months while

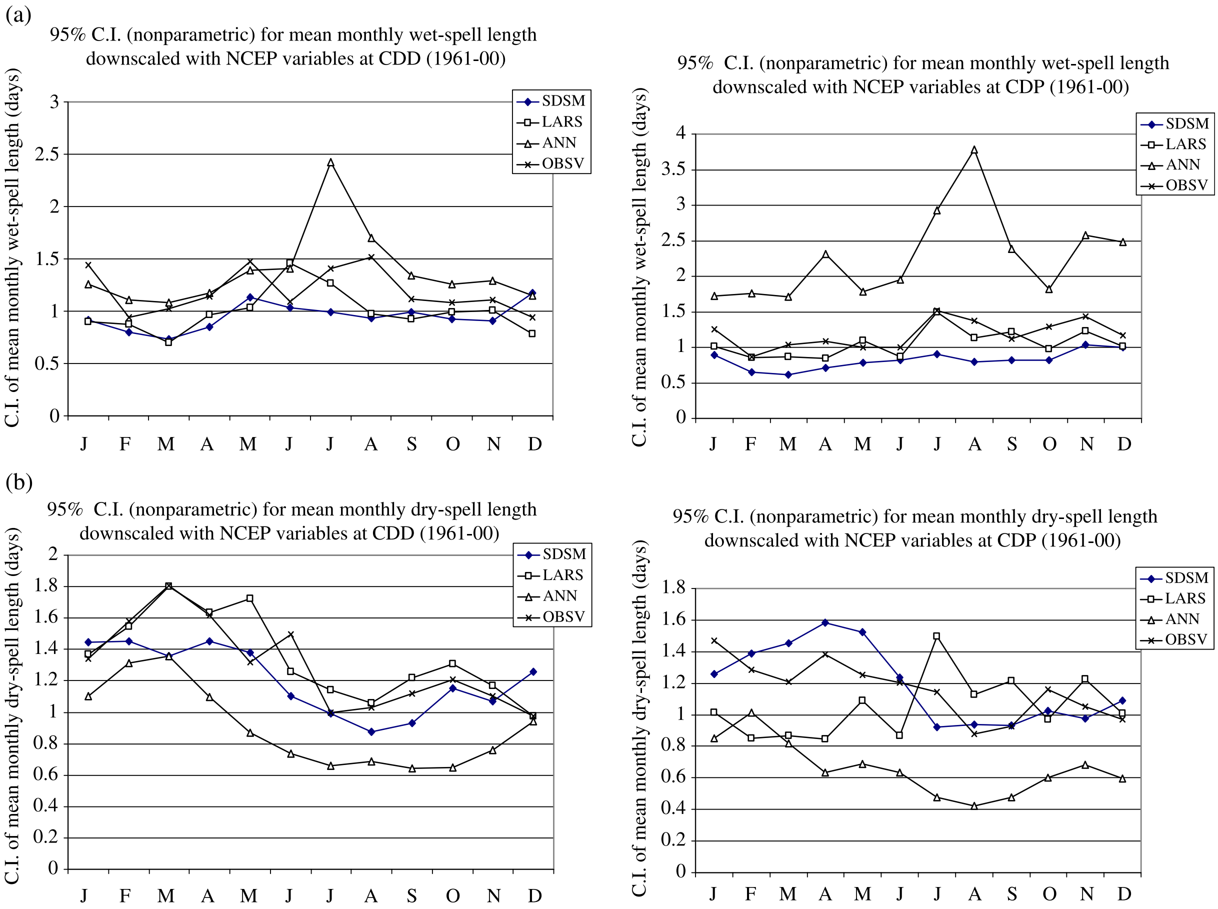

Fig. 9. Ninety-five percent confidence intervals for the estimates of mean monthly wet-spell and dry-spell lengths of observed and downscaled daily precipitation.

the ANN model cannot. These results (Fig. 9) also confirm the findings illustrated by Fig. 8.

5.8. Distribution of monthly dry and wet days

For further assessment of the downscaling models in daily precipitation downscaling, distributions of monthly dry and wet days have been considered. In doing that analysis numbers of wet and dry days in each month of a year have been calculated using the 40 years (1961-2000) data. This provides two series of wet and dry days with forty data points in each series for each month. The cumulative distribution of frequency (CDF) of those calculated wet and dry series has been estimated in each month for observed and downscaled daily precipitation, and compared those distributions using Kolmogorov–Smirnov goodness-of-fit test. The test results are provided in Table 9. In case of SDSM and LARS models, the p-values of the goodness-of-fit test for wet and dry days distributions are all found above 0.05, while the ANN model shows different results. For the station CDD, only in six months the p-values are found above 0.05 for both wet and dry days, while at the station CDP, the p-values in all twelve months are found at zero. This indicates that the ANN model has a strong limitation in reproducing distribution of wet and dry days as found in the observed daily precipitation data, while the SDSM and the LARS models can

Table 9
Kolmogorov-Smirnov goodness-of-fit test results (p-values) for comparing distribution of forty years (1961–2000) observed and downscaled monthly wet and dry days

	CDD						CDP					
	SDSM		LARS		ANN		SDSM		LARS		ANN	
	Wet-days	Dry-days	Wet-days	Dry-days	Wet-days	Dry-days	Wet-days	Dry-days	Wet-days	Dry-days	Wet-days	Dry-days
J	0.765	0.765	0.99	0.999	0.765	0.765	0.165	0.165	0.404	0.404	0.0	0.0
F	0.404	0.265	0.265	0.918	0.765	0.918	0.918	0.765	0.918	0.918	0.0	0.0
M	0.404	0.404	0.999	0.999	0.165	0.165	0.165	0.165	0.918	0.918	0.0	0.0
A	0.765	0.765	0.765	0.765	0.765	0.765	0.097	0.097	0.165	0.165	0.0	0.0
M	0.578	0.578	0.165	0.165	0.006	0.008	0.097	0.097	1.0	1.0	0.00	0.00
J	0.99	0.99	0.097	0.097	0.00	0.00	0.765	0.765	0.99	0.99	0.00	0.00
J	0.09	0.097	0.765	0.765	0.00	0.00	0.054	0.054	0.404	0.404	0.00	0.00
A	0.404	0.404	0.918	0.918	0.028	0.026	0.165	0.165	0.918	0.918	0.00	0.00
S	0.165	0.165	0.578	0.578	0.165	0.165	0.097	0.097	0.097	0.097	0.00	0.00
O	0.404	0.404	0.999	0.999	0.014	0.014	0.265	0.265	0.918	0.918	0.00	0.00
N	0.578	0.578	0.765	0.765	0.014	0.014	0.054	0.054	0.918	0.918	0.00	0.00
D	0.054	0.0541	0.765	0.765	0.918	0.918	0.265	0.265	0.765	0.765	0.00	0.00

reproduce observed distribution of wet and dry days in their downscaled daily precipitation data.

6. Summary and conclusion

The three statistical downscaling models namely SDSM, LARS-WG and ANN have been compared by assessing uncertainties in their downscaled results of daily precipitation, daily maximum and minimum temperatures. In the cases of daily maximum and minimum temperature, uncertainty is assessed by comparing monthly means and variances of downscaled and observed daily maximum and minimum temperature for each month of the year at 95% confidence level. In addition, uncertainties of the monthly means and variances of downscaled daily temperature have been calculated using 95% confidence intervals, which are compared with the observed uncertainties of means and variances. In daily precipitation downscaling, in addition to comparing means and variances, uncertainties have been assessed by comparing monthly mean dry and wet spell lengths and their confidence intervals, the cumulative frequency distributions (cdfs) of monthly mean of daily precipitation, and the distributions of monthly wet and dry days for observed and downscaled daily precipitation.

In comparing means of daily maximum and minimum temperatures, the SDSM model errors (difference between observed and downscaled data) are found insignificant in all months at 95% confidence level but the ANN and LARS model errors are found significant in most of the months based on that criteria. In confidence interval comparison of mean daily maximum and minimum temperature, the SDSM was able to reproduce observed uncertainty in its downscaled results in almost all months of the year, the LARS performs well in some months of the year, and the ANN in few months of the year. In comparing variances of observed and downscaled daily maximum and minimum temperatures at each month of the year, the SDSM model errors were insignificant in all months of the year at 95% confidence level, the LARS model errors were not insignificant in all months, in some months they were insignificant and in some months they were significant, and the ANN model errors were insignificant in few months but significant in most months of the year. In confidence interval comparison of variances of daily maximum and minimum temperatures in each month, the SDSM and the LARS models were able to reproduce observed uncertainty in their downscaled temperature in almost all months of the year but the ANN model was able only in few months of the year.

For the daily precipitation downscaling, in comparison of means of observed and downscaled daily precipitation in each month, the errors were found insignificant in all months at 95% confidence level for all three models. In confidence interval comparison of daily precipitation in each month, the SDSM and the LARS models were able to reproduce uncertainty very close to the observed one in all months, while the ANN model was not able to reproduce uncertainty close to the observed one in all months. In variance comparison of daily precipitation in each month, the SDSM and the LARS models errors were insignificant in all months at 95% confidence level but the ANN model errors were not found insignificant in all months, in some months they are significant and in some months they were insignificant. In 95% confidence interval comparison of daily precipitation variances in each month, the SDSM and the LARS models were able to reproduce uncertainty closer to the observed data in most months of the year while the ANN model was not able to reproduce observed uncertainty in most months of the year. In cumulative frequency distribution comparison of monthly mean of daily precipitation, all three models were able to reproduce observed distribution at 95% confidence level. In comparison of monthly mean dry and wet spell lengths and their uncertainty estimates with 95% confidence intervals, the SDSM and the LARS models were able to reproduce spell statistics closer to the observed data in all months while the ANN model was not able to reproduce spell statistics closer to the observed even in few months of the year. Finally, in comparison of cdfs of monthly dry and wet days, the SDSM and the LARS were able to reproduce observed distributions in their downscaled daily precipitation at 95% confidence level while the ANN model was not able to reproduce observed distributions in its downscaled daily precipitation at that confidence level.

Based on this comprehensive and rigorous assessment of uncertainty of downscaled daily precipitation, daily maximum and minimum temperature, it can be concluded that the SDSM is the best statistical downscaling model, the LARS is the second and the ANN is the third. The SDSM is capable of reproducing almost all statistical characteristics in its downscaled information at 95% confidence level as found in the observed data. The LARS is also capable of reproducing statistical characteristics of the observed data in its downscaling results but not at the same level as the SDSM can do. The ANN is the least capable of reproducing the observed statistical characteristics in its downscaling results specifically in daily precipitation downscaling. The reasons for superiority of one model with respect to the other can be attributed to their respective modeling techniques. For instance, in SDSM downscaling a regression relationship is developed between sensitive large-scale predictors and local scale predictand. That relationship is further tuned by adjusting means and variances of downscaled data by using bias correction and variance inflation so that the model can generate outputs closer to the observed data. During variance inflation a white noise is added, which makes the model stochastic and provides capability of generating a number of ensembles of downscaling results. Moreover, in the modeling process, the SDSM considers daily precipitation downscaling as a conditional process, in which precipitation amount are conditioned on the occurrence of wet-days, which in turn is linked with large scale atmospheric variables. On the other hand, during downscaling with LARS model no large scale atmospheric variables are used in the modeling process. Instead, the model analyzes local scale observed precipitation and temperature data to derive statistical characteristics representing observed data, and then change those statistical parameters proportionately based on changes found in the large-scale climate model respective variables. Similar to the SDSM, the LARS also considers precipitation downscaling as a conditional process in which case, empirical distributions of dry and wet days and precipitation amount are created, and random precipitation amounts are generated by conditioning empirical distribution of precipitation amount on the empirical distribution of wet-days. However, this stochastic model can not supersede the SDSM model, the reasons may be due to hybrid nature of the SDSM model, which not only considers deterministic relationship between predictors and predictand but also refine that relationship by using white noise in the model to account for the relationship of the inputs and outputs unexplained by the deterministic model. LARS model also produced significant errors in temperature downscaling in some months. The reason may be due to the smooth curve fitting to the average daily mean values for maximum and minimum temperature, which caused large departures from the observed data. However, overall, the ANN exhibits the most noticeable shortcomings as compared to SDSM and LARS. The ANN model did not consider precipitation downscaling as a conditional process, rather it established a direct nonlinear relationship between large-scale predictors and local scale predictand suppressing the precipitation occurring process, which causes significant errors in downscaled daily precipitation. Moreover, the ANN model considered here is deterministic, restricting to create only one time series. There is a scope of improvement of the ANN model by making the network stochastic and considering the precipitation downscaling as two step process in which dry and wet days and precipitation amount would be modeled separately. However, one should question whether these uncertainty assessment results remain stable under future climate forcing scenario. In the case of SDSM and ANN, the reliability of the results would be at the same level because model parameters would remain constant for future climate change condition, only a different set of large scale predictors will be used to represent climate changes. The same is true for the LARS-WG model as well because under future climate forcing, the observed statistics of the climate variable for the site in question would be altered by using changes found in the respective climate model variables between current and future periods produced by a global circulation model. Therefore, under future climate forcing, the respective performance of the three downscaling models would likely remain the same because under future conditions the uncertainty of their results would be mostly governed by the uncertainty of the GCM outputs.

Acknowledgements

This work was made possible through a grant from the Canadian Climate Change Action Fund, Environment Canada, and a grant from the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank the Aluminum Company of Canada (Alcan) for providing the study data. The authors also appreciated very much the valuable comments and suggestions received from two anonymous reviewers.

References

Ang, A.H-S., Tang, W.H., 1975. Probability Concepts in Engineering Planning and Design. Wiley, New York.

Brown, M.B., Forsythe, A.B., 1974. Robust tests for the equality of variances. J. Am. Stat. Assoc. 69, 364–367.

Caruso, B.S., 1999. Attributes and uncertainty of dissolved zinc data from a mined catchment. Hydrol. Processes 13, 169–190.

Coleman, H.W., Coleman Jr., H.W., Steele, W.G., 1989. Experimentation and Uncertainty Analysis for Engineers. Wiley, New York.

Conover, W.J., 1980. Practical Nonparametric Statistics, second ed. Wiley, New York.

Coulibaly, P., Dibike, Y.B., Anctil, F., 2005. Downscaling precipitation and temperature with temporal neural networks. J. Hydrometeorol.

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals (with discussion). Stat. Sci. 11, 189–228.

Kaleris, V., Papanastasopoulos, D., Lagas, G., 2001. Case study on impact of atmospheric circulation changes on river basin hydrology: uncertainty aspects. J. Hydrol. 245, 137–152.

Kistler, et al., 2001. The NCEP/NCAR 50-Year Reanalysis. Bull. Am. Meteorol. Soc., 226–247.

Lehmann, E.L., Lehmann, E.L., 1975. Nonparametrics: Statistical Methods Based on Ranks. Holden and Day, San Francisco.

Levene, H., 1960. Contributions to Probability and Statistics. Stanford University Press.

Loukas, A., Loukas, A., Vasiliades, L., Dalezios, N.R., 2002. Climatic impacts on the runoff generation processes in British Columbia, Canada. Hydrol. Earth Syst. Sci. 211, 228.

Minitab Inc, 2003. MINITAB Statistical Software, Release 14 for Windows. State College, Pennsylvania.

Montanari, A., Brath, A., 2004. A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. Water Resour. Res. 40. doi:10.1029/2003WR002540. W01106.

Principle, J.C., Euliano, N.R., Lefebvre, W.C., 2000. Neural and Adaptive Systems: Fundamentals through Simulations. Wiley, New York.

Semenov, M.A., Barrow, E.M., 1997. Use of stochastic weather generator in the development of climate change scenarios. Climate Change 35, 397–414.

Semenov, M.A., Barrow, E.M., 2002. LARS-WG: A Stochastic Weather Generator for Use in Climate Impact Studies, Version 3.0, User Manual.

S-PLUS 6 for Windows Guide to Statistics, 2001. Insightful Corporation, vol. 2. Seattle, WA, USA.

Wahl, T.L., 2004. Uncertainty of prediction of embankment dam breach parameters. J. Hydraul. Eng. 130, 389–397.
Wilby, R.L., Wigley, T.M.L., Conway, D., Jones, P.D., Hewitson, B.C., Main, J., Wilks, D.S., 1998. Statistical downscaling of general circulation model output: a comparison of methods. Water Resour. Res. 34, 2995–3008.
Wilby, R.L., Dawson, C.W., Barrow, E.M., 2002. SDSM—a decision support tool for the assessment of regional climate change impacts. Environ. Model. Software 17, 147–159.
Yar, M., Chatfield, C., 1990. Prediction intervals for Holt–Winters forecasting procedure. Int. J. Forecast. 6, 1–11.