Comparison of statistical methods for downscaling daily precipitation

Getnet Y. Muluye

ABSTRACT

There are several statistical downscaling methods available for generating local-scale meteorological variables from large-scale model outputs. There is still no universal single method, or group of methods, that is clearly superior, particularly for downscaling daily precipitation. This paper compares different statistical methods for downscaling daily precipitation from numerical weather prediction model output. Three different methods are considered: (i) hybrids; (ii) neural networks; and (iii) nearest neighbor-based approaches. These methods are implemented in the Saguenay watershed in northeastern Canada. Suites of standard diagnostic measures are computed to evaluate and inter-compare the performances of the downscaling models. Although results of the downscaling experiment show mixed performances, clear patterns emerge with respect to the reproduction of variation in daily precipitation and skill values. Artificial neural network-logistic regression (ANN-Logst), partial least squares (PLS) regression and recurrent multilayer perceptron (RMLP) models yield greater skill values, and conditional resampling method (SDSM) and *K*-nearest neighbor (KNN)-based models show the potential to capture the variability in daily precipitation.

Key words | nearest neighbor, neural network, numerical weather prediction model output, precipitation, statistical downscaling

Getnet Y. Muluye
Department of Civil Engineering,
McMaster University,
Hamilton,
ON L8S 4L7,
Canada
E-mail: *yayehgm@mcmaster.ca*

INTRODUCTION

The accuracy of a river forecast system largely depends on the quality of hydrologic inputs. The lack of reliable and accurate precipitation forecasts is one of the many challenges confronting the hydrologic community, particularly for use in short- to medium-range hydrological forecasts. Recent studies have shown that the use of information from meteorological forecasts and climate model simulations can help improve the accuracy of hydrological forecasts (e.g. Clark & Hay 2004; Schaake *et al.* 2006; Roulin 2007; Shi *et al.* 2008; Li *et al.* 2009; Muluye 2011a). As an example of the possibilities of the Medium-Range Forecast (MRF) model output for prediction of streamflow, Clark & Hay (2004) examined atmospheric forecasts over the contiguous United States from the National Centers for Environmental Prediction (NCEP) reanalysis project. Their findings indicated that forcing MRF output into a distributed-hydrologic model can yield useful results.

Obviously, large-scale model outputs possess substantial skill at global and regional scales (e.g. Harpham & Wilby 2005; Schmidli *et al.* 2007). Nevertheless, their usefulness for local or basin-scale hydrological studies are greatly limited owing to the coarse spatial resolution as well as the inability of large-scale model outputs to resolve important sub-grid scale features such as clouds and topography (Wilby *et al.* 2002). Bridging the gap between the resolution of large-scale model output and local-scale hydrological processes signifies a considerable difficulty for the effective assessment of grid and sub-grid scale hydrological responses (Fowler *et al.* 2007).

There exist two broad approaches to translate information from large-scale model outputs to the local scale. One way to achieve this is through dynamical downscaling, in which a regional climate model (RCM) uses large-scale model output as initial and lateral boundary conditions for

doi: 10.2166/hydro.2012.197

much more spatially detailed climatological simulations over a region of interest (Hay *et al*. 2002). The skill of such regional models in deriving local-scale variables from large-scale model output has been successfully demonstrated in various regions (e.g. Mearns *et al*. 1999; Hay & Clark 2003; Schmidli *et al*. 2007; Spak *et al*. 2007). It is important that although dynamical downscaling techniques are based on a strong physical realism, they do suffer from several limitations (see Mearns *et al*. 1999). An alternative approach to dynamical downscaling is to use statistical methods, in which an empirical relationship is established between large-scale model output and local-scale station variables (e.g. von Storch *et al*. 1993; Wilby *et al*. 2002; Hay & Clark 2003; Muluye 2011b). Wilby & Wigley (1997) further classify statistical downscaling techniques into regression methods, weather pattern-based approaches and stochastic weather generators. The general theory, applications, advantages and shortcomings of common downscaling methods are described in the literature (Wilby & Wigley 1997; Xu 1999; Yarnal *et al*. 2001; Fowler *et al*. 2007).

Statistical downscaling methods are becoming increasingly popular due to: (1) their simplicity in design and implementation; (2) computational efficiency; and (3) comparable or better suitability for short-term forecasts when compared to dynamical downscaling. Given a pool of statistical downscaling models available in the scientific literature, a limited number of articles are devoted to rigorous intercomparison of downscaling station meteorological variables (e.g. Schoof & Pryor 2001; Dibike & Coulibaly 2005; Mehrotra & Sharma 2005; Wetterhall *et al*. 2006; Wilby & Harris 2006; Liu *et al*. 2007). Furthermore, even fewer studies have reported downscaling of station daily precipitation from the NCEP MRF numerical weather prediction model output (Clark & Hay 2004; Gangopadhyay *et al*. 2005; Liu *et al*. 2007). Another important consideration is the demonstrated failure of current downscaling methods to effectively describe precipitation characteristics, such as occurrence and intensity (Clark & Hay 2004). Outputs from these downscaling models generally yield modest or poor performance when forced into a distributed-hydrologic model (Clark & Hay 2004). This clearly underscores the need for novel and innovative downscaling methods, particularly for daily precipitation. Therefore, the main objectives of this paper are: (i) to develop promising statistical methods for downscaling daily precipitation from a dataset of historical weather forecasts generated with a fixed numerical model – a 1998 version of NCEP's Global Forecasting System (GFS, formerly known as MRF) for the Chute-du-Diable weather station in northern Quebec, Canada; (ii) to assess the skill of these downscaling models through suites of diagnostic measures; and (iii) to compare the newly developed models with the commonly used statistical downscaling models. It is noteworthy to mention that the commonly used models which are used as a benchmark for comparison are selected based on preliminary analysis and literature review.

The remainder of the material is organized as follows. A brief description of statistical downscaling techniques used in the downscaling experiment is provided, followed by a description of the experimental setup and calibration of the downscaling methods. A comparison of the performance of the developed methods with that of the commonly used methods is given and the results of the downscaling experiment discussed. Finally, conclusions and recommendations are given based on the results obtained.

MODEL DESCRIPTION

This section presents a brief description of the statistical downscaling techniques selected for evaluation and intercomparison of downscaling daily precipitation fields. The methods considered are: (i) the Statistical DownScaling Model, (ii) partial least squares regression, (iii) hybrids, (iv) nearest neighbor-based models, and (v) a family of artificial neural networks.

Conditional resampling method

The most common regression-based technique used to map global climate model outputs to individual sites or localities is the Statistical DownScaling Model (SDSM; Wilby *et al*. 2002). The SDSM is best described as a hybrid of the stochastic weather generator and regression-based method. In this model, local-scale weather generator parameters, such as daily precipitation occurrence and intensity are linearly conditioned on the basis of large scale circulation patterns and atmospheric moisture variables. The variance of the

downscaled daily time series is synthetically inflated by means of stochastic methods to reasonably represent the observed time series. The downscaling algorithm of SDSM has been extensively applied to a wide spectrum of meteorological, hydrological and environmental assessments across the world including Africa, Europe, North America and Asia (Wilby & Dawson 2007). Technical details of the SDSM are found in Wilby & Dawson (2007).

Partial least squares regression

Multiple linear regression (MLR) is a common statistical method used to link large-scale model outputs to station-scale variables (Clark *et al.* 2004; Harpham & Wilby 2005). The major limitation of MLR is that when the predictor variables are not independent and are collinear, the model predictions can be poor (Wold *et al.* 1987). Because of this collinearity issue, principal component regression (PCR) is preferred over the MLR method. Partial least squares (PLS) regression, a technique that generalizes and combines good features of the principal component analysis and the MLR method (Lorber *et al.* 1987; Wold *et al.* 1987), is becoming popular. PLS regression attempts to identify factors (called latent variables) that maximize the amount of variation explained in **X** (predictors) that are relevant for predicting **Y** (predictands). This is an advantage when compared to the PCR regression where the factors (called Principal Components) are selected solely based on the amount of variation that they explain in **X** (Wold *et al.* 1987). As a result, the PLS regression is considered a better alternative to both the MLR and the PCR methods in terms of producing improved forecast skill.

There are several ways of computing the PLS model parameters. The most widely used algorithms are (de Jong 1993): NIPALS (Non-Iterative Partial Least Squares) and SIMPLS (Simple PLS). The latter algorithm was used in the present study as it is faster and less memory-intensive than the NIPALS. A detailed description and working procedure of SIMPLS can be found in de Jong (1993).

Hybrid models

Statistical models such as MLR and artificial neural networks (ANN) generally have a tendency to overestimate precipitation occurrences and underestimate precipitation amounts (Clark *et al.* 2004). It is not unusual to get negative precipitation forecasts from such models. To address this serious weakness, hybrid models are proposed in this study. In hybrid models, precipitation is modeled in a two stage process: Logistic regression is used to identify the occurrence of wet days, and PLS or ANN is used to model the amount of precipitation. A similar two stage procedure has been employed for modeling with SDSM; however, the procedure is different from the approach used in this study (e.g. Wilby *et al.* 2002).

The intermittent and skewed nature of daily precipitation data requires some preprocessing prior to developing the downscaling models. In order to model the occurrence of wet days, the site precipitation time series $\{d_1, d_2, \ldots, d_n\}$ is converted into a binary time series $\{y_1, y_2, \ldots, y_n\}$, with 0 representing dry days and 1 representing wet days. Given a data set $\{(x_1^t, x_2^t, \ldots, x_m^t)/t = 1, 2, \ldots, n\}$ of m large-scale predictor variables x_i, $i = 1, 2, \ldots, m$, the probability of occurrence of wet days $\{\hat{p}_t/t = 1, 2, \ldots, n\}$ can be modeled by performing the logistic regression:

$$\hat{p}_t = \frac{1}{1 + \exp\left(-\left(\alpha + \sum_{i=1}^{m} \beta_i x_i^t\right)\right)} \tag{1}$$

where α and β are model parameters of the logistic regression. The time series $\{\hat{p}_t/t = 1, 2, \ldots, n\}$ is further processed and transformed into a binary time series $\{\hat{y}_t/t = 1, 2, \ldots, n\}$ using the relation:

$$\hat{y}_t = \begin{cases} 1 & \text{if } \hat{p}_t - p \geq 0; \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The parameter p is a threshold of probability according to which $\{\hat{p}_t/t = 1, 2, \ldots, n\}$ is transformed into binary time series, $\{\hat{y}_t/t = 1, 2, \ldots, n\}$, and this transformation makes mismatches between $\{y_t/t = 1, 2, \ldots, n\}$ and $\{\hat{y}_t/t = 1, 2, \ldots, n\}$ minimal (Hua & Zhang 2006). The value of p can be set at some fixed number or a random number (for each simulation) from a uniform distribution ranging between 0 and 1 (e.g. Clark *et al.* 2004). In the present study, threshold values from 0.1 to 0.9 were considered and the value of $p = 0.5$ provided the best result. The results were inter-compared against p values assigned from a uniformly generated random number, and the threshold value

of $p = 0.5$ was found consistently superior. Intuitively, when the probability of precipitation occurrence is over 50%, then there is a greater chance that precipitation will occur, and vise versa. Thus, the selected p value makes practical sense.

Once wet and dry days are identified using logistic regression, a regression or neural network-based model can be used to model precipitation amounts. The large-scale predictor variables used were the same as for the logistic regression. Two hybrid models were developed: (i) PLS-Logst – in which the logistic regression was used to model precipitation occurrence and the PLS was used to develop models for precipitation amounts; and (ii) ANN-Logst – in which the logistic regression was used to model precipitation occurrence and the conventional multilayer perceptron (simply ANN) was used to develop models for precipitation amounts. It should be noted that it is possible to develop several hybrid models from similar coupling.

K-nearest neighbors

The nearest neighbor approaches work on the principle of classical bootstrapping techniques (e.g. Yakowitz 1993; Yates *et al*. 2003; Gangopadhyay *et al*. 2005). The *K*-nearest neighbor (KNN) algorithm is a very intuitive downscaling technique that identifies the next day's weather based on information given in the large-scale numerical weather prediction model output (Gangopadhyay *et al*. 2005). KNN-based models have been successfully applied to generating synthetic weather data on various water resources and environmental problems (e.g. Rajagopalan & Lall 1999; Buishand & Brandsma 2001; Yates *et al*. 2003; Mehrotra & Sharma 2006; Sharif & Burn 2006; Muluye 2011b).

Among others, one of the key issues that dictate effectiveness of nearest neighbor-based models is the type of distance measure or nearness used in the modeling. The problem is more apparent particularly when the feature vector consists of multiple variables. The measure of closeness is typically quantified either using the Euclidean or Mahalanobis distance (Mehrotra & Sharma 2006; Sharif & Burn 2006). The Mahalanobis distance formulation has the following advantages over the traditional resampling models which utilize the Euclidean distance (Mehrotra *et al*. 2004): (i) the use of Mahalanobis distance obviates the standardizing of the predictor variables; and (ii) the Mahalanobis distance measure considers the existing dependence amongst the predictor variables. The Euclidean distance formulation discussed in Gangopadhyay *et al*. (2005) overcomes the major limitation in (ii) by considering less significant, correlated or redundant predictors through the use of principal component analysis (PCA), and the subsequent PCs are then weighted according to their eigenvalues. The major weakness of both distance metrics is that they ignore the dependence between the predictors and predictands. Because of their inherent strengths and weaknesses, both distance metrics are considered in the present study. The present study employs methods as discussed in Gangopadhyay *et al*. (2005) and Yates *et al*. (2003) for modeling with Euclidian and Mahalanobis distance formulations, respectively. Thus, the two KNN-based models considered are: (i) KNN model based on Mahalanobis distance (KNN-M), and (ii) KNN model based on the Euclidean distance (KNN-E). Further description of these models can be found in Muluye (2011b).

Artificial neural networks

The utility of artificial neural networks in various hydrology-related areas, including precipitation forecasting, has been extensively studied. There exists a diverse family of neural networks in the scientific literature, among which, feedforward neural networks or multilayer perceptron (MLP) is the most widely used architecture (e.g. ASCE 2000; Haykin 2008). Pure MLP along with time-lagged feedforward network architectures have been applied to represent the spatial and temporal information of dynamic systems. But it has been shown that a neural network containing a state feedback possesses more computational advantages and uses fewer numbers of parameters than input-output models (Palma *et al*. 2001). One way of effective characterization of a state feedback is through the use of a recurrent multilayer perceptron (RMLP). The application of feedback enables the RMLP to acquire state representations, which makes the architecture suitable to describe a large class of nonlinear dynamic systems (Choi *et al*. 2005).

The two most widely used algorithms to perform supervised training of the RMLP are back-propagation through time and real-time recurrent learning. Both algorithms are based on the gradient method which utilizes first-order

derivative information. The real-time recurrent learning algorithm that uses continuous learning based on gradient descent has the following drawbacks (Haykin 2008): (i) it is typically slow due to its reliance on instantaneous estimates of gradients, compared with a learning algorithm that uses second-order derivative information; and (ii) it suffers from the vanishing gradient problem. These serious weaknesses can be overcome through the use of second-order training algorithms such as a linearized recursive least squares or an extended Kalman filter. The algorithms applied in the present study to perform supervised training of the MLP and the RMLP models are, respectively, the classical back-propagation algorithm and the decoupled version of an extended Kalman filter. For detailed treatment of the subject, the reader is referred to one of many sources (e.g. Puskorius & Feldkamp 1994; Haykin 2008; Muluye 2011c).

EXPERIMENTAL SETUP AND DATA

Study area and data

The study area selected for the evaluation and investigation of downscaling models was the Saguenay-Lac-Saint-Jean (SLSJ) hydrologic system in northern Quebec, Canada. The Chute-du-Diable meteorological station, located at 48.75° N and 71.7° W was considered for the present study. Alcan Company operates and maintains the Chute-du-Diable meteorological station primarily for the purpose of hydroelectric power generation. Station daily precipitation data were collected from Alcan hydro-meteorological database for the period 1979 to 2001. The station is located at an elevation of 174 m with an average annual temperature of 2 °C and an annual precipitation amount of 935 mm, from which 27% is in the form of snow. The maximum recorded daily precipitation is 70 mm, the average daily precipitation is 2.5 mm and the daily precipitation variability is 5.1 mm based on recorded data since 1950.

The atmospheric predictor variables used for the downscaling experiment were a 'reforecast' dataset generated by the Physical Climate Division of NOAA's Earth System Research Lab (http://www.esrl.noaa.gov/psd/forecasts/reforecast/details.html). A fixed numerical model – a 1998 version of NCEP's GFS – was used to generate an ensemble of 15-day forecasts over a 23-year period from 1979 to 2001. The NCEP's GFS model output ensembles are defined on a global lat-lon grid with 2.5° resolutions both in longitude and latitude (144×73 grid points). The data used in this analysis consist of daily values from a 23-year period from January 1979 through December 2001. Large-scale model outputs corresponding to the first ensemble member (i.e. out of 15) and all 15 forecast ranges were considered (Table 1). The large-scale model outputs were retrieved from the nearest grid to the Chute-du-Diable meteorological station (Muluye 2011b).

Application and calibration

Two data types were available for the present study: (i) local-scale predictands such as total daily precipitation, and (ii) eight large-scale model output predictor variables (Table 1). These datasets were further divided into two parts: the first part of the dataset (1979–1996) was used to estimate the statistical parameters of the models considered and the remaining dataset (1997–2001) was used to evaluate the performances of the downscaling models. All metrics were evaluated with a 0.3 mm threshold for dry and wet day occurrences (e.g. Wilby & Harris 2006).

The first model investigated in the downscaling experiment was the SDSM. A MLR model was developed between large-scale predictor variables (Table 1) and the local-scale predictand (precipitation). The parameters of the regression model were estimated using ordinary least

Table 1 | Numerical weather prediction model output variables

Variable Field	Description	Surface level (mb)	Grid
1. *apcp*	Accumulated precipitation (mm)	Surface	Latlon
2. *heating*	Vertically integrated diabatic heating (K/s/mb)	Vertical average	Latlon
3. *pwat*	Precipitable water	Surface	Latlon
4. *prmsl*	Pressure reduced to mean sea-level (Pa)	Surface	Latlon
5. *t2m*	Temperature at 2 m (K)	Surface	Latlon
6. *rhum*	Relative humidity (%)	700 mb	Latlon
7. *u10m*	Zonal wind at 10 m (m/s)	Surface	Latlon
8. *v10m*	Meridional wind at 10 m (m/s)	Surface	Latlon

squares. Precipitation was modeled as a conditional process, in which, large-scale circulation patterns and atmospheric moisture variables were used to linearly condition local-scale weather generator parameters such as precipitation occurrence and intensity. The downscaled daily time series was then adjusted for its mean and variance via bias correction and variance inflation factors, respectively, to better accord with observations.

For the case of modeling with partial least squares regression, the SIMPLS algorithm described in de Jong (1993) was used. The performance of the model was studied over a range of latent variables (factors) that should be retained in order to maximize the amount of variation explained in large-scale predictors relevant for predicting station precipitation. Five latent variables yielded adequate results in the present study. Similarly, the performances of hybrids in downscaling station precipitation were studied. The sensitivity of these models to various choices of threshold probability (p) for the occurrence of wet and dry days were analyzed, and the value of $p = 0.5$ was found to offer optimal results for both PLS-Logst and ANN-Logst models.

The other models considered in the present study were K-nearest neighbors. For modeling with KNN, days similar to each of the 8401 days in the archive were identified using the KNN algorithm as described in Gangopadhyay *et al.* (2005) and Yates *et al.* (2003). The parameters associated with the KNN model were also examined. Two model parameters were considered: moving window (w), and optimal number of nearest neighbors (K). After analyzing the sensitivity of the KNN model to different choices of width of moving window, a value of $w = 14$ days was chosen for use in the downscaling experiment. Similarly, an analysis was performed to find out the optimal value of K, and a value of $K = 19$ was found to offer adequate results in the present study.

To perform modeling using the MLP model, appropriate network architecture was first designed and the various parameters were optimized. The search for an optimal network began with a simple network having one hidden layer, keeping the number of hidden neurons between two and 30. The different learning rules as well as transfer functions were investigated in both the hidden layer and output layer in the search for an optimal network (Muluye & Coulibaly 2007). For the case of modeling with the MLP, one hidden layer with 14 neurons provided the best performing network. Hyperbolic tangent and linear activation function in the hidden layer and output layer, respectively, and the learning rule with a conjugate gradient yielded the optimal network. Similarly, for the case of modeling with the RMLP, one hidden layer with 12 neurons provided the optimal network. In this study, the MLP model was trained with the conventional back-propagation algorithm, whereas the RMLP model was trained with the back-propagation through time algorithm via the decoupled version of an extended Kalman filter (Muluye 2011c).

COMPARISON OF DOWNSCALING RESULTS

Diagnostic measures

Diagnostic measures use logistical metrics in order to evaluate the quality of the forecast system. Two types of diagnostic tools were used: generic and categorical. Wet- and dry-spell length and quantile–quantile (q–q) plots of observed and downscaled daily precipitation amounts were also used. The general statistics employed here provide key statistics such as bias (%), root mean squared error (RMSE), Pearson correlation coefficient (r) and reduction of variance (RV). On the other hand, categorical statistics provided the skill of downscaling models on one of two key features of precipitation (i.e. occurrence) using frequency bias (BiasF), probability of detection (POD) and false alarm ratio (FAR). The contingency table was used to determine the different types of errors made in the downscaled precipitation. Obviously, a perfect forecast system produces only hits and correct rejections, and no misses or false alarms (Stanski *et al.* 1989). Several categorical statistics can be computed from the elements in the contingency table. In this paper, only the major score statistics are computed and discussed.

The BiasF explains how the forecast frequency of 'yes' events compared to the observed frequency of 'yes' events. The range of the score is between zero and infinity, for which a score of one represents a perfect forecast. Basically, the BiasF signifies whether the forecast system has a tendency to underestimate (BiasF < 1) or overestimate (BiasF > 1) occurrences but does not quantify how well

the forecasts correspond to observations; i.e. BiasF only measures relative frequencies (Stanski *et al.* 1989). The Bias score is given by Wilks (1995):

$$\text{BiasF} = \frac{(N_{11} + N_{10})}{(N_{11} + N_{01})} \tag{3}$$

where N_{11} is the number of correct wet days, N_{10} is modeled wet and observed dry days, and N_{01} is modeled dry and observed wet days.

The POD explains the fraction of the observed 'yes' events with correct forecasts. The range of the score is between 0 and one, with a score of one representing a perfect forecast. The POD is sensitive to hits, but ignores false alarms. For this reason, the POD is usually used in conjunction with the FAR (Stanski *et al.* 1989), and is given by Wilks (1995):

$$\text{POD} = \frac{(N_{11})}{(N_{11} + N_{01})} \tag{4}$$

Conversely, the FAR explains the fraction of the predicted 'yes' events which did not occur. The range of the score is between 0 and one, with a score of zero representing a perfect forecast. The FAR is sensitive to false alarms, but ignores hits, and is given by Wilks (1995):

$$\text{FAR} = \frac{(N_{10})}{(N_{11} + N_{10})} \tag{5}$$

The downscaled precipitation was further assessed using a deterministic skill score (SS). The SS represents the improvement in the downscaled precipitation over some reference forecast. Climatology was employed as the reference forecast in this work. When the mean squared error (MSE) is used as a score in the SS computation, the resulting statistic is referred to as a reduction of variance (RV). The RV is given by Stanski *et al.* (1989):

$$\text{RV} = 1 - \frac{\text{MSE}_{\text{forecast}}}{\text{MSE}_{\text{reference}}} \tag{6}$$

An RV of zero indicates no improvement over the reference forecast, one indicates a perfect forecast, and a negative value indicates that the reference forecast is better than a forecast. A more detailed description of model performance statistics is provided by Stanski *et al.* (1989) and Wilks (1995).

Discussion of results

Table 2 shows the error metrics associated with the various models in downscaling daily precipitation (Prec). In most of the cases, the downscaling models tended to underestimate (negative bias) the mean precipitation amounts. The smallest mean bias (−6.18%) over the 15 forecast ranges was simulated by the RMLP model. The ANN-Logst and PLS-Logst models underestimated the mean precipitation amounts during the first five forecast ranges, and overestimated significantly thereafter. The Pearson correlation coefficients between the downscaled and the observed daily precipitation indicated that the RMLP had an advantage over the other downscaling techniques in the vast majority of cases although the ANN-Logst showed competency beyond the five day forecast. This observation was further supported by the overall RMSE and RV statistics. On the other hand, while the SDSM model performed poorly in terms of these statistics, the KNN-E and the KNN-M models yielded the worst statistics. In comparison, the KNN-M model performed relatively better, yielding smaller RMSE and greater r values, compared to the KNN-E model, in spite of relatively larger biases. Note that the bold numeric statistics in Table 2 represent the best statistics associated with each forecast range.

Figure 1 presents RV plots of the various downscaling models against forecast ranges in downscaling daily precipitation. The comparative RV plots show that the skill of precipitation forecasts was generally poor and inadequate. In particular, the SDSM, KNN-M and KNN-E models performed poorly when compared with the other competing downscaling models. These models yielded RV values less than zero, suggesting lack of skill. Further analysis of Figure 1 indicates that there is a decreasing trend in skill with forecast lead time. This clearly underscores the great difficulties associated with the long-range precipitation forecasting. Overall, the RV statistics associated with the RMLP, ANN-Logst and PLS models yielded relatively better skills for all forecast ranges.

Table 2 | Comparative performance statistics for downscaling daily precipitation. The bold numeric statistics represent the best statistics corresponding to each forecast range

Model	Forecast range / Diagnostic	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
SDSM	Bias (%)	−9.71	−7.25	−13.44	−11.26	−8.70	−9.71	−11.60	−11.73	**−0.41**	**−5.36**	−12.23	−6.03	−8.95	**−3.46**	−8.89
	RMSE	5.73	6.11	6.33	6.55	6.82	5.73	6.74	6.87	7.23	7.22	7.05	7.06	6.92	6.99	7.20
	r	0.40	0.27	0.18	0.14	0.07	0.40	0.05	0.07	0.05	0.00	0.00	0.02	0.01	0.04	−0.01
	RV	−0.04	−0.18	−0.27	−0.36	−0.47	−0.04	−0.44	−0.49	−0.65	−0.65	−0.57	−0.58	−0.52	−0.55	−0.64
	POD	0.62	0.59	0.59	0.58	0.56	0.62	0.51	0.49	0.52	0.52	0.52	0.54	0.51	0.53	0.5
	FAR	0.37	0.42	0.41	0.44	0.45	0.37	0.49	0.48	0.49	0.5	0.49	0.49	0.5	0.5	0.49
	BiasF	**0.99**	**1.02**	**1**	**1.03**	**1.01**	**0.99**	**1.01**	**0.94**	**1.01**	**1.04**	**1**	**1.07**	**1.02**	**1.06**	**0.98**
PLS	Bias (%)	−12.01	−10.49	−10.82	−11.28	−10.38	−8.82	−8.87	−9.33	−9.61	−9.46	−9.39	−8.96	−8.64	−8.93	−8.3
	RMSE	4.54	4.87	5.19	5.3	5.43	5.55	5.57	5.6	5.6	5.6	5.6	5.6	5.59	5.59	5.58
	r	0.61	0.51	0.39	0.35	0.27	0.16	0.14	0.1	0.1	0.09	0.1	0.09	0.11	0.12	0.12
	RV	0.35	0.25	0.15	0.11	0.07	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	POD	**0.93**	**0.96**	**0.98**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	FAR	0.44	0.46	0.49	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	BiasF	1.66	1.78	1.9	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99
PLS-Logst	Bias (%)	−8.45	−5.73	−3.17	−6.56	−3.33	**2.4**	16.97	22.42	16.83	6.95	25.58	7.24	17.37	21.82	38.07
	RMSE	4.53	4.88	5.26	5.41	5.62	5.83	5.9	5.95	6.01	6.04	5.97	5.98	5.95	5.9	5.95
	r	0.59	0.5	0.39	0.33	0.25	0.16	0.13	0.1	0.08	0.07	0.09	0.1	0.11	0.12	0.1
	RV	0.35	0.25	0.12	0.07	0	−0.07	−0.1	−0.12	−0.14	−0.16	−0.13	−0.13	−0.12	−0.1	−0.12
	POD	0.69	0.68	0.67	0.61	0.61	0.61	0.67	0.69	0.63	0.57	0.67	0.56	0.63	0.66	0.74
	FAR	**0.22**	**0.26**	**0.3**	**0.36**	**0.39**	**0.43**	**0.45**	**0.46**	**0.48**	**0.48**	**0.49**	**0.48**	**0.48**	**0.48**	**0.49**
	BiasF	0.89	0.92	0.95	0.95	0.99	1.06	1.23	1.29	1.21	1.09	1.31	1.09	1.21	1.28	1.45
KNN-E	Bias (%)	−14.29	−8.13	−8.91	0.41	5.15	−4.21	−14.35	−7.89	−6.67	−13.89	−11.73	−12.66	−7.2	−7.99	−17.46
	RMSE	7.13	7.63	7.35	7.21	7.97	7.68	7.08	7.58	7.58	7.07	7.31	7.31	7.6	7.55	7.07
	r	0.03	−0.01	0.03	0.07	0.05	−0.01	0.03	0.02	0.01	0.03	0.03	0.01	0	0.03	0.02
	RV	−0.63	−0.87	−0.73	−0.66	−1.04	−0.89	−0.6	−0.84	−0.84	−0.6	−0.71	−0.71	−0.85	−0.83	−0.6
	POD	0.5	0.5	0.52	0.54	0.52	0.53	0.55	0.53	0.5	0.5	0.52	0.49	0.5	0.51	0.47
	FAR	0.49	0.5	0.48	0.48	0.5	0.51	0.46	0.49	0.5	0.5	0.47	0.52	0.49	0.48	0.5
	BiasF	**0.98**	**1.02**	**1.01**	**1.05**	**1.03**	**1.07**	**1**	**1.03**	**1**	**0.99**	**0.98**	**1.02**	**0.98**	**0.98**	**0.94**

(*Continued*)

Table 2 | Continued

Model	Forecast range / Diagnostic	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
KNN-M	Bias (%)	−13.58	−17.02	−14.2	−8.79	−15.07	−6.54	−12.47	−7.39	−9.65	−9.45	−15.44	−10.58	−8.48	−8.23	−5.31
	RMSE	6.33	6.35	6.61	7.02	7.5	7.75	7.25	7.54	7.39	7.32	7.17	7.33	7.4	7.27	7.56
	r	0.27	0.24	0.17	0.17	0.03	0.03	0.03	−0.01	0.06	0	0.03	0.01	0.02	0.06	0.01
	RV	−0.28	−0.29	−0.4	−0.58	−0.8	−0.92	−0.68	−0.82	−0.75	−0.71	−0.65	−0.72	−0.76	−0.69	−0.83
	POD	0.58	0.61	0.56	0.58	0.53	0.53	0.53	0.5	0.53	0.53	0.51	0.53	0.52	0.53	0.5
	FAR	0.42	0.39	0.44	0.43	0.46	0.48	0.48	0.51	0.48	0.49	0.48	0.5	0.47	0.48	0.51
	BiasF	**0.99**	**1**	**1**	**1.02**	**0.99**	**1.01**	**1.02**	**1.02**	**1.02**	**1.04**	**0.98**	**1.07**	**0.99**	**1.03**	**1.02**
ANN-Logst	Bias (%)	−10.57	**−5.7**	**−2.05**	−7.14	**−3.16**	3.31	14.98	21.94	15.51	6.94	23.34	8.1	17.69	21.76	36.47
	RMSE	4.56	4.86	5.26	5.4	5.59	5.84	5.88	5.95	6.02	6.06	5.96	5.99	5.97	5.92	5.93
	r	0.59	0.51	0.39	0.33	0.26	0.16	0.13	0.1	0.07	0.06	0.09	0.1	0.11	0.12	0.11
	RV	0.34	0.25	0.13	0.08	0.01	−0.08	−0.1	−0.12	−0.15	−0.16	−0.13	−0.14	−0.13	−0.11	−0.11
	POD	0.69	0.68	0.67	0.61	0.61	0.61	0.67	0.69	0.63	0.56	0.67	0.56	0.63	0.66	0.74
	FAR	**0.22**	**0.26**	**0.3**	**0.36**	**0.39**	**0.43**	**0.45**	**0.46**	**0.48**	**0.48**	**0.49**	**0.48**	**0.48**	**0.48**	**0.49**
	BiasF	0.89	0.92	0.95	0.95	0.99	1.06	1.23	1.29	1.21	1.09	1.31	1.09	1.21	1.28	1.45
MLP	Bias (%)	−10.37	−12.61	−12.18	−10.99	−11.13	−8.61	−8.12	−9.67	−10.39	−10.63	−10.02	−8.39	−8.21	−8.84	−4.13
	RMSE	4.58	4.86	5.21	5.32	5.44	5.55	5.57	5.58	5.64	5.63	5.6	5.6	5.58	5.61	5.68
	r	0.59	0.52	0.38	0.33	0.25	0.16	0.13	0.12	0.06	0.06	0.09	0.09	0.12	0.07	0.04
	RV	0.34	0.25	0.14	0.1	0.06	0.02	0.02	0.01	−0.01	0	**0.01**	**0.01**	**0.01**	0	−0.02
	POD	**0.97**	**0.98**	**1**	**1**	**0.98**	**0.99**	**0.99**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.97**
	FAR	0.47	0.49	0.49	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	BiasF	1.83	1.91	1.97	1.98	1.93	1.97	1.97	1.99	1.98	1.98	1.99	1.99	1.99	1.99	1.94
RMLP	Bias (%)	**−7.81**	−7.68	−7.59	−10.23	−8.58	−5.61	**−5.29**	**−5.44**	**−4.96**	−6.43	**−5.24**	**−4.24**	**−4.39**	−5.3	**−3.92**
	RMSE	**4.39**	**4.77**	**5.11**	**5.26**	**5.41**	**5.52**	**5.56**	**5.59**	**5.58**	**5.59**	**5.59**	**5.59**	**5.59**	**5.59**	**5.58**
	r	**0.63**	**0.53**	**0.42**	**0.36**	**0.27**	**0.19**	**0.14**	**0.11**	**0.11**	**0.1**	**0.1**	**0.11**	**0.11**	**0.11**	**0.12**
	RV	**0.39**	**0.28**	**0.17**	**0.12**	**0.07**	**0.03**	**0.02**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
	POD	**0.92**	**0.94**	**0.93**	**0.97**	**0.99**	**0.99**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	FAR	0.42	0.43	0.46	0.48	0.49	0.49	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	BiasF	1.6	1.64	1.71	1.85	1.94	1.96	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99

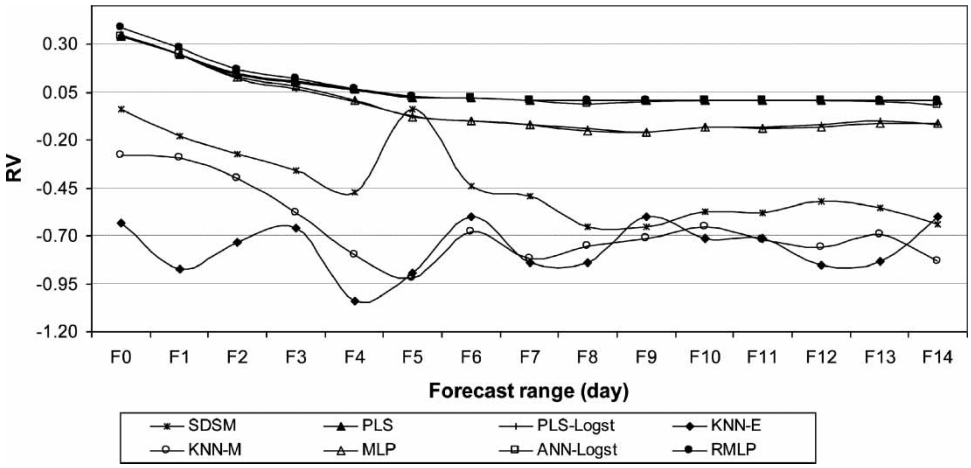

Figure 1 | Reduction of variance (RV) with forecast range (FR) as downscaled by various models, 1997–2001.

Table 2 shows the BiasF simulated by the various models. The BiasF of wet periods simulated by the SDSM, KNN-E and KNN-M models were reasonably accurate (BiasF close to one). The PLS, MLP and RMLP models consistently overestimated the wet periods, by more than 90% on average over the 15 forecast ranges. It is interesting to note that the coupled models (PLS-Logst and ANN-Logst) appeared to perform better than their counterparts (PLS and MLP) in terms of representing the occurrence of wet periods. On average, only about 13% of the wet periods were overestimated by the coupled models, over the 15 forecast ranges. Model performance was further evaluated through the POD and the FAR statistics. The ability of the RMLP, MLP and PLS models to capture the POD statistics were generally adequate (POD close to unity), however about 50% of the cases were incorrectly forecasted (Table 2). In comparison, the smallest FAR (on average 42%) was attained by the PLS-Logst and the ANN-Logst models although some difficulties were clearly noticed in simulating the POD (on average 65%).

While the performances of the RMLP, MLP and PLS models were reasonably adequate in representing the POD, very poor performances were observed in representing the BiasF. This can be explained in part by the structure of the model and the technique used in modeling the precipitation process. In the case of modeling with the SDSM, PLS-Logst and ANN-Logst, precipitation was modeled as a conditional process, in which, local precipitation amounts were correlated with the occurrences of wet days and, which in turn, were correlated with large-scale atmospheric predictors; and the KNN was modeled on the principle of random sampling techniques. Such methods enabled these models to preserve the stochasticity of daily precipitation, thus permitting the occurrence of both dry and wet days. Conversely, the MLP, RMLP and PLS models performed poorly in this regard, due to direct linkage or conditioning of large-scale atmospheric predictors with precipitation amounts and occurrences, thereby leaving few dry days to occur. This phenomenon is clearly evident from the relative superior performances demonstrated by their counterparts, particularly on BiasF and FAR statistics. Furthermore, as discussed in the subsequent sections, these models (i.e. MLP, RMLP and PLS) have great difficulties in characterizing key precipitation features, such as dry-spell, wet-spell and variance statistics in downscaling daily precipitation fields.

The various downscaling models are further intercompared using more general statistics such as mean and variance of daily precipitation. Such generic tests provide the overall performances of the different models over the first and the second order moments. Figure 2 compares the observed and the downscaled daily precipitation amounts for each forecast range. The statistics associated with the observed and the downscaled precipitation values were computed annually. Figure 2(a) shows that all downscaling models (with the exception of ANN-Logst and PLS-Logst) yielded comparable performances in terms of representing the mean observed precipitation. Also, these

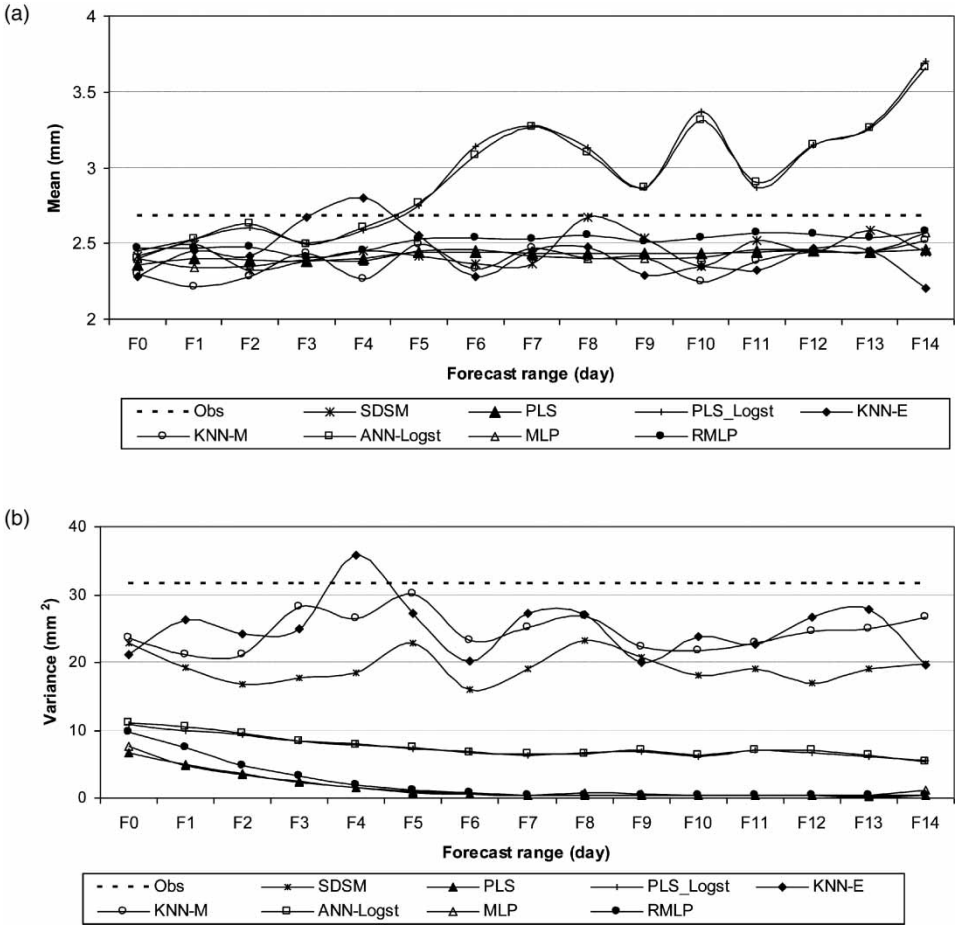

Figure 2 | Comparison of downscaled precipitation derived from numerical forecast model: (a) mean precipitation totals, and (b) variance of daily precipitation, 1997–2001.

models tended to slightly underestimate the mean observed local precipitation. On the other hand, the ANN-Logst and the PLS-Logst models overestimated significantly the mean observed local precipitation, beyond the five day forecast. A similar comparative plot of the variances of the observed and the downscaled daily precipitation amounts is shown in Figure 2(b). Despite a slight and a consistent underestimation of the observed precipitation variance, the KNN-E and the KNN-M models demonstrated the best performance in terms of representing the variability in daily precipitation. The SDSM model also demonstrated a reasonably adequate performance. The remaining downscaling models were, however, unable to capture the variability in daily precipitation and even showed further deterioration with increased forecast ranges. The relative poor performances of these models can be explained in part by the technique used in modeling the precipitation process (i.e. the method has a tendency to smoothen and deflate the variance of daily precipitation). The failure to capture the variability in daily precipitation is one of the acknowledged limitations of most statistical models (e.g. Clark & Hay 2004; Harpham & Wilby 2005).

The average lengths of dry- and wet-spells in each month are important diagnostic measures commonly used for evaluating the accuracy of the downscaled precipitation. In the present study, dry days were considered when days had precipitation intensities of 0.3 mm or less; and dry-spell length in a given month was computed as the maximum number of consecutive dry days in that month (Khan *et al.* 2006). In this study, results are presented only for mean wet-spell lengths, for forecast ranges of 3, 7 and 10. The comparative plots of the average wet-spell lengths for each month of observed and downscaled precipitation are shown in Figure 3. The MLP, RMLP and PLS-based downscaled

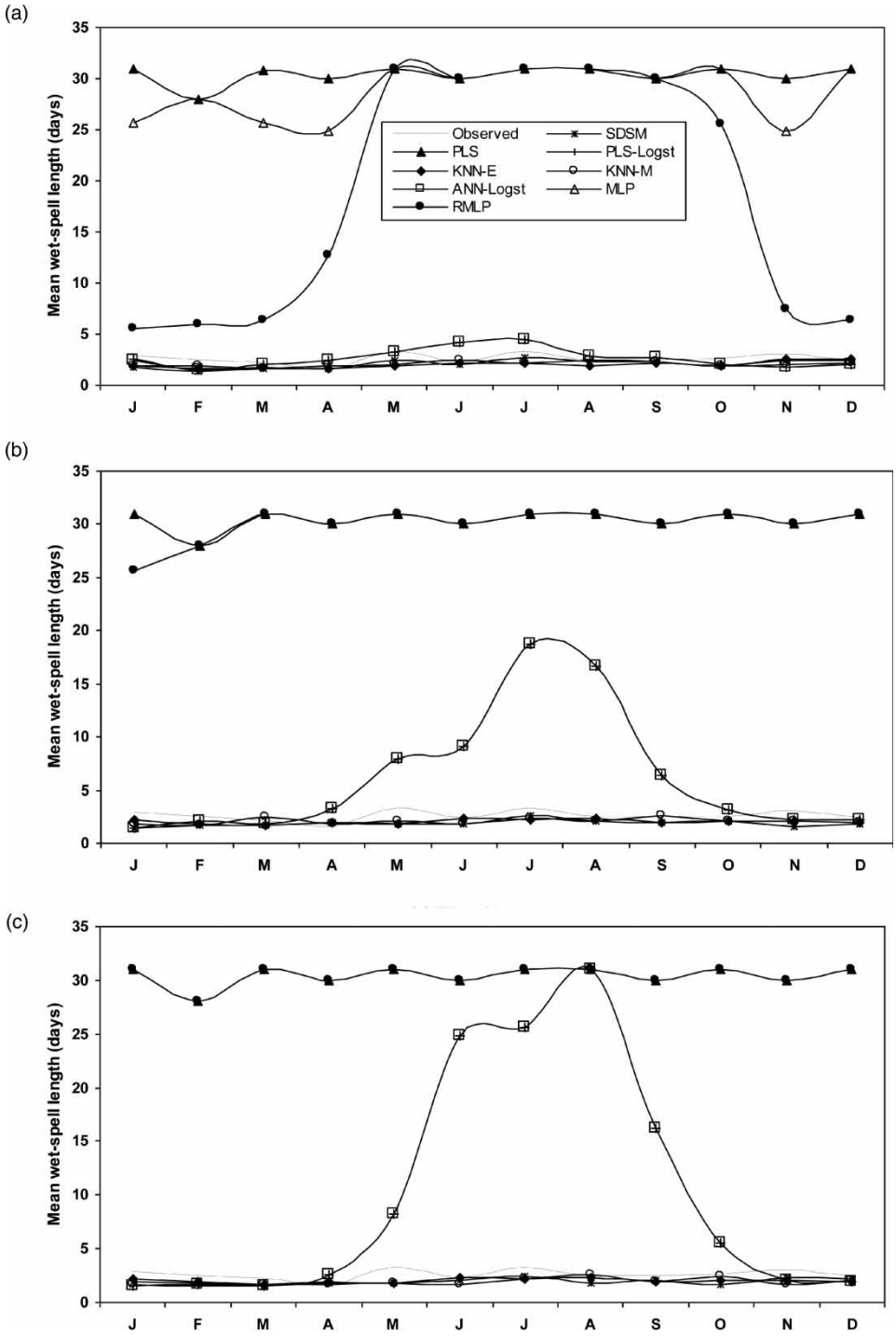

Figure 3 | Mean length of wet-spells derived from numerical forecast model for forecast range (FR) (a) 3, (b) 7, and (c) 10, 1997–2001.

precipitation data appeared to overestimate the mean wet-spells significantly. The other competing models reproduced the portion of wet-spells reasonably, for all months, in spite of a slight underestimation of observed precipitation. The coupled models (PLS-Logst and ANN-Logst) represented the mean wet-spell length reasonably well for the first and

the last few forecast ranges, and significantly overestimated the remaining forecast ranges. Such overestimation is a result of the technique used in modeling the precipitation process, as described earlier. Similar but opposite model performances were observed for mean dry-spell characteristics (results not shown).

Figures 4–7 depict quantile–quantile plots of the observed and the downscaled daily precipitation, for each season (only results of forecast range 7 are shown). The q–q plots were constructed using the quantiles of the observed precipitation and the quantiles of the downscaled precipitation. The purpose of the q–q plot was to determine

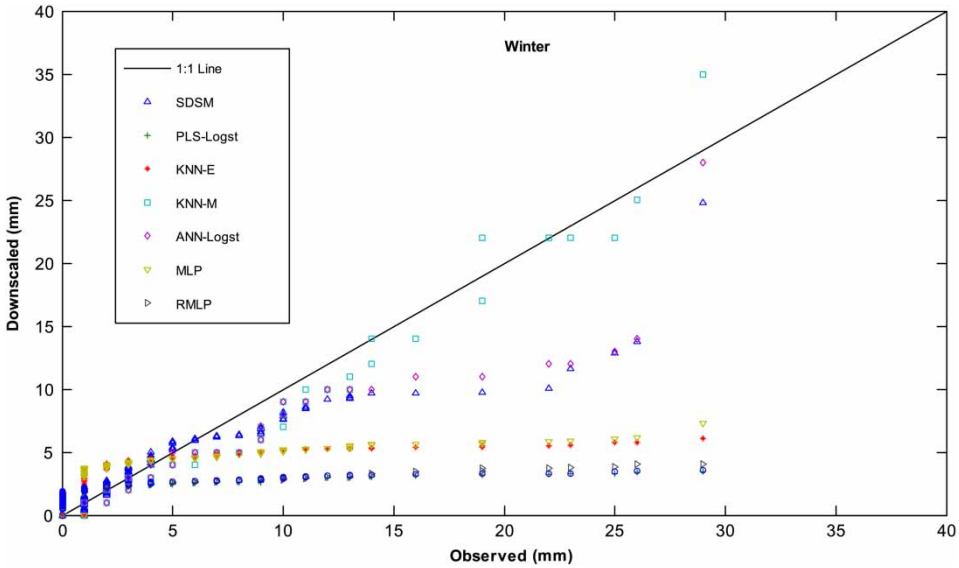

Figure 4 | Quantile–quantile (q–q) plots of the quantiles of the observed precipitation (mm) against the quantiles of the simulated precipitation (mm) as downscaled by the different models, Jan 1997 to Dec 2001, for forecast range of 7 day for winter (JFM).

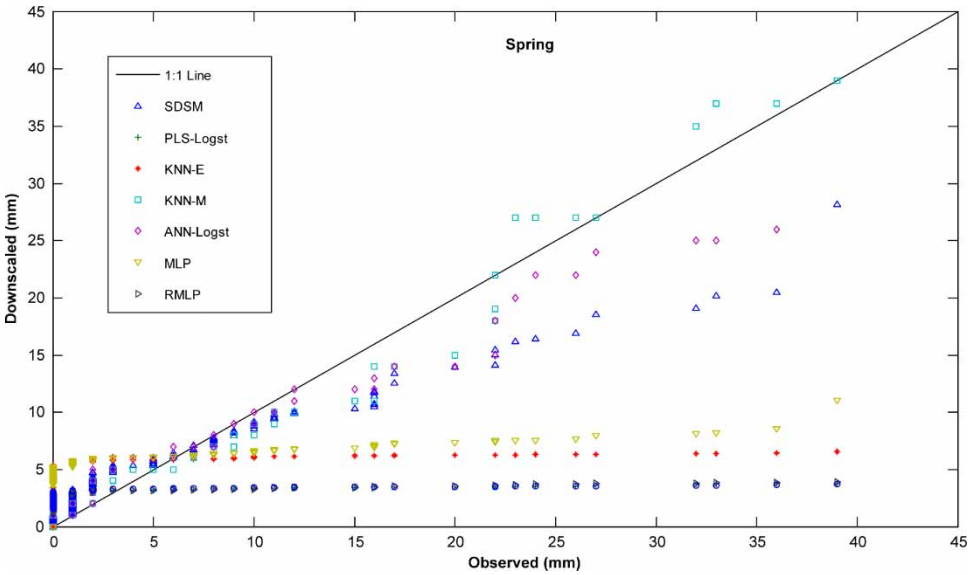

Figure 5 | Quantile–quantile (q–q) plots of the quantiles of the observed precipitation (mm) against the quantiles of the simulated precipitation (mm) as downscaled by the different models, Jan 1997 to Dec 2001, for forecast range of 7 day for spring (AMJ).

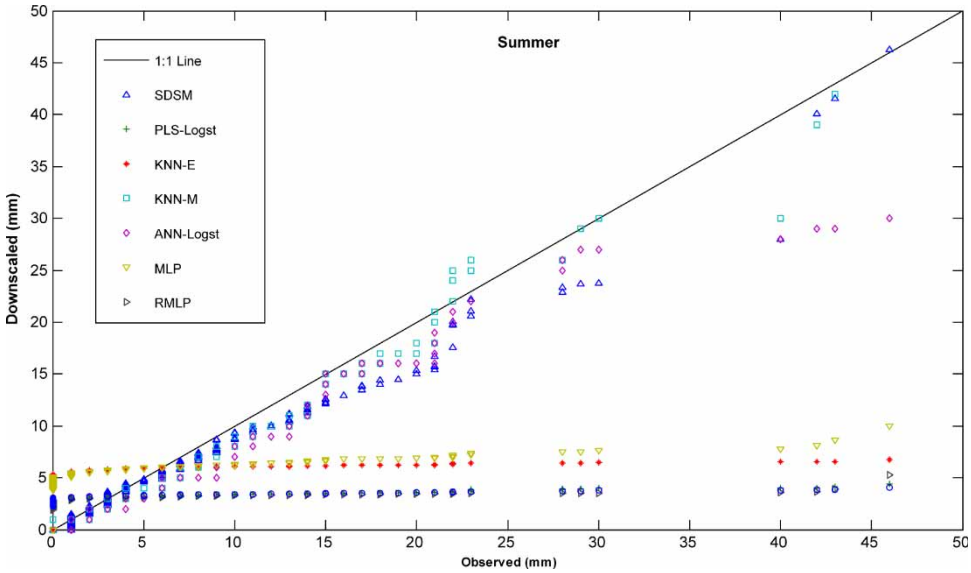

Figure 6 | Quantile–quantile (q–q) plots of the quantiles of the observed precipitation (mm) against the quantiles of the simulated precipitation (mm) as downscaled by the different models, Jan 1997 to Dec 2001, for forecast range of 7 day for summer (JAS).

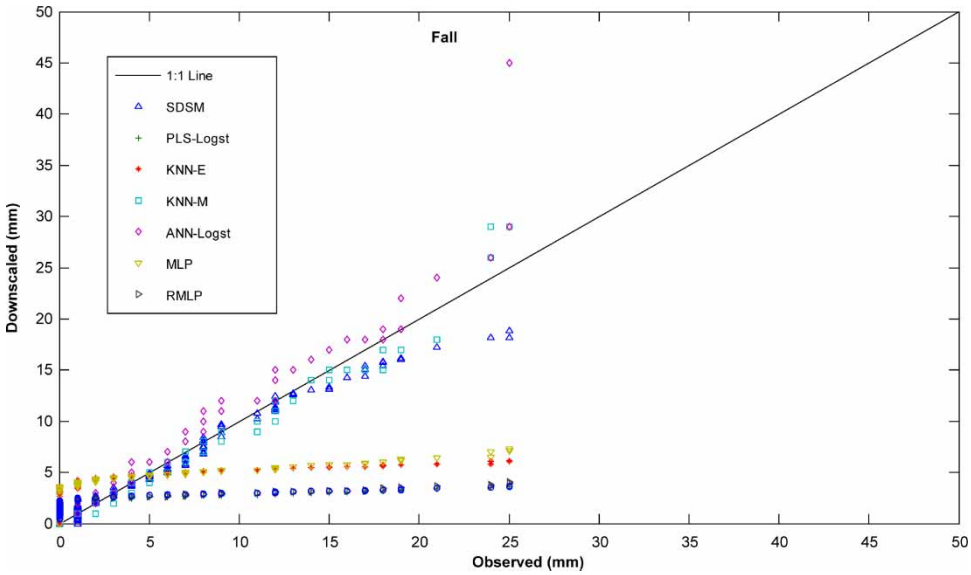

Figure 7 | Quantile–quantile (q–q) plots of the quantiles of the observed precipitation (mm) against the quantiles of the simulated precipitation (mm) as downscaled by the different models, Jan 1997 to Dec 2001, for forecast range of 7 day for fall (OND).

whether the samples of the observed and the downscaled precipitation came from populations with a common distribution. If the samples come from the same distribution then the points of the q–q plot should fall approximately along a 45-degree reference line. Examination of Figures 4–7 indicated a relatively good performance for fall and summer seasons by most downscaling models. In comparison, performance was poor for spring and worst for winter. The SDSM, KNN-E and KNN-M models exhibited superior performances, when compared to the other competing models, the latter two representing the best performance overall. Regardless of the seasons under consideration, all models

(with the exception of SDSM, KNN-E and KNN-M) overestimated dry and light precipitation events, and underestimated significantly heavy events. Performance was particularly poor for heavy events. The demonstrated poor performance could be associated with the method used in the process of modeling precipitation as described earlier.

In addition to the above diagnostic measures, statistical tests were also performed between the observed and the downscaled precipitation to determine if there is evidence of difference in the population locations and variances without assuming a parametric model for the distributions. Two statistical tests were conducted: (i) Mann-Whitney test (Lehmann 1975) of the equality of two population medians; and (ii) Levene's test (Levene 1960) of the equality of two population variances. Table 3 presents the results of the statistical tests conducted. The values in this table represent *p*-values. A significance level of 5% is chosen for the purpose of comparison. A *p*-value greater than the chosen significance level suggests the downscaled and the observed precipitation medians or variances are not significantly different. The two statistical tests were conducted on five of the eight downscaling models investigated in the present study. The MLP, the ANN-Logst and the PLS models were not considered as these models did not show substantial differences in performance when compared to their associate downscaling models.

The results of the Mann-Whitney test (Table 3) confirm that the SDSM, KNN-M and KNN-E-based precipitation medians and the observed precipitation medians are not significantly different for most of the months (reported *p*-values greater than the rejection level of significance). With regard to the other downscaling models, except for a few months in fall and winter, the observed and the downscaled precipitation medians are significantly different. The results of the Levene's test show that the reported *p*-values associated with the SDSM, KNN-E and PLS-Logst models were greater than the rejection level of significance for most of the months, suggesting the variances of the observed and the downscaled precipitation are not significantly different (Table 3). The RMLP model generally reproduced the observed variances reasonably well, except for September and December where the downscaled and the observed precipitation variances are significantly different. The *p*-values associated with the KNN-M model were, however, less than the chosen level of significance for most of the months (except for June, August and September) indicating the observed and the downscaled precipitation variances are significantly different.

Table 3 | Test for position (i.e. median) and variance using Mann-Whitney and Levene, respectively, for the test period Jan 97 to Dec 01. The dataset used for the statistical test corresponds to the first member and zero forecast range. Note that the values in this table represent *p*-values

	Test for Position (Mann-Whitney)					Test for Variance (Levene)				
Month	**SDSM**	**RMLP**	**PLS-Logst**	**KNN-M**	**KNN-E**	**SDSM**	**RMLP**	**PLS-Logst**	**KNN-M**	**KNN-E**
Jan	0.219	0.158	0.375	0.020	0.950	0.357	0.517	0.483	0.000	0.986
Feb	0.156	0.226	0.067	0.971	0.026	0.221	0.147	0.128	0.003	0.237
Mar	0.249	0.000	0.605	0.264	0.467	0.578	0.291	0.327	0.001	0.635
Apr	0.059	0.000	0.522	0.939	0.010	0.917	0.973	0.423	0.000	0.651
May	0.071	0.000	0.028	0.542	0.231	0.080	0.903	0.260	0.001	0.016
Jun	0.411	0.000	0.004	0.535	0.115	0.855	0.896	0.576	0.141	0.394
Jul	0.703	0.000	0.004	0.852	0.128	0.244	0.191	0.505	0.013	0.851
Aug	0.156	0.226	0.067	0.982	0.290	0.956	0.014	0.363	0.250	0.951
Sep	0.315	0.038	0.732	0.826	0.224	0.851	0.035	0.193	0.571	0.982
Oct	0.854	0.001	0.641	0.501	0.336	0.934	0.101	0.503	0.472	0.734
Nov	0.921	0.756	0.067	0.476	0.873	0.650	0.065	0.141	0.000	0.971
Dec	0.604	0.538	0.262	0.541	0.860	0.480	0.027	0.089	0.000	0.813

SUMMARY AND CONCLUSIONS

The present study undertook a rigorous inter-comparison of daily precipitation statistics as downscaled by eight different statistical downscaling models for the Chute-du-Diable sub-basin located in northeastern Canada. A 'reforecast' dataset generated by a fixed numerical model – a 1998 version of NCEP's GFS model – was used in the present study. An ensemble of 15-day forecasts over a 23-year period from 1979 through 2001 was available for analysis. Eight model output variables corresponding to the first ensemble member (i.e. out of 15) and all 15 forecast ranges were considered in the downscaling experiment. Concurrent observed station daily precipitation was collected from Alcan Company. The dataset from 1979–1996 was used to calibrate the models, and the remaining dataset from 1997–2001 was used to evaluate the performances of the downscaling models. A range of standard suites of diagnostic measures were employed to evaluate and inter-compare the downscaling results.

The MLR and PCR models are the two common statistical methods used to link the large-scale model output to station-scale variables. The PLS regression was, however, used in the present study as it generalizes and combines good features from both models. In order to properly characterize precipitation occurrences by the PLS and the ANN models, hybrid models were proposed. In these hybrid models, precipitation was modeled in a two stage process: logistic regression was used to identify the occurrence of wet days, and the PLS and the ANN models were used to model the amounts. The hybrid models generally showed improvements in representing precipitation occurrence, but their overall skills were not superior to their counterparts.

To link the dynamics of the large-scale predictors to station-scale precipitation, two ANN-based models were investigated. In comparison, the RMLP model outperformed the MLP model in the vast majority of model evaluation measures. The demonstrated performance of the RMLP was not a coincidence given its strong mathematical foundation. The RMLP model used in the present study utilized an extended Kalman filter approach to perform supervised training. The improved statistics shown in this study was in part due to the optimal filtering capability of the model.

The other statistical models considered in downscaling daily precipitation were nearest neighbors. Two types of KNN-based model were designed on the basis of distance formulation. In general, the KNN model which used Euclidean distance yielded less skill and showed poor statistics, compared to the KNN model which used Mahalanobis distance. Compared to the other competing models, both KNN-based models exhibited the best performance in terms of representing the variability of daily precipitation.

The present study considered a number of statistical models in downscaling daily precipitation in order to provide better insights into the nature of the problem. In view of the downscaling results, the following general conclusions and recommendations have been made.

(i) The skill values in the downscaled daily precipitation were generally inadequate.
(ii) Large differences were observed in performance from season to season. Performance was better in fall followed by summer. On average, winter was the season with the lowest skill values based on the q–q plot.
(iii) A considerable distinction was observed among models in terms of reproducing the variability in daily precipitation. The KNN and the SDSM models showed a clear advantage over their counterparts in reproducing the variability in daily precipitation.
(iv) No method has consistently shown superior performance in terms of all diagnostic measures. Nevertheless, a clear pattern emerged with respect to the reproduction of variations in daily precipitation and RV values. The PLS, PLS-Logst, ANN-Logst, MLP and RMLP models showed modest RV values, whereas the SDSM and KNN models showed considerable potential to capture the variability in daily precipitation.
(v) The performance of models may be quite different from season to season and from region to region. Thus, care should be exercised while models are inter-compared.

(vi) The analysis and discussion in this study were based on only one station. The performance of the different statistical downscaling models could respond differently than what has been observed in the present study when these models are applied to stations with different geographic, hydrologic and climatic characteristics.

ACKNOWLEDGEMENTS

This research was supported by the School of Graduate Studies at McMaster University. Drs Brian Baetz and Sarah Dickson are gratefully acknowledged for their proof reading, review comments and suggestions. The author is grateful to Dr Paulin Coulibaly and to Alcan Company for making the experiment data available. The author is also grateful to Dr Noel Evora for providing the pre processed ensemble weather predictors for the study area. The ensemble reforecast data were made available by NOAA at: http://www.cdc.noaa.gov/reforecast/. Minitab Statistical Software was used to conduct statistical tests. The author is grateful to the Editor-in-Chief and the two anonymous reviewers whose suggestions strongly improved the quality of the original manuscript.

REFERENCES

ASCE 2000 Task committee on application of artificial neural network in hydrology. artificial neural network in hydrology. I: preliminary concepts. *J. Hydrol. Eng.* **5**, 115–123.

Buishand, T. A. & Brandsma, T. 2001 Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Water Resour. Res.* **37**, 2761–2776.

Choi, J., Yeap, T. H. & Bouchard, M. 2005 Online state–space modeling using recurrent multilayer perceptrons with unscented kalman filter. *Neural Process. Lett.* **22**, 69–84.

Clark, M. P. & Hay, L. E. 2004 Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *J. Hydrometeorol.* **5**, 15–32.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. & Wilby, R. 2004 The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5**, 243–262.

de Jong, S. 1993 SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **18**, 251–263.

Dibike, Y. B. & Coulibaly, P. 2005 Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models. *J. Hydrol.* **307**, 145–163.

Fowler, H. J., Blenkinsopa, S. & Tebaldib, C. 2007 Review linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modeling. *Int. J. Climatol.* **27**, 1547–1578.

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

Harpham, C. & Wilby, R. L. 2005 Multi-site downscaling of heavy daily precipitation occurrence and amounts. *J. Hydrol.* **312**, 235–255.

Hay, L. E. & Clark, M. P. 2003 Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States. *J. Hydrol.* **282**, 56–75.

Hay, L. E., Clark, M. P., Wilby, R. L., Gutowski, W. J., Arritt, R. W., Takle, E. S., Pan, Z. & Leavesley, G. H. 2002 Use of regional climate model output for hydrologic simulations. *J. Hydrometeor.* **3**, 571–590.

Haykin, S. 2008 *Neural Networks and Learning Machines*. 3rd edition. Prentice-Hall, Upper Saddle River, NJ.

Hua, Z. & Zhang, B. 2006 A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts. *Appl. Math. Compt.* **18**, 1035–1048.

Khan, M. S., Coulibaly, P. & Dibike, Y. 2006 Uncertainty analysis of statistical downscaling methods. *J. Hydrol.* **319**, 357–382.

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

Levene, H. 1960 Robust tests for the equality of variance. In: *Contributions to Probability and Statistics* (I. Olkin, ed.). Stanford University Press, Palo Alto, California, USA, pp. 278–292.

Li, H., Luo, L., Wood, E. F. & Schaake, J. 2009 The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *J. Geophys. Res.* **114**, D04114.

Liu, X., Coulibaly, P. & Evora, N. 2007 Comparison of data-driven methods for downscaling ensemble weather forecasts. *Hydrol. Earth Syst. Sci. Discuss.* **4**, 189–210.

Lorber, A., Wangen, L. E. & Kowalski, B. R. 1987 A theoretical foundation for the PLS algorithm. *J. Chemometrics* **1**, 19–31.

Mearns, L. O., Bogardi, I., Giorgi, F., Matyasovszky, I. & Palecki, M. 1999 Comparison of climate change scenarios generated from regional climate model experiments and statistical downscaling. *J. Geophys. Res.* **104**, 6603–6621.

Mehrotra, R. & Sharma, A. 2005 A nonparametric nonhomogeneous hidden Markov model for downscaling of multi-site daily rainfall occurrences. *J. Geophys. Res.* **110**, D16108.

Mehrotra, R. & Sharma, A. 2006 Conditional resampling of hydrologic time series using multiple predictor variables: A k-nearest neighbour approach. *Adv. Water Resour.* **29**, 987–999.

Mehrotra, R., Sharma, A. & Cordery, I. 2004 Comparison of two approaches for downscaling synoptic atmospheric patterns to multisite precipitation occurrence. *J. Geophys. Res –Atmospheres* **109**, D14107.

Muluye, G. Y. 2011a Implications of medium-range numerical weather model output in hydrologic applications: Assessment of skill and economic value. *J. Hydrol.* **400**, 448–464.

Muluye, G. Y. 2011b Deriving meteorological variables from numerical weather prediction model output: a nearest neighbor approach. *Water Resour. Res.*

Muluye, G. Y. 2011c Improving long-range hydrological forecasts with extended Kalman filters. *Hydrol. Sci. J.* **56**, 1118–1128.

Muluye, G. Y. & Coulibaly, P. 2007 Seasonal reservoir inflow forecasting with low frequency climatic indices: a comparison of data-driven methods. *Hydrol. Sci. J.* **52**, 508–522.

Palma, L., Gil, P., Henriques, J., Dourado, A. & Duarte-Ramos, H. 2001 Application of an Extended Kalman Filter for On-line Identification with Recurrent Neural Networks. In *Proc. of the 7th Jornadas Hispano-Lusas de Ingenieria Electrica*, Madrid, Spain.

Puskorius, G. V. & Feldkamp, L. A. 1994 Neurocontrol of nonlinear dynamical systems with kalman filter trained recurrent networks. *IEEE Trans. Neural Networks* **5**, 279–297.

Rajagopalan, B. & Lall, U. 1999 A k-nearest neighbour simulator for daily precipitation and other weather variables. *Water Resour. Res.* **35**, 3089–3101.

Roulin, E. 2007 Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrol. Earth Syst. Sci.* **11**, 725–737.

Schaake, J., Franz, K., Bradley, V. & Buizza, R. 2006 The hydrologic ensemble prediction EXperiment (HEPEX). *Hydrol. Earth Syst. Sci.* **3**, 3321–3332.

Schmidli, J., Goodess, C. M., Frei, C., Haylock, M. R., Hundecha, Y., Ribalaygua, J. & Schmith, T. 2007 Statistical and dynamical downscaling of precipitation: an evaluation and comparison of scenarios for the European Alps. *J. Geophys. Res.* **112**, D04105.

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

Sharif, M. & Burn, D. 2006 Simulating climate change scenarios using an improved K-nearest neighbor model. *J. Hydrol.* **325**, 179–196.

Shi, X., Wood, A. W. & Lettenmaier, D. P. 2008 How essential is hydrologic model calibration to seasonal streamflow forecasting. *J. Hydrometeorol.* **9**, 1350–1363.

Spak, S., Holloway, T., Lynn, B. & Goldberg, R. 2007 A comparison of statistical and dynamical downscaling for surface temperature in North America. *J. Geophys. Res.* **112**, D08101.

Stanski, H. R., Wilson, L. J. & Burrows, W. R. 1989 *Survey of Common Verification Methods in Meteorology*. WMO World Weather Watch Tech. Report No. 8, WMOI TD No. 358, pp. 114.

von Storch, H., Zorita, E. & Cubash, U. 1993 Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. *J. Clim.* **6**, 1161–71.

Wetterhall, F., Bardossy, A., Chen, D., Halldin, S. & Xu, C. Y. 2006 Daily precipitation-downscaling techniques in three Chinese regions. *Water Resour. Res.* **42**, W11423.

Wilby, R. L. & Wigley, T. M. L. 1997 Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography* **21**, 530–548.

Wilby, R. L. & Harris, I. 2006 A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the river Thames, UK. *Water Resour. Res.* **42**, W02419.

Wilby, R. L. & Dawson, C. W. 2007 *SDSM 4.1 – A Decision Support Tool for the Assessment of Regional Climate Change Impacts*. User Manual, UK.

Wilby, R. L., Dawson, C. W. & Barrow, E. M. 2002 SDSM–A decision support tool for the assessment of regional climate impacts. *Environ. Modell. Software* **17**, 145–157.

Wilks, D. S. 1995 *Statisistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, California.

Wold, S., Geladi, P., Esbensen, K. & Ohman, J. 1987 Multi-way principal components and PLS-analysis. *J. Chemometrics* **1**, 41–56.

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

Yakowitz, S. 1993 Nearest neighbor regression estimation for null-recurrent Markov time series. *Stochastic Process. Appl.* **48**, 311–318.

Yarnal, B., Comrie, A. C., Frakes, B. & Brown, D. P. 2001 Developments and prospects in synoptic climatology. *Int. J. Climatol.* **21**, 1923–1950.

Yates, D., Gangopadhyay, S., Rajagopalan, B. & Strzepek, K. 2003 A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resour. Res.* **39**, 1199.

First received 15 December 2011; accepted in revised form 27 March 2012. Available online 15 June 2012