Statistical tool for modeling of a daily precipitation process in the context of climate change

Myeong-Ho Yeo, Hoang-Lam Nguyen and Van-Thanh-Van Nguyen

ABSTRACT

The present study proposes a climate change assessment tool based on a statistical downscaling (SD) approach for describing the linkage between large-scale climate predictors and observed daily rainfall characteristics at a local site. The proposed SD of the daily rainfall process (SDRain) model is based on a combination of a logistic regression model for representing the daily rainfall occurrences and a nonlinear regression model for describing the daily precipitation amounts. A scaling factor (S_R) and correction coefficient (C_R) are suggested to improve the accuracy of the SDRain model in representing the variance of the observed daily precipitation amounts in each month without affecting the monthly mean precipitation. SDRain facilitates the construction of daily precipitation models for the current and future climate conditions. The tool is tested using the National Center for Environmental Prediction re-analysis data and the observed daily precipitation data available for the 1961–2001 period at two study sites located in two completely different climatic regions: the Seoul station in subtropical-climate Korea and the Dorval Airport station in cold-climate Canada. Results of this illustrative application have indicated that the proposed functions (e.g. logistic regression, S_R, and C_R) contribute marked improvement in describing daily precipitation amounts and occurrences. Furthermore, the comparison analyses show that the proposed SD method could provide more accurate results than those given by the currently popular SDSM method.

Key words | assessment tool, climate change, daily precipitation, statistical downscaling

Myeong-Ho Yeo (corresponding author)
Water and Environmental Research Institute of the Western Pacific,
University of Guam,
Mangilao, GU 96923,
USA
E-mail: *yeom@triton.uog.edu*

Hoang-Lam Nguyen
Van-Thanh-Van Nguyen
Department of Civil Engineering and Applied Mechanics,
McGill University,
817 Sherbrooke Street West, Montreal, Quebec,
Canada

INTRODUCTION

Understanding the variations in the precipitation process in time and in space is essential for the planning, design, and management of various water resources systems. For instance, daily/monthly precipitation time series are commonly used for the assessment of the available water resources in a region, and the extreme rainfall amount for a given return period is required for the estimation of flood for the design of hydraulic structures. Recently, climate change has been recognized as having a profound impact on the hydrologic cycle at different temporal and spatial scales (Pachauri *et al.* 2014). Global climate models (GCMs) have been extensively used in many studies for assessing this impact on the precipitation process. However, outputs from these models are not suitable for these hydrological impact studies at a regional or local scale due to too coarse spatial resolutions (generally greater than 200 km). Various downscaling methods have hence been proposed for associating GCM predictions of climate change with hydrologic processes at the desired space and time scales (Yarnal *et al.* 2001; Nguyen & Nguyen 2008). Of particular importance for hydrologic applications are those procedures dealing with the linkage of the large-scale climate variability to the historical observations of the daily precipitation process at a given location or over a given watershed. Once the linkage is established, the projected change of

doi: 10.2166/wcc.2019.403

climate conditions given by a GCM could be used to project the resulting change of the local precipitation and the resulting runoff characteristics.

In general, there are two different types of downscaling approaches (Wilby & Wigley 1997; Xu 1999; Nguyen *et al.* 2006). The first type is dynamical downscaling (DD) methods that are based on high-resolution regional climate models (RCMs) (Laprise 2008). The RCM uses GCM variables as boundary conditions for a specific location and time to capture a higher spatial resolution of 20–50 km. The main advantage of DD is that it is able to provide a physical understanding of synoptic systems and relationships between atmospheric conditions and weather conditions (Yarnal *et al.* 2001). The spatial resolution (20–50 km) is still too coarse to conduct frequency analyses despite their extensive and costly computing requirements. The second type is statistical downscaling (SD) methods that are privileged by their ease of implementation and use. The simple computational requirement by these SD methods permits the consideration of different GCMs in the development of climate change scenarios and their associated uncertainties. Further, the SD techniques could be adapted to the climatic conditions for a specific site based on some established statistical relationships between large-scale atmospheric variables (predictors) and local weather variables (predictands). Through these practical advantages, SD methods have been commonly used in many different climate change impact studies (Nguyen & Nguyen 2008; Yeo 2014).

The SD methods can be further sub-divided into three types with respect to the used statistical techniques: weather typing (Bárdossy 1997; Goodess & Palutikof 1998), stochastic weather generators (Richardson & Wright 1984), and regression-based methods (Kilsby *et al.* 1998; Wilks & Wilby 1999; Wilby *et al.* 2002; Harpham & Wilby 2005). The main characteristic of weather typing techniques is to classify days into some number of discrete weather conditions or states on the basis of synoptic similarity. However, this classification of climatic conditions may not be quite accurate, depending often on some subjective assessments. Stochastic weather generators, such as WGEN (Richardson & Wright 1984), could well describe the statistical characteristics of climate processes at a local scale. The most challenging of them is how to transmit the information from GCMs' outputs to the parameters of these stochastic models. Finally, regression models use empirical relationships between predictors and observed weather variables (e.g. temperature and precipitation) to describe weather conditions (Wilby *et al.* 2002). Their main limitation is the stationary assumption of the regression coefficients, which means that the statistical relationships developed for the current climate hold also under the different climatic conditions of the future climate.

The present study proposes a climate change assessment tool based on regression-based SD methods for describing the statistical linkage between the large-scale climate variables and rainfall characteristics at a local site. In general, there is no general agreement as to which downscaling method is the most suitable approach for accurately describing the observed precipitation characteristics for a given study site, depending mainly on the specific study objectives and on the specific climatology of the particular study area (Nguyen & Nguyen 2008). Because regression models reflect the information from GCMs' outputs to local weather conditions with ease, the tool is developed based on two different regression approaches. In particular, the suggested SD model is a combination of a logistic regression model for representing the daily rainfall occurrences and a nonlinear regression model for describing the daily precipitation amounts. In this study, the proposed assessment tool is tested using the US National Center for Environmental Prediction (NCEP) re-analysis data and the observed daily precipitation data available for the 1961–2001 period from two raingauge networks located in two completely different climatic regions: the Seoul station in subtropical-climate South Korea and the Dorval Airport station in cold-climate Canada.

SD OF DAILY RAINFALL PROCESS – SDRAIN

Modeling of the daily precipitation occurrence process

For modeling the precipitation occurrences, two common types of procedures have been used in general: (1) the use of *Markov chain* for representing the wet or dry status of a given day (Gabriel & Neumann 1962; Wilks 1998; Serinaldi 2009; Mehrotra & Sharma 2010) and (2) the application of *alternating renewal processes* to represent the relative frequencies of wet- and dry-day spells (Sharma & Lall 1999). In the context of climate change, it is a well-known issue to link the parameters of these Markov chain and alternating renewal processes to the GCM climate predictors. The multiple linear regression method as an alternative has often been employed to describe the occurrence process with large-scale climate predictors (Wilby *et al.* 2002; Cannon 2008; Hessami *et al.* 2008). Nevertheless, the application of the linear regression model to binary variables (such as precipitation occurrence process) is highly likely problematic as it violates the strongly required assumptions of the linear regression model as indicated in the following. Hence, it is more appropriate to suggest an alternative model based on the logistic regression for describing more accurately the statistical characteristics of the binary precipitation occurrences using the large-scale climate predictors given by GCMs.

Let O_i be the random variable representing the daily precipitation occurrence ($O_i = 0$ if day i is dry and $O_i = 1$ if day i is wet). The probability of a wet day is prob(wet at day i) $= \pi_{i,}$ and the probability of a dry day is prob(dry at day i) $= 1 - \pi_i$. The expected value of a wet day, E(wet$_i$), can be computed as follows:

$$\mathrm{E}(\mathrm{wet}_i) = 0 \times (1 - \pi_i) + 1 \times (\pi_i) \tag{1}$$

or using the climate predictor variables, X_i, it can be calculated by the following equation:

$$\mathrm{E}(\mathrm{wet}_i|X_i) = a_0 + \sum a_i X_i \tag{2}$$

Additionally, the variance of this binary variable is given by

$$\mathrm{Var}(\mathrm{wet}_i) = \pi_i(1 - \pi_i) \tag{3}$$

In a linear regression model, the wet-day probability could be described as follows:

$$\pi_i = \hat{\pi}_\imath + \varepsilon_i \tag{4}$$

in which $\hat{\pi}_\imath$ and ε_i are the deterministic component and the modeling error (or residual), respectively. According to linear regression theory, the residual (ε_i) should satisfy the following three fundamental assumptions: normally distributed with zero mean and unit variance, homoscedastic (i.e. constant variance), and independent from each independent variable. However, as indicated by Equation (3), the variance of precipitation occurrence is dependent on the value of the wet-day probability. It implies that the application of a linear regression model for describing the precipitation occurrence process would violate the required homoscedastic assumption. In addition, this model has a high risk of violating the normality assumption because the error term can be expressed as $\varepsilon_i = O_i - (\alpha_0 + \sum \alpha_i X_i)$. The last issue is that the linear regression model for calculating the wet-day probability has a high chance of providing estimated values that are outside range between 0 and 1 even though the estimated values must be between 0 and 1 by the definition of the probability.

In the end, the logistic curvilinear model is used for the precipitation occurrence process since it is developed for figuring out the probability of a certain class with a binary dependent variable. The theory of logistic regression (Kleinbaum *et al.* 2002) defines the probability of daily precipitation occurrence as follows:

$$\mathrm{Prob}(\text{wet at } i \text{ day}) = \hat{\pi}_i = \frac{e^{a_0 + a_1X_1 + a_2X_2 + \cdots + a_mX_m}}{1 + e^{a_0 + a_1X_1 + a_2X_2 + \cdots + a_mX_m}} \tag{5}$$

or by logit transformation, it can be written as follows:

$$\ln\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = a_0 + a_1X_1 + a_2X_2 + \cdots + a_mX_m \tag{6}$$

where the regression parameters (a_i) can be estimated by the maximum likelihood method. By comparing the estimated probability of precipitation occurrence $\hat{\pi}_i$ and a uniformly distributed random number r_i ($0 \le r_i \le 1$), the proposed SD of daily rainfall process (SDRain) model determines wet/dry processes. For example, if $r_i \le \hat{\pi}_i$ on a given day i, the precipitation occurs on this day i.

Modeling of daily precipitation amounts

As suggested by Kilsby *et al.* (1998), the daily precipitation amount is a non-zero and right-skewed distributed random variable, and hence it can be described by the following model:

$$R_i = \exp(b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m + \eta_i) \tag{7}$$

in which R_i is the daily precipitation, X_i values are the large-scale atmospheric predictors given by GCM simulations, b values are regression parameters, and η_i is the modeling error. Because log-transformation-based models would underestimate the estimated variance of daily precipitation amount, a variance inflate factor ($\mathrm{VIF_R}$) is suggested for increasing the underestimated estimate without changing the mean in order to improve the accuracy of the estimated variance (Hay et al. 1991; Wilby et al. 1994). In the SDRain, an improved $\mathrm{VIF_R}$ is suggested for artificially controlling variances. The modeling error (η_i) in Equation (7) is defined as follows:

$$\eta = Z \times S_e \tag{8}$$

in which S_e is the standard error of each monthly regression model, and Z is a normally distributed random number with a mean of 0 and a standard deviation of $\mathrm{VIF_R}$:

$$Z \sim N(0, \mathrm{VIF_R}) \tag{9}$$

By changing the value of $\mathrm{VIF_R}$, an SDRain user could artificially inflate or diminish the amplitude of white noise.

For synthesizing daily precipitation series for the future, SDRain uses a fraction factor (f) for reflecting a bias between GCMs and the NCEP re-analysis data. The f comes from the deviation of the simulated mean given by GCMs from the estimated mean given by the NCEP re-analysis data. The fraction factor can be defined as the ratio such as follows:

$$f = \frac{\text{total mean generated using NCEP for calibration period}}{\text{total mean generated using GCM for calibration period}} \tag{10}$$

The value of this coefficient is set to 1 in the calibration step of the SDRain model. Consequently, the daily precipitation amount model in SDRain can be expressed as follows:

$$R_i = f \times \exp(b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m + \eta_i) \tag{11}$$

Due to the log-transformation of daily precipitation amounts, the proposed model should take into account the bias correction related to the re-transformation from $\ln(R)$ into R. The expected value of the precipitation amount in the proposed SDRain model can be derived as follows:

$$E(R_i) = f \times C_R \times S_R \times \exp\left(b_0 + \sum_{j=1}^{n} b_jX_{i,j}\right) \tag{12}$$

where S_R is the scaling factor for artificially controlling the variance of R, and C_R is the correction coefficient associated with the total amounts of observed to those of downscaled. The scaling factor can be estimated as follows:

$$\exp\left(\frac{[S_e \times \mathrm{VIF_R}]^2}{2}\right) \tag{13}$$

One of the well-known weaknesses of the regression method is to underestimate the variation of daily precipitation. An SDRain is able to add/reduce the amplitude of white noise by changing the value of $\mathrm{VIF_R}$. The proposed model allows the user to match the variability of the downscaled daily precipitation to of the observed. The value of $\mathrm{VIF_R}$ can be estimated by the ratio of the monthly/annual standard deviations of the downscaled to the observed precipitations. However, an adjustment of the default value (i.e. 12) of $\mathrm{VIF_R}$ results in the variation in average values of daily precipitation amount because of the re-transformation from $\ln(R)$ to R. The value of C_R is estimated by the ratio of monthly or annual average of the precipitation amounts generated with the updated $\mathrm{VIF_R}$ to the observed. Thus, C_R plays a role in constraining the total amount of downscaled precipitation to those of observed. In summary, the S_R and C_R coefficients can be used to improve the accuracy of the SDRain model in representing the variance of the observed daily precipitation amounts in each month without affecting the monthly mean precipitation.

Figures 1 and 2 show the main menu of SDRain and two main functions for modeling daily precipitation amounts and occurrences. The functions are coded in MATLAB 2014a. After compiling them, Visual Basic .NET is used to establish a graphic user interface (GUI) environment so that users can easily execute this tool. The requirement for running this software is to install MATLAB Runtime version 8.3. The detailed instruction to the SDRain software

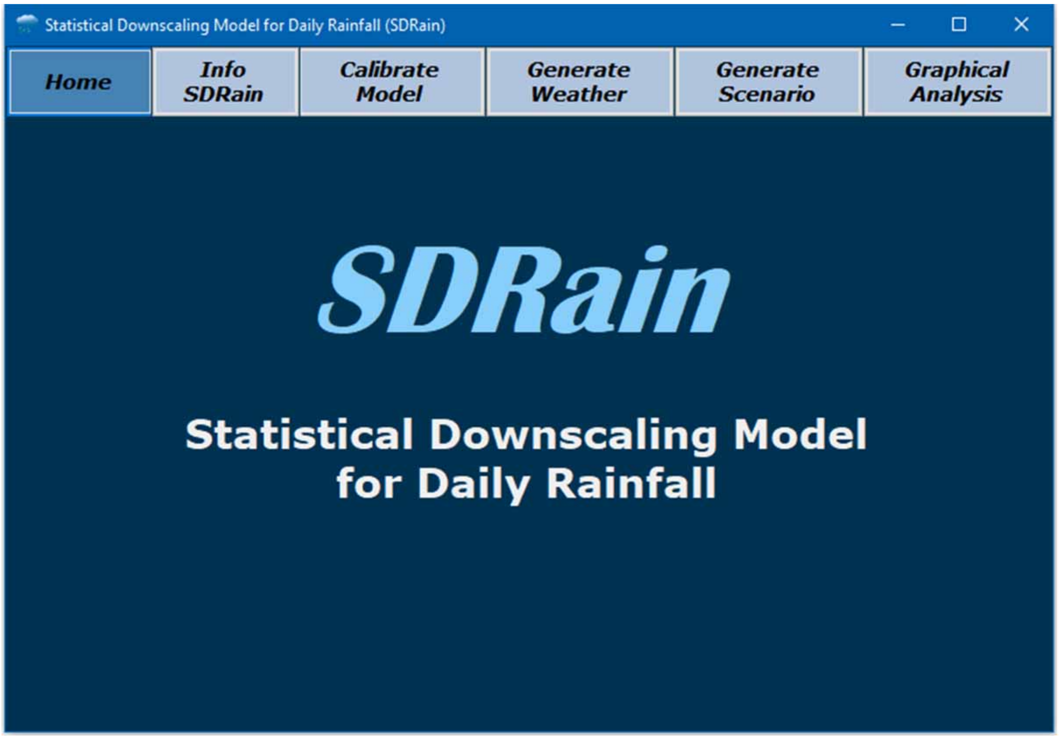

Figure 1 | Main menu of SDRain.

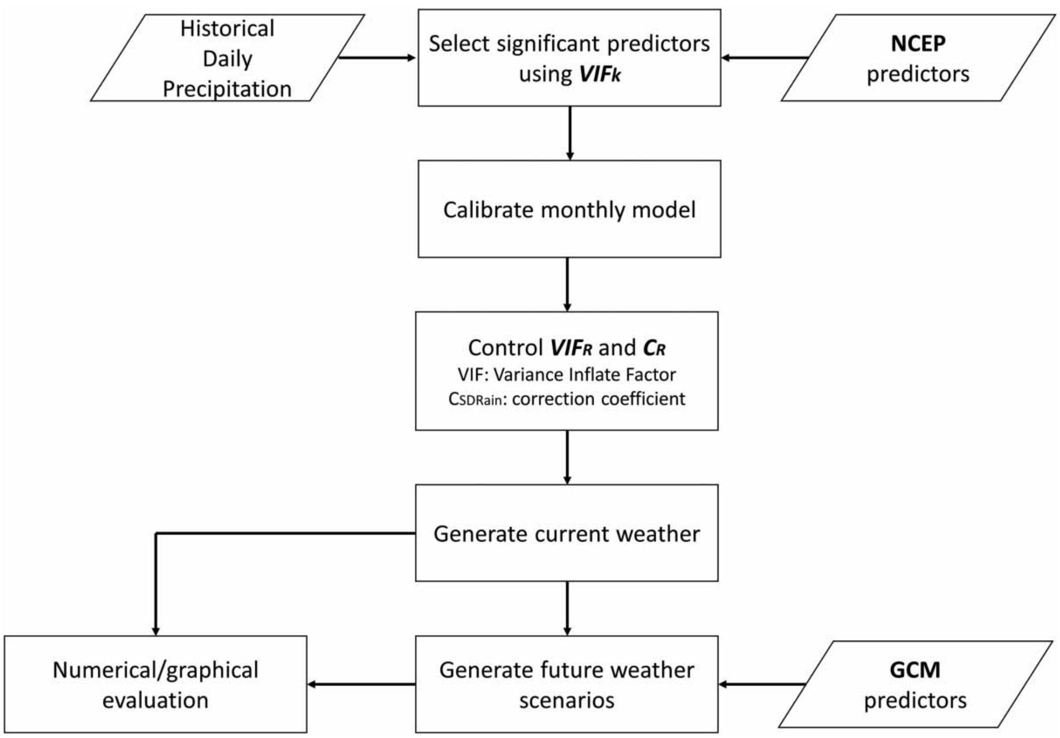

Figure 2 | Scheme of SDRain.

(e.g. installation, executive procedures, and so on) is illustrated in the manual.

Selection of large-scale climate variables in SDRain

In linear regression models, the estimated coefficients and variances could be unreliable if there is the presence of a multicollinearity condition (i.e. there is a significant correlation between independent variables). The atmospheric predictors from NCEP re-analysis data and from GCM simulations are expected to have a strong statistical/physical correlation (e.g. humidity is a function of temperature). Hence, to obtain a reliable estimation of regression model parameters, the software automatically computes the variation inflation factor ($\mathrm{VIF_k}$) for each predictor, X_k, in each monthly precipitation regression model for removing the multicollinearity effects over the 'Screening Variable' step. The $\mathrm{VIF_k}$ can be expressed by the following formula:

$$\mathrm{VIF_k} = \frac{1}{(1 - R_k^2)} \tag{14}$$

where R_k^2 is the coefficient of determination for a regression model for a given predictor X_k as follows:

$$X_k = c_0 + c_1 X_1 + \cdots + c_{k-1} X_{k-1} + c_{k+1} X_{k+1} + \cdots + c_m X_m + \gamma_k \tag{15}$$

ILLUSTRATIVE APPLICATION

To assess the accuracy and feasibility of the proposed tool, case studies are carried out using the NCEP re-analysis data (Kalnay *et al.* 1996) and the observed daily precipitation data available at two networks of raingauges located in two completely different climatic regions: the Seoul station in subtropical-climate Korea and the Dorval station in the cold-climate southern Quebec region in Canada. To evaluate the performance of the proposed downscaling model, the comparison studies are conducted with another regression-based model: the currently popular statistical downscaling model (SDSM). More specifically, historical daily precipitation data sets for the period from 1961 to 2001 (Seoul, South Korea) and from 1961 to 1990 (Dorval, Canada) and the same predictors from the NCEP are used for verifying the models' performances. The evaluation of SDRain for the daily precipitation process is carried out in two steps: the feasibility test with historical data and comparison test with those generated by SDSM. The feasibility test of SDRain is implemented based on the evaluation statistics and indices (as shown in Table 1) to figure out statistical characteristics of the precipitation processes, which are: the average and variance of precipitation, frequency of precipitation occurrence, intensity of precipitation amount, and extreme events.

RESULTS

To compare the performance of representing observed precipitation characteristics by SDSM and SDRain, the same sets of significant large-scale NCEP predictors identified are used for both the models for each given station. The SD model for the Seoul (K1) station consists of the following significant climate predictors: the mean sea level pressure, the vorticity at the near surface, the 500 hPa geopotential height, and the zonal velocity, divergence, and relative humidity at 800 hPa geopotential height. The identified significant predictors for the Dorval (C1) station are the

Table 1 | Evaluation statistics and indices

Categories	Indices	Definition	Unit	Time scale
Basic variable	Precip_m	Average of precipitation	mm/day	Month
	Precip_std	Standard deviation of precipitation	mm/day	Month
Frequency	PRCP1	Percentage of wet days (threshold $\geq$ 1 mm)	%	Season
Intensity	SDII	Mean precipitation amount at wet days	mm/day	Season
Extreme	CDD	Maximum number of CDDs (threshold <1 mm)	Days	Season
	PREC90P	90th percentile of rain day amount	mm	Season

mean sea level pressure, the vorticity, and relative humidity at 850 hPa, the surface zonal velocity and meridional velocity at 850 hPa, the relative humidity at 500 hPa, and the surface specific humidity. Based on the graphical or numerical comparison between observed and generated monthly precipitation means and standard deviations, the values of VIF_R and C_R are determined. In the application studies, the values of VIF_R and C_R used for the Seoul station model are 11 and 1.03, and those for the Dorval station model are 8.5 and 1.05, respectively.

Numerical analysis

The suggested assessment tool calculates and provides McKelvey–Zavoina's measure for presenting the performance of the precipitation occurrence model. Although the percentage of explained variance (R^2) is widely used to measure the extent to which the proportion of variability is accounted for by the regression model, this metric is not suitable for evaluating the performances of the regression model for binary responses or discrete variables. DeMaris (2002) discusses that McKelvey–Zavoina's measure could provide the best estimating explained variances for logistic regression amongst many previous approaches. For either a logit or probit model, McKelvey–Zavoina's measure for measuring the goodness of fit is expressed by the following formula:

$$R^2 = \frac{V(\sum a_k X_k)}{V(\sum a_k X_k) + (\pi^2/3)} \tag{16}$$

where a_k is logistic regression coefficient and $\pi^2/3$ is the underlying error variance for the logistic regression model.

Figure 3 shows the comparison results of the explained variances and McKelvey–Zavoina's measures of the monthly precipitation amount and occurrence models for the two SDs, respectively. The results show that the modeling performance of monthly amount

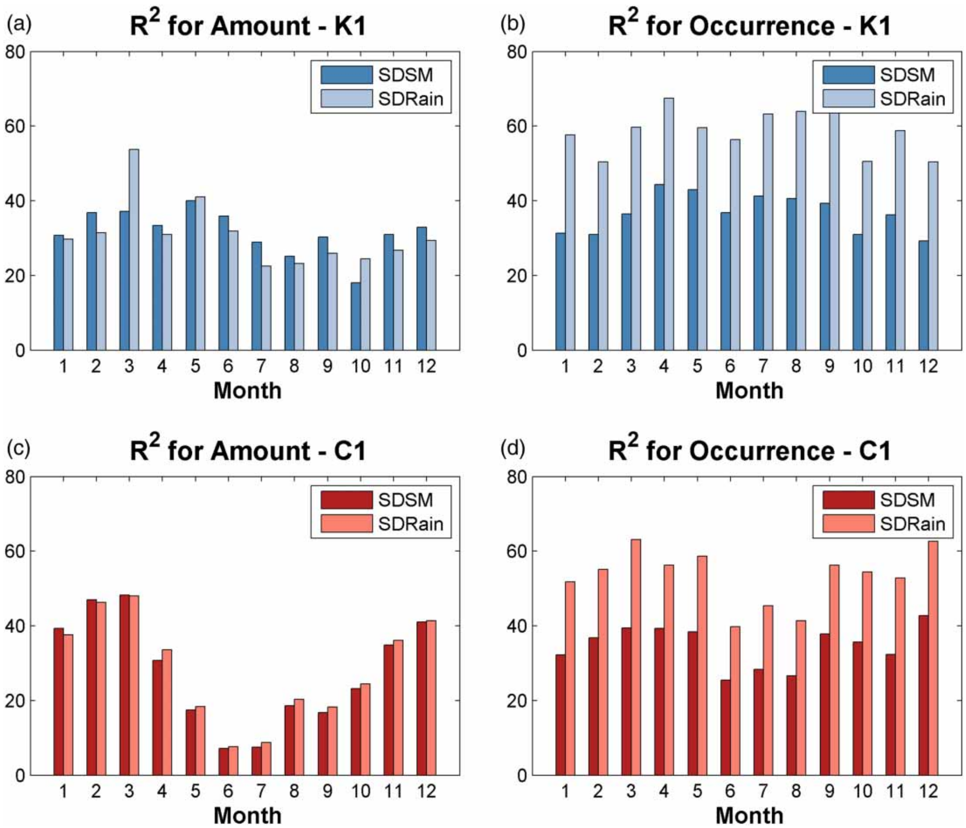

Figure 3 | Comparison of the explained variances (%) of monthly precipitation amount and occurrence models by SDSM and SDRain of two representative stations for each climatic region: Seoul (K1), Korea, for subtropical and Dorval (C1), Quebec, Canada, for a cold climatic region. Explained variances (a) of the precipitation amount model for K1, (b) of amount model for C1, (c) of occurrence model for K1, and (d) of the occurrence model for C1.

models by the suggested SDRain is comparable to those by SDSM. In contrast, it is found that the precipitation occurrence model based on the logistic regression has a better interpretation power regarding the wet–dry process than the multiple linear regression model. Wilby *et al.* (2002) mention that less than 40% is a usual value for precipitation occurrence and amount process, and Hessami *et al.* (2008) address the range of 13–32% of the values of R^2 for the precipitation model. It can be seen that the proposed SDRain model can account for the monthly variance of the wet/dry-day process in the range of 36.91–63.10% and for the monthly variance of the precipitation amount in the range of 7.57–53.76% (shown in Table 2). R^2s of precipitation amount models are relatively low during the rainy season, whereas those of occurrences are not significantly different from each other. It could infer that exploration powers of precipitation amount models are more sensitive regarding seasonal precipitation properties than those of occurrence models.

For an objective assessment, the root-mean-square error (RMSE) is used to compare the performance of the two SD models. Since RMSEs quantify how different the estimated values are from the estimator, these values have been used to assess the quality of the built model (Pandey & Nguyen 1999). The values of the RMSE can be computed by the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum (\mathrm{SI}_{\mathrm{model}} - \mathrm{SI}_{\mathrm{observed}})^2} \tag{17}$$

where SI indicates the value of the evaluation statistics and indices as shown in Table 1 and N is the number of sample size. The smaller RMSE indicates better accuracy of the model considered.

Tables 3–8 provide the results of the monthly and seasonal evaluation SIs for the calibration (Tables 3–5) and validation (Tables 6–8) periods, respectively. In these tables, a bold letter denotes the case when values of RMSEs provided by SDRain are higher than those by SDSM. The proposed SD model represents more accurately monthly means of precipitation for two stations than the SDSM as shown in Table 3. Only for the month of July at K1, it is found that the RMSE value for the mean precipitation by the SDRain is higher than the value given by the SDSM. For the other three-monthly precipitation models (February, April, and November at C1), the difference of the RMSE is less than 10%. As shown in Table 4, in general, the proposed SDRain at C1 provided more accurate results for the variance of precipitation, while the results at K1 are comparable with those given by the SDSM. Regarding the accurate simulations of representing monthly means and standard deviations, it could be a reason that the proposed $\mathrm{VIF_R}$ term plays a role

Table 3 | RMSEs of monthly means of precipitation over the calibration period of Seoul (K1), Korea, and of Dorval (C1), Quebec, Canada, respectively

	Seoul		Dorval	
Month	SDSM	SDRain	SDSM	SDRain
Jan	0.165	0.081	0.242	0.214
Feb	0.171	0.149	0.233	**0.246**
Mar	0.345	0.179	0.254	0.249
Apr	0.317	0.307	0.224	**0.233**
May	0.307	0.287	0.184	0.197
Jun	0.539	0.477	0.301	0.260
Jul	0.750	**0.932**	0.340	0.257
Aug	1.141	0.899	0.363	0.320
Sep	1.025	0.651	0.342	0.283
Oct	0.309	0.202	0.273	0.214
Nov	0.157	0.143	0.255	**0.264**
Dec	0.141	0.090	0.283	0.250

The bold values denote the case when the RMSE value of SDRain is higher than that of SDSM.

Table 2 | Explained variance (%) of precipitation amount and occurrence models by SDRain for Seoul (South Korea) and Dorval (Quebec, Canada)

		Month											
Site	Model	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Seoul	Amount	29.71	31.39	53.76	31	41.05	31.83	22.53	23.2	25.94	24.45	26.72	29.34
	Occurrence	49.58	43.16	48.46	54.92	54.99	50.68	55.67	55.67	58.13	39.48	45.47	36.91
Dorval	Amount	37.61	46.35	47.96	33.56	18.32	7.57	8.76	20.3	18.21	24.46	36.07	41.39
	Occurrence	51.8	55.07	63.1	56.21	58.65	39.79	45.34	41.34	56.25	54.42	52.8	62.66

Table 4 | RMSEs of monthly standard deviations of precipitation over the calibration period of Seoul (K1), Korea, and of Dorval (C1), Quebec, Canada, respectively

	Seoul		Dorval	
Month	**SDSM**	**SDRain**	**SDSM**	**SDRain**
Jan	0.578	**0.641**	0.619	**1.467**
Feb	0.864	**1.601**	1.193	**1.704**
Mar	1.371	1.256	1.190	**1.210**
Apr	2.373	2.076	0.780	0.449
May	1.460	**2.035**	0.751	0.461
Jun	2.680	2.104	1.214	0.811
Jul	4.185	**6.005**	1.441	1.041
Aug	8.055	3.897	1.682	0.944
Sep	6.423	**6.567**	1.477	1.358
Oct	1.679	**1.743**	1.091	0.381
Nov	1.111	0.692	1.219	0.592
Dec	0.401	**0.823**	0.925	0.805

The bold values denote the case when the RMSE value of SDRain is higher than that of SDSM.

Table 6 | RMSEs of monthly means of precipitation over the validation period of Seoul, Korea, and of Dorval, Quebec, Canada, respectively

	Seoul		Dorval	
Month	**SDSM**	**SDRain**	**SDSM**	**SDRain**
Jan	0.257	0.145	0.257	0.214
Feb	0.433	0.373	0.365	0.196
Mar	0.323	**0.404**	0.229	**0.261**
Apr	1.142	1.059	0.238	0.208
May	0.644	**0.820**	0.375	0.367
Jun	0.654	**0.706**	0.332	0.265
Jul	2.922	2.168	0.298	0.269
Aug	4.345	2.496	0.415	0.331
Sep	3.183	2.211	0.435	**0.592**
Oct	0.621	0.499	0.318	0.267
Nov	0.257	**1.096**	0.669	0.454
Dec	0.200	**0.287**	0.234	0.197

The bold values denote the case when the RMSE value of SDRain is higher than that of SDSM.

of control both monthly average and variation concurrently. In addition to means and variations, Table 5 provides the RMSE values for the frequency, intensity, and extreme values of precipitation. It is found that SDRain on the basis of the logistic regression is able to accurately account for the precipitation occurrence process. For the SDII index, the SDRain can provide a significant improvement over the SDSM (Tables 5 and 8) because of the substantially improved accuracy for the number of wet days. The maximum number of consecutive

Table 5 | RMSEs of seasonal indices about the frequency, intensity, and extreme of precipitation over the calibration period for Seoul (K1) and Dorval (C1), respectively

		Seoul		Dorval	
	Season	**SDSM**	**SDRain**	**SDSM**	**SDRain**
Prcp1 (%)	Spring	0.963	0.447	1.126	1.018
	Summer	1.342	0.665	1.201	1.137
	Autumn	1.283	0.612	1.592	1.096
	Winter	0.570	0.544	3.312	1.468
SDII (mm/wet day)	Spring	1.753	0.820	0.360	**0.382**
	Summer	1.395	1.218	0.673	0.473
	Autumn	2.787	1.231	0.723	0.388
	Winter	1.039	0.413	0.415	0.377
CDD (days)	Spring	8.749	7.747	3.652	3.476
	Summer	6.079	**6.438**	6.220	5.824
	Autumn	7.591	**7.944**	3.988	**5.027**
	Winter	13.606	**11.338**	7.653	8.518
Prec90p (mm/day)	Spring	1.214	0.447	0.412	**0.475**
	Summer	7.505	3.852	0.678	0.582
	Autumn	0.910	0.353	0.522	0.476
	Winter	0.326	0.325	0.506	0.509

The bold values denote the case when the RMSE value of SDRain is higher than that of SDSM.

Table 7 | RMSEs of monthly standard deviations of precipitation over the validation period of Seoul, Korea, and of Dorval, Quebec, Canada, respectively

	Seoul		Dorval	
Month	**SDSM**	**SDRain**	**SDSM**	**SDRain**
Jan	1.024	0.757	0.653	**1.598**
Feb	1.762	1.338	1.060	**1.435**
Mar	2.029	**3.013**	2.080	**2.171**
Apr	4.179	3.330	0.831	0.571
May	2.533	2.070	0.780	**0.912**
Jun	4.141	3.516	1.527	0.839
Jul	10.091	8.812	1.436	1.034
Aug	21.069	9.203	2.220	0.976
Sep	8.259	**10.130**	1.197	**1.479**
Oct	2.816	2.301	0.918	**1.303**
Nov	0.868	**2.351**	0.948	0.787
Dec	0.747	**1.318**	0.970	0.808

The bold values denote the case when the RMSE value of SDRain is higher than that of SDSM.

Table 8 | RMSE of seasonal indices about the frequency, intensity, and extreme of precipitation over validation period for Seoul and Dorval, respectively

		Seoul		Dorval	
	Season	**SDSM**	**SDRain**	**SDSM**	**SDRain**
Prcp1 (%)	Spring	1.651	1.160	6.493	5.642
	Summer	4.190	1.952	2.769	1.445
	Autumn	2.279	1.820	3.088	**4.464**
	Winter	1.544	0.983	1.651	1.486
SDII (mm/wet day)	Spring	2.261	1.321	1.205	1.046
	Summer	3.155	2.589	1.299	0.589
	Autumn	4.128	3.007	0.452	**0.519**
	Winter	0.820	0.552	0.395	0.381
CDD (days)	Spring	14.024	**16.078**	4.737	3.945
	Summer	7.747	**8.263**	5.051	**5.479**
	Autumn	10.914	**11.988**	4.312	2.967
	Winter	15.250	14.008	4.169	4.177
Prec90p (mm/day)	Spring	1.460	**1.809**	0.719	0.527
	Summer	2.621	**2.795**	0.656	0.506
	Autumn	1.455	1.132	0.855	0.484
	Winter	0.355	0.323	0.461	**0.809**

The bold values denote the case when the RMSE value of SDRain is higher than that of SDSM.

dry days (CDDs) has been regarded as one of the most difficult indices in the modeling process. The performance of representing daily precipitation by SDRain is found to be comparable to that of the SDSM. With respect to the extreme precipitation indices (Prec90p), the RMSE values from SDRain are generally lower than those from SDSM. In brief, the proposed SD model is able to capture well seasonal statistics of the extreme precipitation, as well as its frequency and intensity for both calibration and validation periods for two raingauge stations.

Graphical analysis

The SDRain software provides the boxplots for assessing graphically the accuracy (the closeness between the estimated median value of the model and the observation) and the robustness of the model results (the size of the inter-quartile range box).

Figures 4 and 5 show the boxplots for the monthly mean of precipitation (Precip-Mean) and for the percentage of wet days (Prcp1), respectively. It can be seen that the proposed model is able to reproduce more accurate statistics than SDSM does for both the stations. Moreover, the accuracy of the results for the percentage of wet days index (Prcp1) by

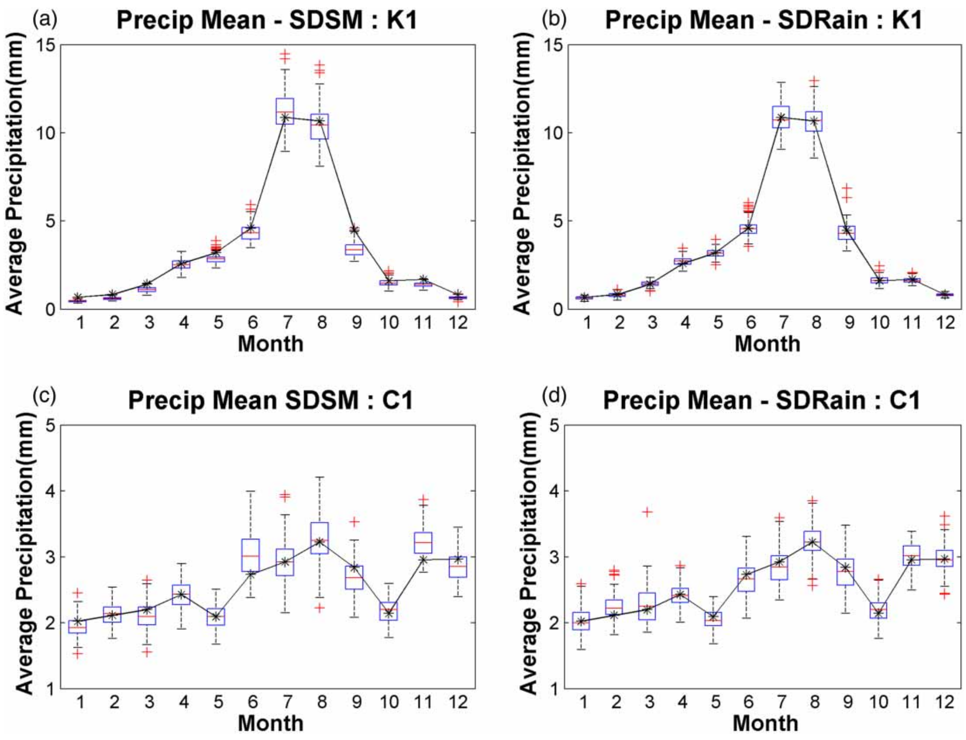

Figure 4 | Boxplot of monthly means of precipitation for SDSM and SDRain for Seoul, Korea, and Dorval, Quebec, Canada. (a) SDSM for Seoul, (b) SDRain for Seoul, (c) SDSM for Dorval, and (d) SDRain for Dorval. (Black star markers indicate monthly average values of observed precipitation data and boxplots indicate model results.)

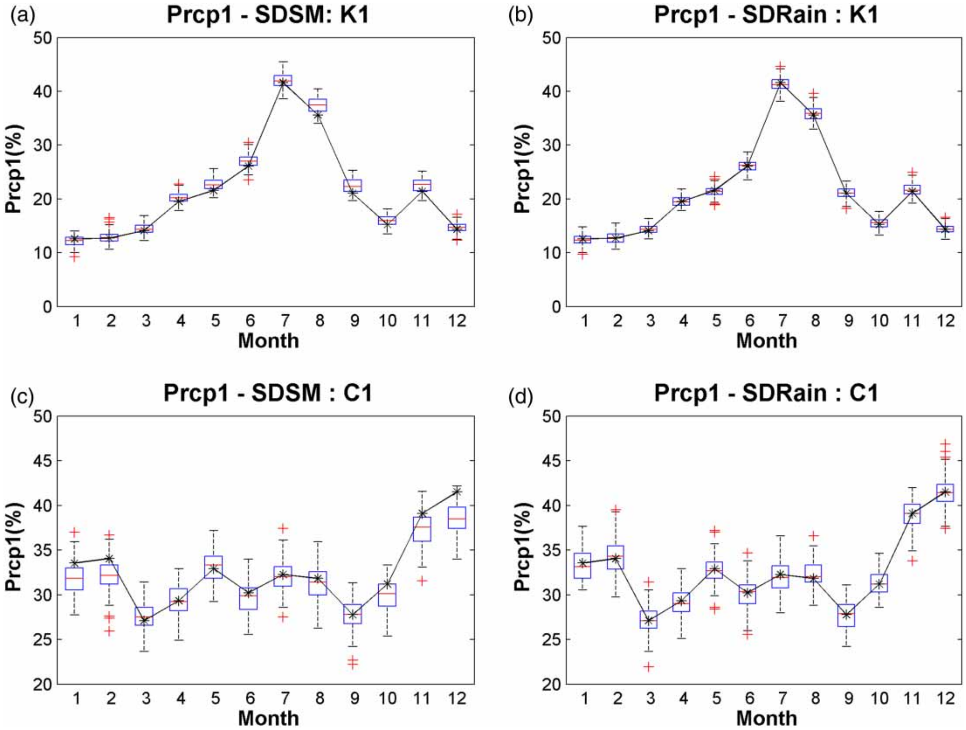

Figure 5 | Boxplots of the monthly percentage of wet day for SDSM and SDRain for Seoul, Korea, and Dorval, Quebec, Canada. (a) SDSM for Seoul, (b) SDRain for Seoul, (c) SDSM for Dorval, and (d) SDRain for Dorval. (Black star markers indicate monthly average values of observed precipitation data and boxplots indicate model results.)

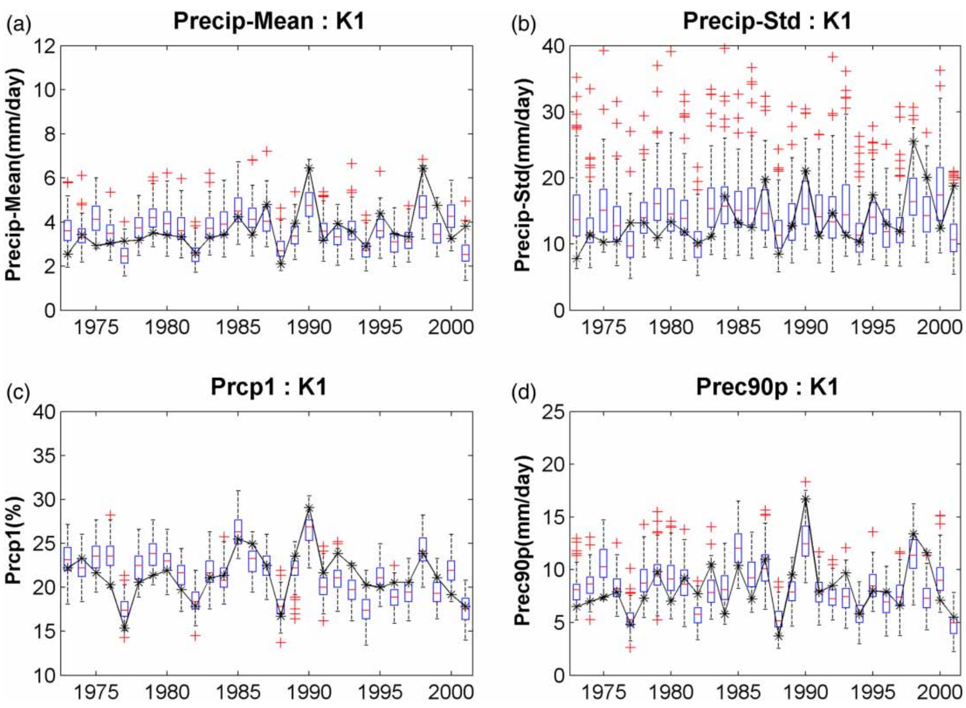

Figure 6 | Boxplots of annual statistics and indices for SDRain for Seoul, Korea. For each station, Precip-mean, Precip-Std, Prcp1, and Prec90p are presented. (Black star markers indicate monthly average values of observed precipitation data and boxplots indicate model results.)

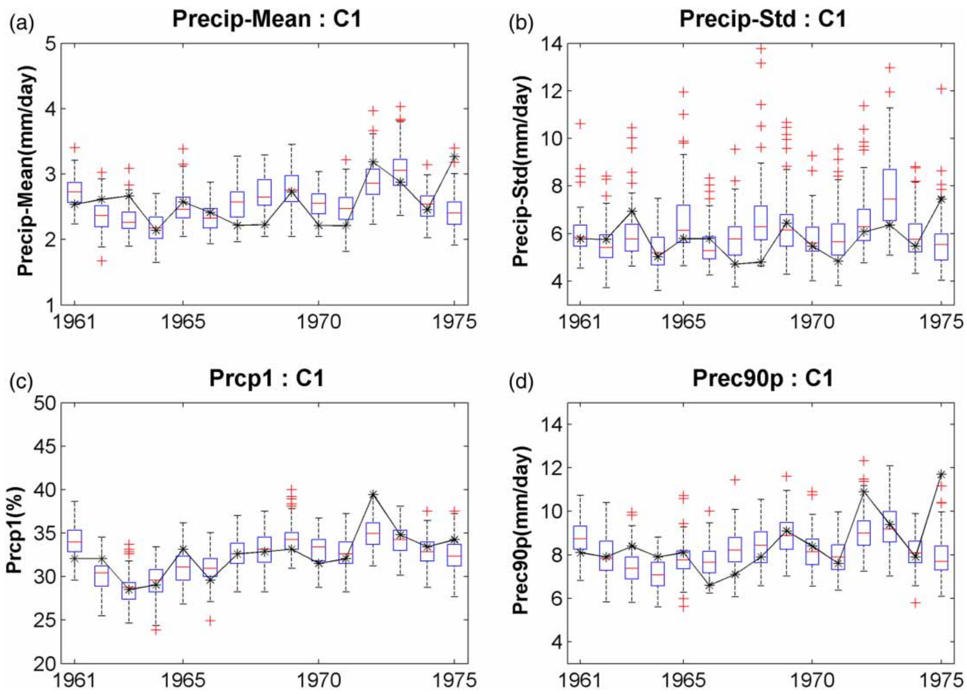

Figure 7 | Boxplots of annual statistics and indices for SDRain for Dorval, Quebec, Canada. For each station, Precip-mean, Precip-Std, Prcp1, and Prec90p are presented. (Black star markers indicate monthly average values of observed precipitation data and boxplots indicate model results.)

the SDRain indicates that the use of the logistic regression approach is more suitable than the linear regression used in the SDSM for modeling of the precipitation occurrence process. To assess the performance of the SDRain model in representing the annual variability of precipitation, the annual indices of Precip-Mean, Precip-Std, Prcp1, and Prec90p are computed in Figures 6 and 7. It could be concluded that the proposed assessment tool can describe quite well the annual characteristics of the daily precipitation series.

SUMMARY AND CONCLUSION

In the present study, an improved statistical downscaling model (SDRain) has been developed to accurately simulate precipitation processes at a single site for the current and climate change conditions. More specifically, the proposed tool is based on the combination of two main components: (i) a logistic regression model for representing the precipitation occurrence process and (ii) a nonlinear regression model for the precipitation amount. As a GUI environment-software, SDRain can be used for generating daily precipitation series easily. Results of the illustrative applications using data from two raingauge stations located in two different climatic regions in Korea and in Canada have demonstrated the feasibility and accuracy of the proposed assessment tool. Furthermore, it has been demonstrated that the suggested SDRain model could provide more accurate results than those given the existing SDSM model on the basis of the numerical and graphical comparisons of the model results with the observed data. In common with regression models, the calibrated model is also highly sensitive to the choice of predictors. In subsequent works, we will further refine our SD model for providing an objective method to select the best set of significant predictors. In addition, the next version of SDRain will include automatic procedures to determine the values of VIF_R and C_R.

ACKNOWLEDGEMENTS

The development of SDRain Ver. 1.1 was supported by FloodNet-NSERC (National Sciences and Engineering Research Council of Canada) from November 2014 to January

2017. The detailed information about FloodNet-NSERC is available at the website: http://www.nsercfloodnet.ca.

SUPPLEMENTARY DATA

The Supplementary Data for this paper is available online at http://dx.doi.org/10.2166/wcc.2019.403.

REFERENCES

Bárdossy, A. 1997 Downscaling from GCMs to local climate through stochastic linkages. *Journal of Environmental Management* **49** (1), 7–17.

Cannon, A. J. 2008 Probabilistic multisite precipitation downscaling by an expanded Bernoulli–gamma density network. *Journal of Hydrometeorology* **9** (6), 1284–1300.

DeMaris, A. 2002 Explained variance in logistic regression: a Monte Carlo study of proposed measures. *Sociological Methods & Research* **31** (1), 27–74.

Gabriel, K. & Neumann, J. 1962 A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society* **88** (375), 90–95.

Goodess, C. M. & Palutikof, J. P. 1998 Development of daily rainfall scenarios for southeast Spain using a circulation-type approach to downscaling. *International Journal of Climatology: A Journal of the Royal Meteorological Society* **18** (10), 1051–1083.

Harpham, C. & Wilby, R. L. 2005 Multi-site downscaling of heavy daily precipitation occurrence and amounts. *Journal of Hydrology* **312** (1–4), 235–255.

Hay, L. E., McCabe Jr., G. J., Wolock, D. M. & Ayers, M. A. 1991 Simulation of precipitation by weather type analysis. *Water Resources Research* **27** (4), 493–501.

Hessami, M., Gachon, P., Ouarda, T. B. & St-Hilaire, A. 2008 Automated regression-based statistical downscaling tool. *Environmental Modelling & Software* **23** (6), 813–834.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L. M., Iredell, S., Saha, G., White, J., Woollen, Y., Zhu, M., Chelliah, W., Ebisuzaki, W., Higgins, J., Janowiak, K. C., Mo, C., Ropelewski, J., Wang, A., Leetmaa, R., Reynolds, R., Jenne, R. & Joseph, D. 1996 The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77** (3), 437–472.

Kilsby, C., Cowpertwait, P., O'connell, P. & Jones, P. 1998 Predicting rainfall statistics in England and Wales using atmospheric circulation variables. *International Journal of Climatology: A Journal of the Royal Meteorological Society* **18** (5), 523–539.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M. & Klein, M. 2002 *Logistic Regression*. Springer, Berlin.

Laprise, R. 2008 Regional climate modelling. *Journal of Computational Physics* **227** (7), 3641–3666.

Mehrotra, R. & Sharma, A. 2010 Development and application of a multisite rainfall stochastic downscaling framework for climate change impact assessment. *Water Resources Research* **46** (7).

Nguyen, V. & Nguyen, T. 2008 Statistical downscaling of daily precipitation process for climate-related impact studies. In: *Hydrology and Hydraulics* (V. P. Singh, ed.). Chapter 16 Water Resources Publications, Highlands Ranch, Colorado, pp. 587–604.

Nguyen, V.-T.-V., Nguyen, T.-D. & Gachon, P. 2006. On the linkage of large-scale climate variability with local characteristics of daily precipitation and temperature extremes: an evaluation of statistical downscaling methods. In: *Advances in Geosciences: Volume 4: Hydrological Science (HS)* (N. Park, ed.), World Scientific, Singapore, pp. 1–9.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q. & Dasgupta, P. 2014 Climate change 2014: synthesis report. In: *Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (R. Pachauri & L. Meyer, eds.). IPCC, Geneva, Switzerland, p. 151.

Pandey, G. R. & Nguyen, V.-T.-V. 1999 A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology* **225** (1–2), 92–101.

Richardson, C. W. & Wright, D. A. 1984 WGEN: a model for generating daily weather variables. U.S. Department of Agriculture, Agricultural Research Service, ARS-8, p. 83.

Serinaldi, F. 2009 A multisite daily rainfall generator driven by bivariate copula-based mixed distributions. *Journal of Geophysical Research: Atmospheres* **114** (D10).

Sharma, A. & Lall, U. 1999 A nonparametric approach for daily rainfall simulation. *Mathematics and Computers in Simulation* **48** (4–6), 361–371.

Wilby, R. L. & Wigley, T. 1997 Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography* **21** (4), 530–548.

Wilby, R., Greenfield, B. & Glenny, C. 1994 A coupled synoptic-hydrological model for climate change impact assessment. *Journal of Hydrology* **153** (1–4), 265–290.

Wilby, R. L., Dawson, C. W. & Barrow, E. M. 2002 SDSM – a decision support tool for the assessment of regional climate change impacts. *Environmental Modelling & Software* **17** (2), 145–157.

Wilks, D. 1998 Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology* **210** (1–4), 178–191.

Wilks, D. S. & Wilby, R. L. 1999 The weather generation game: a review of stochastic weather models. *Progress in Physical Geography* **23** (3), 329–357.

Xu, C.-y. 1999 From GCMs to river flow: a review of downscaling methods and hydrologic modelling approaches. *Progress in Physical Geography* **23** (2), 229–249.

Yarnal, B., Comrie, A. C., Frakes, B. & Brown, D. P. 2001 Developments and prospects in synoptic climatology. *International Journal of Climatology: A Journal of the Royal Meteorological Society* **21** (15), 1923–1950.

Yeo, M.-H. 2014 *Statistical Modeling of Precipitation Processes for Gaged and Ungaged Sites in the Context of Climate Change*. Thesis, McGill University Libraries.

First received 23 September 2019; accepted in revised form 12 November 2019. Available online 29 November 2019