Journal of Hydrology 385 (2010) 13–22

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Statistical downscaling of daily precipitation using support vector machines and multivariate analysis

Shien-Tsung Chen, Pao-Shan Yu *, Yi-Hsuan Tang

Department of Hydraulic and Ocean Engineering, National Cheng Kung University, Tainan 701, Taiwan

ARTICLE INFO

Article history:
Received 25 February 2009
Received in revised form 15 January 2010
Accepted 25 January 2010

This manuscript was handled by
K. Georgakakos, Editor-in-Chief, with the assistance of A.A. Tsonis, Associate Editor

Keywords:
Statistical downscaling
Daily precipitation
Support vector machine
SDSM

SUMMARY

Downscaling local daily precipitation from large-scale weather variables is often necessary when studying how climate change impacts hydrology. This study proposes a two-step statistical downscaling method for projection of daily precipitation. The first step is classification to determine whether the day is dry or wet, and the second is regression to estimate the amount of precipitation conditional on the occurrence of a wet day. Predictors of classification and regression models are selected from large-scale weather variables in NECP reanalysis data based on statistical tests. The proposed statistical downscaling method is developed according to two methodologies. One methodology is support vector machine (SVM), including support vector classification (SVC) and support vector regression (SVR), and the other is multivariate analysis, including discriminant analysis (for classification) and multiple regression. The popular statistical downscaling model (SDSM) is analyzed for comparison. A comparison of downscaling results in the Shih-Men Reservoir basin in Taiwan reveals that overall, the SVM reproduces most reasonable daily precipitation properties, although the SDMS performs better than other models in small daily precipitation (less than about 10 mm). Finally, projection of local daily precipitation is performed, and future work to advance the downscaling method is proposed.

Introduction

Climate change and its impacts on water resources have gained significant attention in hydrology. General circulation models (GCMs) provide reasonable simulations of weather variables at global scales for climate change study, but cannot generate local climate details for impact studies of water resources at a drainage basin. Accordingly, downscaling methods have been developed to link GCM outputs at coarse resolutions with surface weather variables at finer resolutions. Downscaling methods are commonly classified as statistical downscaling and dynamic downscaling. Statistical downscaling methods construct a statistical relationship between large-scale GCM outputs and local weather variables, and dynamic downscaling methods employ high-resolution regional climate models nested in a GCM to obtain local weather variables. Although both statistical and dynamic downscaling methods have their own advantages, statistical downscaling is more widely adopted in hydrological studies because it is less computationally demanding.

Statistical downscaling can be roughly grouped into four categories, weather typing method (e.g., Bárdossy and Plate, 1992; von Storch et al., 1993; Bárdossy, 1997), stochastic weather generators (e.g., Selker and Haith, 1990; Tung and Haith, 1995; Yu et al., 2002), resampling methods (e.g., Murphy, 2000; Buishand and Brandsma, 2001; Palutikof et al., 2002) and regression methods. A regression method constructs a linear or non-linear empirical function between local-scale variables and large-scale GCM variables, and is preferred among statistical downscaling methods because it is easy to implement. The downscaling regression function is derived by artificial neural networks (Hewitson and Crane, 1996; Olsson et al., 2001; Dibike and Coulibaly, 2006), canonical correlation analysis (Kaas et al., 1996; Landman et al., 2001), principal components (Burger, 1996; Menzel and Burger, 2002; Chu et al., 2008) or support vector machines (Tripathi et al., 2006; Anandhi et al., 2008).

Various downscaling models and software have been developed. One popularly used model is the statistical downscaling model (SDSM) (Wilby et al., 2002). For example, Wilby et al. (2006) combined SDSM with a conceptual water balance model and a mass-balance water quality model to investigate climate change impact assessment and uncertainty in river flow and water quality. Additionally, SDSM is often compared with statistical downscaling models. Harpham and Wilby (2005) compared SDSM with two artificial neural networks (ANNs), and concluded that SDSM yields better daily precipitation quantiles and inter-site correlation than the ANNs. Khan et al. (2006) compared SDSM, a weather generator model and an ANN model to assess the

* Corresponding author.
E-mail address: yups@ncku.edu.tw (P.-S. Yu).

0022-1694/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2010.01.021

uncertainty of downscaling results, and concluded that SDSM reproduces the best statistical characteristics of observed data.

Research of climate change impacts on water resources often requires downscaling daily precipitation. This study proposes a two-step statistical downscaling method for daily precipitation projection. The first step is to classify the day as dry or wet, and the second is regression to calculate daily precipitation amount if the day is classified as wet. The proposed statistical downscaling method is based on two methodologies, namely support vector machines (SVMs), including support vector classification (SVC) and support vector regression (SVR), and multivariate analysis, including discriminant analysis (for classification) and multiple regression. These methodologies are described in detail in Section "Methodologies". Section "Study Area and Data" describes the local daily precipitation data and the large-scale predictors from reanalysis and GCM data. Section "Results and Discussion" presents downscaling model development and results by SVM and multivariate model. Comparison is also made with the SDSM downscaling results. Finally, projected daily precipitations in the study area are presented.

Study area and data

Shih-Men Reservoir basin

The Shih-Men Reservoir, located in the Danshuei River basin in northern Taiwan, was completed in 1964 as a multipurpose reservoir for agriculture, water supply, flood control and hydropower generation. The Shih-Men Reservoir has a storage capacity of about 3×10^8 m^3, and is a major reservoir in Taiwan. Shih-Men Reservoir basin (Fig. 1) encloses an area of 763 km^2, and its elevation ranges from 209 m to 2609 m above sea level. The mean annual precipitation on the basin is about 2250 mm. The long-term daily precipitations since 1964 are available from 10 raingauges. The ratio of precipitation during the wet period (May–October) and dry period is about 7:3 (Fig. 1). The daily areal mean precipitations on the Shih-Men Reservoir basin were calculated using the Thiessen polygons method. The areal precipitation series was divided into the calibration period (1964–1990), to develop statistical downscaling models, and the validation period (1991–2000), to examine and compare downscaling results. Fig. 1 also presents average monthly precipitation days (daily precipitation > 0 mm) according to the areal precipitation series.

Season separation

Taiwan is in the northeast Pacific off the Asian Continent. Its climate is marine tropical, and is governed by the East Asian Monsoon, which is divided into a summer monsoon and a winter monsoon. The summer monsoon is a warm and wet southwest monsoon, and the winter monsoon is a cold and dry northeast monsoon. Therefore, separating a year into wet and dry seasons helps the precipitation downscaling. Typically, the wet season in Taiwan is from May to October, and the dry season is from November to April. This customary season separation is delineated by the Water Resources Bureau in Taiwan and is directly used in this work. Consequently, precipitation downscaling models were modeled for the dry and wet seasons separately.

GCM and predictors

The data set in this work includes both daily areal mean precipitation on the reservoir basin (i.e., local precipitation) and large-scale weather factors. The local precipitation is the predictand in the downscaling model. The large-scale weather factors are candidate predictors from both the US National Centers for Environmental Prediction (NCEP) reanalysis data and the UK Hadley Centre Coupled Model, Version 3 (HadCM3). The reanalysis data were used to develop the downscaling model, and the GCM (HadCM3) A2 and B2 scenario outputs were used to project future precipitation. The reanalysis data were interpolated to fit the grid size of the HadCM3 output to ensure that the downscaling model was consistent when used in projection. Table 1 lists 26 large-scale weather factors of the reanalysis and GCM data, from which the data on the grid closest to the reservoir basin were used. The values of factors were normalized by their respective means and standard deviations.

Methodologies

Proposed downscaling method

The proposed daily precipitation downscaling method consists of both classification and regression models, which are described in following Steps A and B. The classification model (Step A) classifies days as dry or wet, since precipitation occurs only on wet days. Therefore, the dry/wet day classification is the first step in the proposed downscaling method. The precipitation amount is assumed to be zero on a dry day. If the day is classified as wet, then the amount of precipitation is estimated by a suitable regression model (Step B). Fig. 2 presents the flow chart of the proposed downscaling method.

Step A: classification of dry/wet day

A-1: selecting the predictors for classification model. To select suitable predictors for developing the classification model, weather factors from NCEP reanalysis data during the calibration period are split into both dry-day group and wet-day group based on

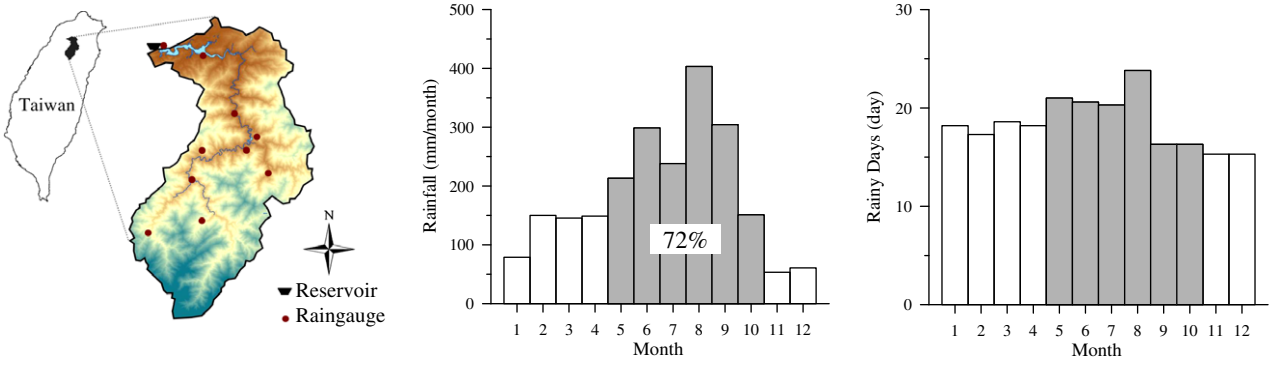

Fig. 1. Shih-Men Reservoir basin and monthly precipitation.

Table 1
Large-scale weather factor (from NCEP).

Variable	Description
Mslp	Mean sea level pressure
P5_f	Geostrophic airflow velocity at 500 hPa height
P5_u	Zonal velocity component at 500 hPa height
P5_v	Meridional velocity component at 500 hPa height
P5_z	Vorticity at 500 hPa height
P5th	Wind direction at 500 hPa height
P5zh	Divergence at 500 hPa height
P8_f	Geostrophic airflow velocity at 850 hPa height
P8_u	Zonal velocity component at 850 hPa height
P8_v	Meridional velocity component at 850 hPa height
P8_z	Vorticity at 850 hPa height
P8th	Wind direction at 850 hPa height
P8zh	Divergence at 850 hPa height
P500	500 hPa geopotential height
P850	850 hPa geopotential height
P_f	Near surface geostrophic airflow velocity
P_u	Near surface zonal velocity component
P_v	Near surface meridional velocity component
P_z	Near surface vorticity
P_th	Near surface wind direction
P_zh	Near surface divergence
R500	Relative humidity at 500 hPa height
R850	Relative humidity at 850 hPa height
Rhum	Near surface relative humidity
Shum	Near surface specific humidity
Temp	Mean temperature at 2 m

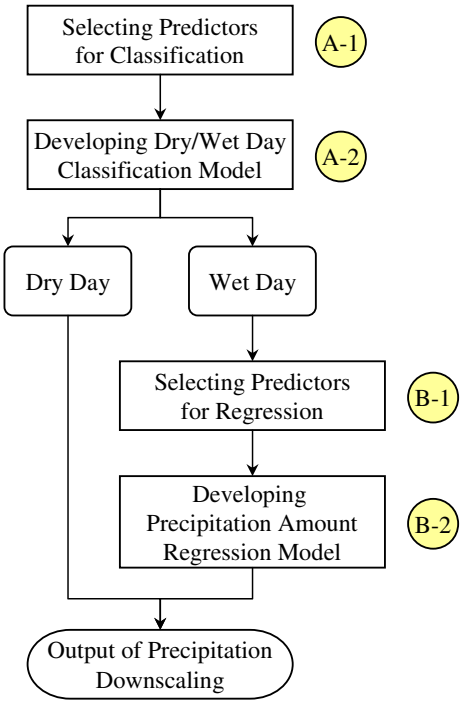

Fig. 2. Flowchart of the proposed downscaling method.

the local daily precipitation. These two groups are statistically tested for each weather factor in Table 1. If the statistical property of a weather factor in the dry-day group is very distinct from that in the wet-day group, then this factor is potentially a suitable predictor for dry/wet day classification. Accordingly, predictors for a classification model can be chosen in many ways. This work uses the Two-sample Kolmogorov–Smirnov test to examine whether the data of a large-scale weather factor between the dry and wet groups are statistically different.

A-2: developing the classification model. A classification model is then developed after the predictor selection Step A-1. The proposed classification model is based on support vector classification (in Subsection "Support vector classification") and discriminant analysis (in Subsection "Discriminant analysis"). If a day is classified as dry, then the local daily precipitation is assumed to be zero. If a day is classified as wet, then the precipitation amount is estimated using a regression model as in Step B.

Step B: regression for precipitation amount on wet days

B-1: selecting the predictors for regression model. The regression model for precipitation estimation is based on the weather factors in Table 1 that are strongly correlated with local precipitation on wet days, since such factors are probably appropriate predictors of the amount of precipitation. This study utilizes Spearman's rank correlation coefficient, rather than the Pearson's correlation coefficient, to choose predictors, because the Spearman's rank correlation coefficient does not require any assumptions about the frequency distribution of the factor, and it allows a significance test to be run.

B-2: developing the regression model. The regression model is based on both the predictors selected in Step B-1 and the amounts of local precipitation. This study adopts both support vector regression (in Subsection "Support vector regression") and multiple regression (in Subsection "Multiple regression"). The established regression model is used to calculate the amount of local precipitation on a wet day, conditional on the occurrence of a wet day, as classified by the classification model. The following sections summarize the support vector machines and multivariate analysis.

Support vector machines

A support vector machine (SVM) that reduces the problem of overfitting by adopting the theory of structural risk minimization, has recently gained popularity in many disciplines. The SVM is mainly utilized in classification and regression problems. Detail principles and algorithms can be found in Vapnik (1995, 1998).

Support vector classification

Consider training data from two classes $[(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)]$ with input vector $\mathbf{x}_i$ and output data y_i with class labels $y_i \in \{+1, -1\}$. The classifier for the problem of binary classification is

$$f(\mathbf{x}) = \text{sign}[\mathbf{w}^T \cdot \Phi(\mathbf{x}) + b] \tag{1}$$

where the input vector $\mathbf{x}$ is mapped into a feature space by a nonlinear function $\Phi(\mathbf{x})$, and $\mathbf{w}$ and b are the classifier parameters. Determining the classifier from the SVM theory is equivalent to solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T \cdot \mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i[\mathbf{w}^T \cdot \Phi(\mathbf{x}_i) + b] \geqslant 1 - \xi_i \\ & \xi_i \geqslant 0, \quad i = 1, \ldots, l \end{aligned} \tag{2}$$

where ξ_i is a non-negative slack variable that exerts an influence on the objective function, when a data point is misclassified, and C is the penalty parameter with a positive value. This optimization problem is solved by introducing Lagrange multipliers α_i, where $0 \leqslant \alpha_i \leqslant C$, and the classifier can be obtained as Eq. (3), by a series of mathematical derivation.

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i,j=1}^{l}\alpha_i y_i \Phi(\mathbf{x}_i)^T \cdot \Phi(\mathbf{x}_j) + b\right) \tag{3}$$

The inner products are calculated using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \cdot \Phi(\mathbf{x}_j)$. A popular and capable kernel is the radial basis function with a parameter γ.

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\gamma|\mathbf{x} - \mathbf{x}_i|^2\right) \tag{4}$$

Only nonzero Lagrange multipliers α_i take part in establishing the final classifier, as indicated in Eq. (3). The data that have nonzero corresponding Lagrange multipliers are called support vectors. Simply, support vectors are those data that "support" the definition of the classifier. The classifier can thus be written as

$$f(\mathbf{x}) = \text{sign}\left(\sum_{k=1}^{m} \alpha_k y_k K(\mathbf{x}_k, \mathbf{x}) + b\right) \tag{5}$$

where $\mathbf{x}_k$ is the support vector, and m is the number of support vectors. The support vector classifier has two parameters, C and γ, to be calibrated. Given parameters C and γ, the Lagrange multipliers α_i and parameter b in Eq. (5) can be determined by the SVM algorithm.

Support vector regression

Let data $[(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)]$ be a given training set, where $\mathbf{x}_i$ is an input vector, and y_i is its corresponding output. Support vector regression (SVR) finds a regression function $f(\mathbf{x}) = \mathbf{w}^T \cdot \Phi(\mathbf{x}) + b$ that best describes the observed output y with an error tolerance ε, where $\mathbf{w}$ and b are parameters, and $\Phi(\mathbf{x})$ is a non-linear function. The penalized losses L_ε, when data are outside of the tube of error tolerance, are defined by the Vapnik's ε-insensitive loss function.

$$L_\varepsilon(y_i) = \begin{cases} 0 & \text{for } |y_i - [\mathbf{w}^T \cdot \Phi(\mathbf{x}_i) + b]| < \varepsilon \\ |y_i - [\mathbf{w}^T \cdot \Phi(\mathbf{x}_i) + b]| - \varepsilon, & \text{for } |y_i - [\mathbf{w}^T \cdot \Phi(\mathbf{x}_i) + b]| \geqslant \varepsilon \end{cases} \tag{6}$$

The SVR problem is then formulated as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi^*} \quad & \tfrac{1}{2}\mathbf{w}^T \cdot \mathbf{w} + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - [\mathbf{w}^T \cdot \Phi(\mathbf{x}_i) + b] \leqslant \varepsilon + \xi_i \\ & [\mathbf{w}^T \cdot \Phi(\mathbf{x}_i) + b] - y_i \leqslant \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geqslant 0, i = 1, 2, \ldots, l \end{aligned} \tag{7}$$

where ξ_i and ξ_i^* indicate slack variables specifying the upper and lower training errors subject to the error tolerance ε, and C is a positive constant that determines the degree of penalized loss when a training error occurs. The optimization problem is solved by a dual set of Lagrange multipliers, α_i and α_i^*. Consequently, the approximate function can be expressed as

$$f(\mathbf{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\Phi(\mathbf{x}_i)^T \cdot \Phi(\mathbf{x}) + b \tag{8}$$

The data with nonzero Lagrange multipliers $(\alpha_i - \alpha_i^*)$ are actively in the regression function, and are the support vectors. With the use of a kernel function (herein, the radial basis function with a parameter γ), the regression function can be rewritten as

$$f(\mathbf{x}) = \sum_{k=1}^{m}(\alpha_k - \alpha_k^*)K(\mathbf{x}_k, \mathbf{x}) + b \tag{9}$$

where $\mathbf{x}_k$ is the support vector, and m is the number of support vectors. The support vector regression has three parameters, C, ε and γ, that need to be calibrated. Given these three parameters, the Lagrange multipliers and parameter b in Eq. (9) can be determined by the SVM algorithm.

Multivariate analysis for classification and regression

Discriminant analysis

Discriminant analysis, originally developed by Fisher (1936), finds a linear discriminant function L to determine the class of a predictand based on a set of n predictors $(x_1, x_2, \ldots, x_n)$.

$$L = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{10}$$

The parameters $(a_0, a_1, a_2, \ldots, a_n)$ are calibrated from the training data of predictors and a predefined class label (for example,+1 and −1) of the predictand. The linear discriminant function L is then used to predict the class of a new predictand according to the estimated class label. Details of the discriminant analysis can be found in textbooks such as McLachlan (1992) and Huberty (1994).

Multiple regression

Multiple regression is a form of regression analysis in which the regression function establishes the relationship between one dependent variable y and more than one independent variables $(x_1, x_2, \ldots, x_n)$. A linear regression equation is in the following form:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \tag{11}$$

Parameters $(b_0, b_1, b_2, \ldots, b_n)$ are estimated using the least squares method. Mertler and Vannatta (2005) describe multiple regression in detail.

Statistical downscaling method (SDSM)

Statistical downscaling model (SDSM), developed by R.L. Wilby and C.W. Dawson (Wilby et al., 2002), is a free decision support tool based on statistical downscaling. SDSM uses multiple linear regression to construct a statistical relationship between large-scale predictors and local predictand, and also generates downscaled data stochastically. Therefore, SDSM is a hybrid downscaling model comprising a stochastic weather generator and a regression method. SDSM is among the most popularly used models for climate change impact studies, and is therefore used as a comparative downscaling model in this study. Wilby et al. (2002) and Wilby and Dawson (2004) describe SDSM in detail.

Results and discussion

This section presents three sets of analytical results. The calibration and validation of classification and regression models are presented in the first two subsections. The following subsection presents the downscaling results of both the proposed downscaling method and SDSM, and compares them with the observed data in the validation period. The final subsection shows the projections of daily precipitation using the proposed downscaling method and SDSM.

Classification models

Predictor selection for classification (Step A-1)

First, calibration data of reanalysis and local precipitations were divided into dry and wet groups according to the local precipitation, pertaining to two seasons. The Two-sample Kolmogorov–Smirnov test was then performed to select predictors of reanalysis data with different values in the dry-day and wet-day groups. Analytical results indicate that most weather factors were statistically significantly different between two groups, under a significance level of 0.01, and therefore were considered as suitable predictors for classification model. However, using too many predictors was

found to cause overfitting. Therefore, the less statistically significant large-scale factors were removed from the predictors. The resulting classification models have six predictors for the dry season, and five for the wet season. The left side of Table 2 lists the predictors used in the classification models. Humidity-related factors (*R500*, *R850* and *Rhum*, cf. Table 1 for description of factors) were selected as predictors to classify a day as dry or wet, as would be rationally expected. Additionally, vorticity (*P8_z* and *P_z*) and 850 hPa geopotential height (*P850*) were also reasonably selected as predictors for the dry season. Predictors for the wet season included meridional velocity (*P5_v*) and divergence (*P5zh*). Although those two factors at 500 hPa height is not supported by meteorological evidence, statistical analysis indicates that they are suitable for dry/wet day classification.

Classification model development (Step A-2)

Local precipitation and reanalysis data in the calibration period were used to develop classification models. The support vector classification (SVC) and the discriminant analysis were applied to the dry/wet day classification. Two classification models for dry and wet seasons were built under the conditions of seasonal separation. The SVC model has two parameters. The left side of Table 3 lists the calibrated SVC parameters, which the SVC adopts to perform the dry/wet day classification. The classification model was also developed by discriminant analysis. The constructed discriminant function based on Eq. (10) for the dry season is

$$L_{\text{dry season}} = 0.59 + 0.62 \cdot P8\_z - 0.60 \cdot P850 - 0.05 \cdot P\_z + 0.59 \cdot R500 + 0.64 \cdot R850 + 0.15 \cdot Rhum \quad (12)$$

The discriminant function for the wet season is

$$L_{\text{wet season}} = 0.06 + 0.04 \cdot P5\_v - 0.32 \cdot P5zh + 0.67 \cdot R500 + 0.58 \cdot R850 + 0.10 \cdot Rhum \quad (13)$$

where the normalized values are used for the predictors in Eqs. (12) and (13), and L is the class label, of which the positive value designates a wet day, and the negative value designates a dry day.

Classification results

The accuracy of dry/wet day classification was measured as in Eq. (14).

$$\text{Accuracy} = \frac{\hat{D}|D + \hat{W}|W}{D + W} \times 100(\%) \quad (14)$$

where D is the number of dry days; W is the number of wet days; $\hat{D}|D$ indicates the number of days that a dry day is correctly classified as a dry day, and $\hat{W}|W$ indicates the number of days that a wet day is correctly classified as a wet day. Table 4 lists the classification accuracies for the calibration period (1964–1990) and validation period (1991–2000), respectively. The classification accuracy was generally higher than 70%, except for the multivariate model (discriminant function) during the wet season, as indicated in Table 4. Fig. 3 shows monthly classification accuracies during the calibration and validation periods, and reveals that the two models performed similarly in the dry season, while the multivariate model is less accurate in the middle of the wet season.

Table 2
Predictors used in classification and regression models.

Classification		Regression	
Dry season	Wet season	Dry season	Wet season
P8_z	*P5_v*	*P8_z*	*P8_z*
P850	*P5zh*	*R500*	*R500*
P_z	*R500*	*R850*	*R850*
R500	*R850*	*Rhum*	*Rhum*
R850	*Rhum*		
Rhum			

Note: See Table 1 for description of predictors.

Table 3
Parameters of SVC and SVR models.

	SVC		SVR		
	C	γ	C	γ	ε
Dry season	10	0.017	4.9	1.292	0.022
Wet season	40	0.091	48.3	0.012	1.296

Table 4
Accuracy (%) of dry/wet day classification.

	Calibration		Validation	
	SVC	Multivariate	SVC	Multivariate
Dry season	73.1	71.7	73.4	72.8
Wet season	75.6	69.4	71.5	65.1
Whole year	74.3	70.6	72.4	69.0

Regression models

Predictor selection for regression (Step B-1)

Spearman's rank correlation coefficient was used to choose correlated factors as predictors for regression. Results show that most large-scale weather factors were statistically correlated with local precipitation under the significance level of 0.01. Factors with higher correlation coefficient were further chosen as the predictors for regression models. The right side of Table 2 lists the predictors used in the regression models. Humidity factors (*R500*, *R850* and *Rhum*) and vorticity (*P8_z*) were logically selected as predictors to calculate the daily precipitation amount in both dry and wet seasons.

Regression model development (Step B-2)

The local precipitation amount and reanalysis data of the calibration period were used to calibrate the parameters of regression models, support vector regression (SVR) and multiple regression. The SVR model has three parameters. Table 3 lists the calibrated parameters, which are adopted by the SVR to apply precipitation amount regression. Multiple regression was also performed to construct the regression model. The established multiple regression function, from Eq. (11), for the dry season is

$$y_{\text{dry season}} = 5.91 + 5.06 \cdot P8\_z + 3.31 \cdot R500 + 1.18 \cdot R850 - 0.09 \cdot Rhum \quad (15)$$

The multiple regression function for the wet season is

$$y_{\text{wet season}} = 7.09 + 13.97 \cdot P8\_z + 2.67 \cdot R500 - 3.86 \cdot R850 + 1.13 \cdot Rhum \quad (16)$$

where y indicates the precipitation amount (normalized value), conditional on a wet day, and the values of predictors used in Eqs. (15) and (16) are normalized values. The precipitation amounts calculated by the SVR and the multiple regression models are normalized values, and are then transformed to their original scale.

Regression results

Wet days in the calibration and validation periods were extracted for the construction and assessment of regression models (the SVR and the multivariate model). The outputs of regression models can be negative values. Under this circumstance, those negative values were set to zero. The percentages of negative

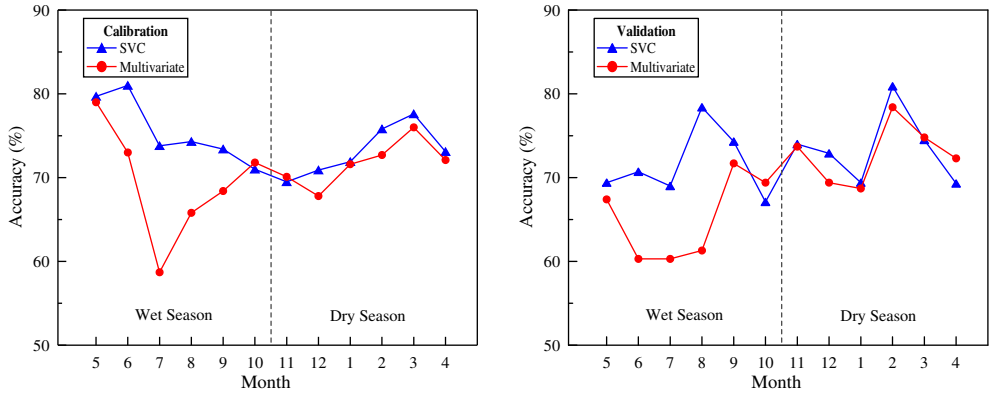

Fig. 3. Classification accuracy for each month (a) Left: calibration (b) Right: validation.

values among the total number of wet days were 10.1% in the SVR, and 14.4% in the multivariate model in the calibration. As for the validation, the percentages of negative values to total number of wet days were 17.6% for the SVR, and 16.4% for the multivariate model.

Statistical properties of daily precipitation data are important for future projection. Table 5 lists the regression results on wet days to compare the performance of the two regression models.

Table 5
Statistics of regression results on wet days.

	Calibration			Validation		
	Mean (mm)	SD (mm)	CS	Mean (mm)	SD (mm)	CS
Dry season						
Observation	**6.03**	**11.39**	**4.18**	**6.83**	**12.65**	**4.26**
SVR	5.47	6.61	6.49	6.16	7.59	2.80
Multivariate	6.23	4.73	0.68	6.43	5.06	0.76
Wet season						
Observation	**13.88**	**34.81**	**7.66**	**14.32**	**34.84**	**7.87**
SVR	13.68	32.14	8.99	14.28	28.56	6.98
Multivariate	14.81	18.67	3.18	14.76	21.83	3.99
Whole year						
Observation	**10.23**	**26.90**	**9.35**	**10.79**	**27.03**	**9.40**
SVR	9.86	24.27	11.63	10.50	21.89	8.78
Multivariate	10.81	14.66	4.18	10.89	16.85	5.23

The mean values in both models were close to the observed values. In general, the mean was underestimated by the SVR model, and overestimated by the multivariate model (multiple regression). However, the multivariate model underestimated the mean in the dry season in the validation dataset. Both models underestimated the standard deviation (SD), but the SVR model yielded better results than the multivariate model. The SD describes the dispersion of the precipitations from their mean. Therefore, downscaled precipitations with smaller SD distribute closer around the mean value than the observations. The coefficient of skewness (CS) relates to the occurrence of extreme values. The SVR had estimated the CS much more accurately than the multivariate model, indicating that the SVR outperforms the multivariate model in downscaling extreme precipitations.

Fig. 4 shows the quantile-to-quantile plot for validation data of daily precipitation regression on wet days. Fig. 4a shows all precipitation data, and demonstrates that the SVR significantly outperformed the multivariate model when precipitations were larger than 100 mm/day. This is consistent with the results of statistics comparison that the SVR better estimates the CS. Fig. 4b enlarges the quantile-to-quantile plot for daily precipitations less than 100 mm, clearly showing that the SVR downscaling data were closer than the multivariate model data to the observations. An inverse S-shape curve at the interval of about 0–20 mm indicates that many downscaled precipitations scattered around 10 mm. This finding also reflects the statistic properties of the mean and SD. Overall, SVR performed better than the multivariate model: it

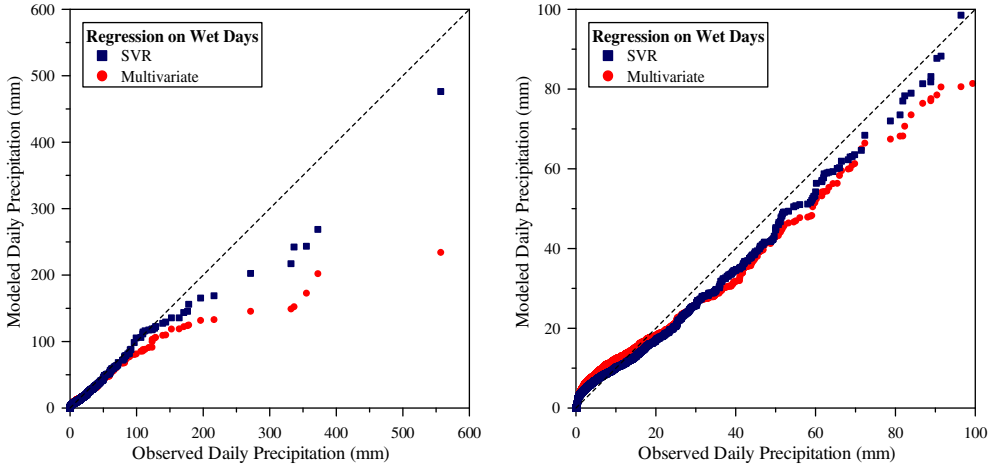

Fig. 4. Quantile-to-quantile plot for daily precipitation regression on wet days. (a) Left: All precipitation data during validation period. (b) Right: Data for precipitation less than 100 mm/day.

accurately computed the daily precipitation, except for underestimating some extreme precipitation amounts.

Downscaling results

This section presents the validation results of the proposed downscaling method, which combines the classification of days as dry or wet with the regression of precipitation amount on wet days. The proposed downscaling method only performed regression on wet days, as determined by the classification model. Therefore, regression results may be influenced by misclassification. Additionally, the downscaling method treats a day with zero precipitation as a dry day, even if the zero value is the result of the regression model outputting a negative value and then replaces this negative value with zero. Moreover, SDSM was also run to downscale daily precipitation in the study area, and its results compared with those of the support vector machine (SVM) and the multivariate model.

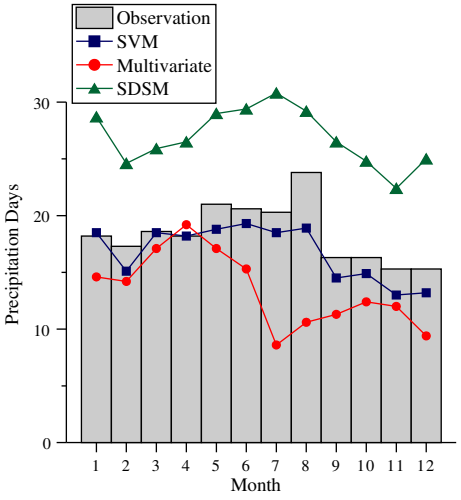

Fig. 5. Average monthly precipitation days during validation period.

Precipitation days

The downscaling models were analyzed using information about precipitation days. Fig. 5 presents downscaled results of monthly precipitation days (daily precipitation > 0 mm) for the validation period. The SVM model outperformed the multivariate model and the SDSM, because the multivariate model underestimated the number of precipitation days, and the SDSM overestimated it.

Precipitation amount

The amount of precipitation is important for the assessment of downscaling results. Figs. 6 and 7 illustrate the downscaled daily precipitation in February of the dry season and August of the wet season, respectively, for the validation period. All downscaling models identified the general tendency of the daily precipitation. However, in February (Fig. 6), the multivariate model did not predict greater daily precipitation (for example, >20 mm) effectively, but the SVR and SDSM predicted some greater daily precipitation. All models performed satisfactorily for the August data (Fig. 7). Some severe precipitation events were successfully reproduced. Table 6 lists the statistical properties of the observed and downscaled daily precipitation regarding the validation period. The

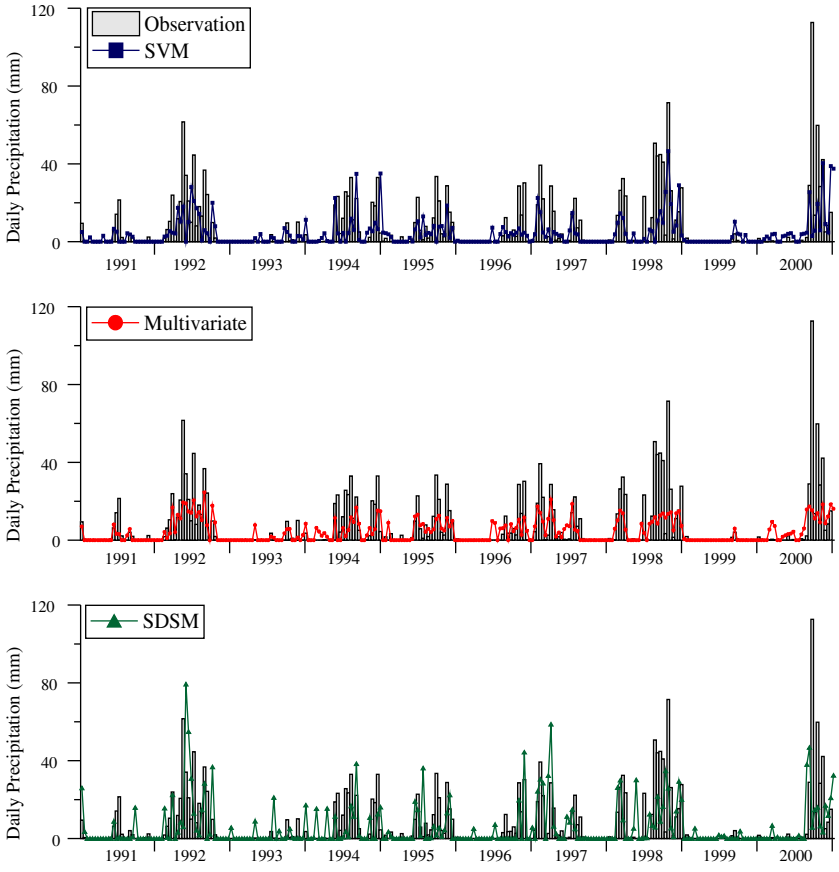

Fig. 6. Downscaling results of daily precipitation in February during validation period.

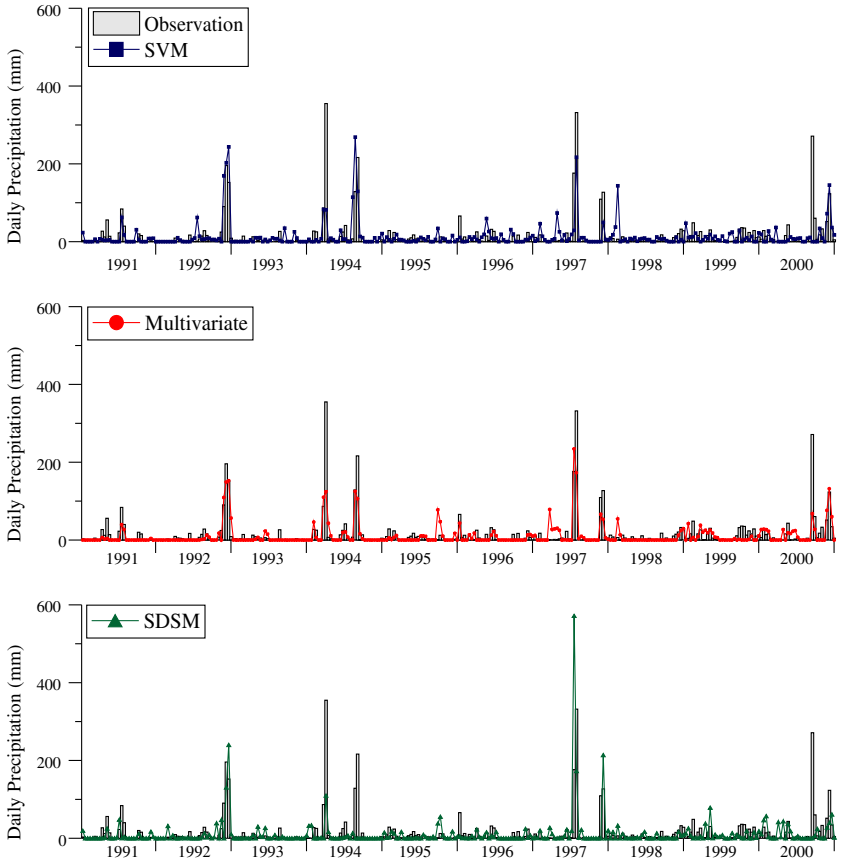

Fig. 7. Downscaling results of daily precipitation in August during validation period.

Table 6
Statistics of downscaling results.

	Mean (mm)	SD (mm)	CS
Dry season			
Observation	**3.86**	**10.11**	**5.48**
SVM	3.76	6.51	3.40
Multivariate	3.94	5.11	1.16
SDSM	3.83	6.50	3.65
Wet season			
Observation	**9.10**	**28.35**	**9.56**
SVM	9.99	24.82	8.07
Multivariate	8.22	18.17	5.09
SDSM	8.12	20.48	11.74
Whole year			
Observation	**6.50**	**21.50**	**11.58**
SVM	6.90	18.47	10.36
Multivariate	6.10	13.56	6.59
SDSM	5.99	15.38	14.40

mean values calculated from three models were comparable to the observation. All models underestimated the standard deviation (SD), but the SVM yielded the most accurate estimation, and thus outperformed other models in predicting precipitation variation. The SVM also estimated the coefficient of skewness (CS) well. The multivariate model underestimated the CS, while the SDSM overestimated it. These findings indicate that the SVM performs well in extreme precipitation.

Fig. 8 shows the quantile-to-quantile plot for daily precipitation downscaling. Fig. 8a displays all validation data, and demonstrates that the SVR significantly outperformed other models. The SVM and the SDSM performed better than the multivariate model for large daily precipitation. Fig. 8b enlarges the quantile-to-quantile plot for daily precipitations less than 100 mm. The SDSM performed better than other models at the interval of about 0–10 mm/day, but underestimated when daily precipitation is larger than 10 mm. Overall, the SVM performed better than other models, and exhibited good agreement with the observations.

Extreme precipitation

This work adopted an extreme precipitation event as daily precipitation exceeding 50 mm, which is a "heavy rain" event as defined by Central Weather Bureau in Taiwan. Totally 75 heavy rain events occurred during the validation period. Table 7 lists the number of statistically accurately predicted extreme events, and the accuracy levels. Fig. 9 shows the distribution of the extreme events, and indicates that the downscaled extreme daily precipitations obtained by SVM had similar statistical properties to the observations, except for slight underestimation of a few events between 150 and 200 mm/day, and severe underestimation of all events larger than 200 mm/day. The SDSM showed better prediction than the multivariate model in events higher than 150 mm/day, but obviously underestimated the number of extreme events. Overall, the SVM outperformed other models, and reproduced extreme daily precipitations well.

Projected daily precipitation

This section presents projected precipitations under the climate change scenarios. The predictors from HadCM3 scenario outputs (A2 and B2) were used in the downscaling models to project future

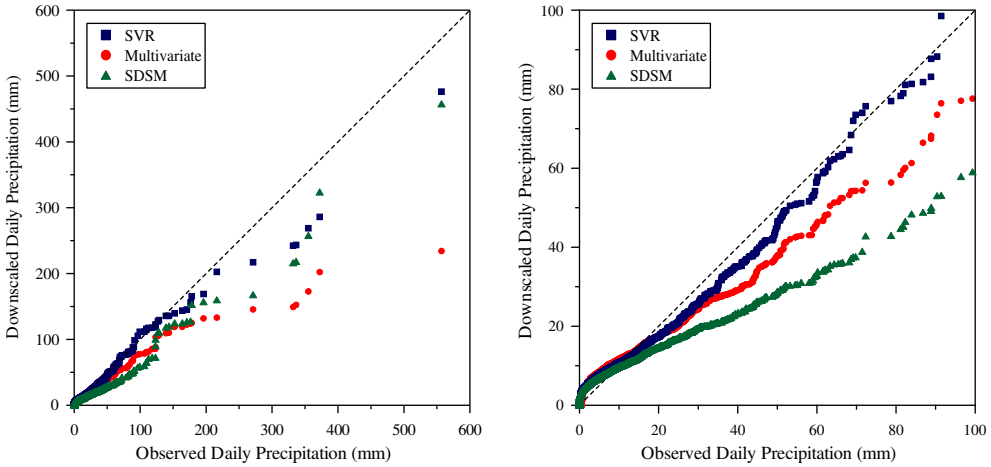

Fig. 8. Quantile-to-quantile plot for downscaling daily precipitation. (a) Left: All precipitation data. (b) Right: Precipitation data less than 100 mm/day.

Table 7
Number of extreme events (daily precipitation > 50 mm) during validation period.

	No. of events	Accuracy (%)
Observation	**75**	–
SVM	65	87
Multivariate	48	64
SDSM	30	40

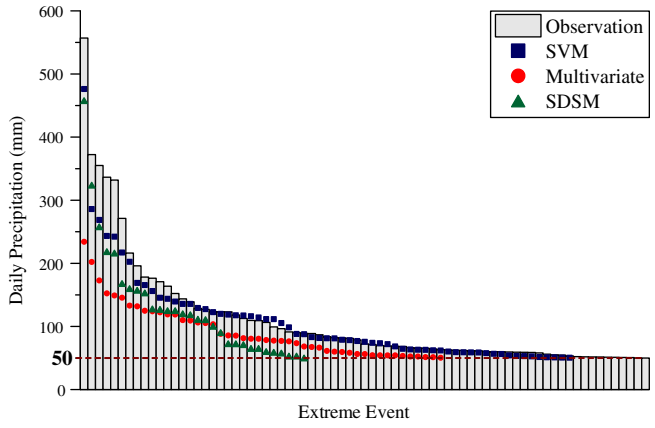

Fig. 9. Downscaling results of extreme events during validation period.

precipitation (2010–2099). Fig. 10 illustrates future annual precipitations downscaled by three downscaling models, and shows a similar pattern of the projected precipitations. Clearly, the SVM projected the highest annual precipitation, and the SDSM projected the lowest. However, the downscaling model possesses bias, i.e., difference of mean values in Table 6. Therefore, the change relative to baseline should refer to the model output in the validation period (1991–2000). Hollow symbols in Fig. 10 represent the mean annual precipitation downscaled by models in the validation period. The multivariate model thus predicted a larger increase in annual precipitation than the SVM. The SDSM projected a slight decrease in precipitation in the A2 scenario, and predicted a small increase followed by a large decrease about B2 scenario.

Fig. 11 illustrates the number of projected extreme precipitation events per year. Only the projection data of the SVM are presented, because the other two downscaling models did not well reproduce extreme precipitation events during the validation period (1991–2000). The observed number of extreme precipitation events per year during the validation period was 7.5, while the number predicted by SVM was 6.5. Both the A2 and B2 scenario data projected an increase in extreme precipitation events. The results of daily precipitation projection can provide information for the authorities about management and operation of Shih-Men Reservoir under climate change.

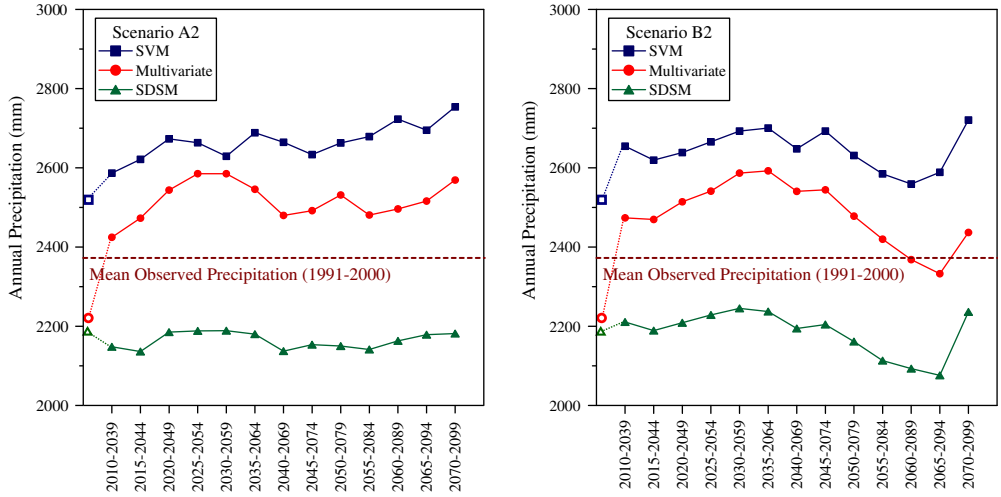

Fig. 10. Projected annual precipitation.

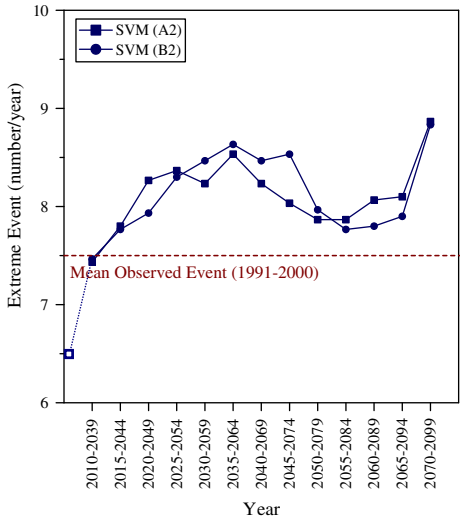

Fig. 11. Projected number of extreme events per year.

Conclusions and future work

This work proposes a daily precipitation downscaling method consisting of classification and regression. The proposed method was developed using SVM and multivariate analysis, and was compared with the SDSM. Downscaled results reveal that the SVM produced more accurate daily precipitation than the SDSM and the multivariate model. Downscaled daily precipitation by the SVM presented comparable quantile precipitations to observations, and showed good agreement with the extreme events. The SDSM was reported to outperform other models in some works (e.g., Harpham and Wilby, 2005; Khan et al., 2006), but did not perform very well for downscaling daily precipitation in this study area. Although the SDSM performed better than other models in relation to small daily precipitation, it underestimated the quantile precipitations.

Daily precipitation was divided into two categories, dry-day and wet-day. Regression was then applied on the wet-day class to obtain the precipitation amount. Future study could categorize wet days more finely, for example, by considering days of light, medium and heavy rain, and then construct different regression models for each class. This approach could improve estimation of precipitation amount on wet days, and reduce the number of negative values calculated by the regression model. Moreover, this work adopted large-scale weather factors at only the nearest GCM data grid to develop the downscaling model, and used A2 and B2 scenario data projected by HadCM3. Future work should consider large-scale weather factors from a region covering more grids in order to select the predictors and then to construct the downscaling model. Also, season separation accounting for the date of onset and duration in each year (e.g., Tripathi et al., 2006; Anandhi et al., 2008) can be adopted rather than the "hard" separation in wet and dry season by the months. Furthermore, additional GCM models, and their updated projection data, could be used to investigate the possible change in future daily precipitation in the study area.

Acknowledgement

The authors would like to thank National Cheng Kung University, Taiwan for financially supporting this study under Landmark Program (Project No. R046).

References

Anandhi, A., Srinivas, V.V., Nanjundiah, R.S., Kumar, D.N., 2008. Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. International Journal of Climatology 28, 401–420.
Bárdossy, A., 1997. Downscaling from GCMs to local climate through stochastic linkages. Journal of Environmental Management 49 (1), 7–17.
Bárdossy, A., Plate, E., 1992. Space-time model for daily rainfall using atmospheric circulation patterns. Water Resources Research 28, 1247–1259.
Buishand, T.A., Brandsma, T., 2001. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest neighbor resampling. Water Resources Research 37 (11), 2761–2776.
Burger, G., 1996. Expanded downscaling for generating local weather scenarios. Climate Research 7 (2), 111–128.
Chu, J.-L., Kang, H., Tam, C.-Y., Park, C.-K., Chen, C.-T., 2008. Seasonal forecast for local precipitation over northern Taiwan using statistical downscaling. Journal of Geophysical Research 113, D12118. doi:10.1029/2007JD009424.
Dibike, Y.B., Coulibaly, P., 2006. Temporal neural networks for downscaling climate variability and extremes. Journal of Hydrology 19, 135–144.
Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188.
Harpham, C., Wilby, R.L., 2005. Multi-site downscaling of heavy daily precipitation occurrence and amounts. Journal of Hydrology 312, 235–255.
Hewitson, B.C., Crane, R.G., 1996. Climate downscaling: techniques and application. Climate Research 7 (2), 85–95.
Huberty, C.J., 1994. Applied Discriminant Analysis. Wiley, New York.
Kaas, E., Li, T.S., Schmith, T., 1996. Statistical hindcast of wind climatology in the North Atlantic and northwestern European region. Climate Research 7 (2), 97–110.
Khan, M.S., Coulibaly, P., Dibike, Y., 2006. Uncertainty analysis of statistical downscaling methods. Journal of Hydrology 319, 357–382.
Landman, W.A., Mason, S.J., Tyson, P.D., Tennant, W.J., 2001. Statistical downscaling of GCM simulations to streamflow. Journal of Hydrology 25 (1–4), 221–236.
McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.
Menzel, L., Burger, G., 2002. Climate change scenarios and runoff response in the Mulde catchment (Southern Elbe, Germany). Journal of Hydrology 267 (1–2), 53–64.
Mertler, C.A., Vannatta, R.A., 2005. Advanced and multivariate statistical methods: practical application and interpretation, third ed. Pyrczak, Glendale, CA.
Murphy, J., 2000. Predictions of climate change over Europe using statistical and dynamical downscaling techniques. International Journal of Climatology 20 (5), 489–501.
Olsson, J., Uvo, C.B., Jinno, K., 2001. Statistical atmospheric downscaling of short-term extreme rainfall by neural networks. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 26 (9), 695–700.
Palutikof, J.P., Goodess, C.M., Watkins, S.J., Holt, T., 2002. Generating rainfall and temperature scenarios at multiple sites: examples from the Mediterranean. Journal of Climate 15 (24), 3529–3548.
Selker, J.S., Haith, D.A., 1990. Development and testing of single-parameter precipitation distributions. Water Resources Research 26 (11), 2733–2740.
Tripathi, S., Srinivas, V.V., Nanjundiah, R.S., 2006. Downscaling of precipitation for climate change scenarios: a support vector machine approach. Journal of Hydrology 330, 621–640.
Tung, C.-P., Haith, D.A., 1995. Global warming effects on New York streamflows. Journal of Water Resources Planning and Management 121 (2), 216–225.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.
Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.
von Storch, H., Zorita, E., Cubasch, U., 1993. Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. Journal of Climate 6 (6), 1161–1171.
Wilby, R.L., Dawson, C.W., 2004. Using SDSM Version 3.1—A Decision Support Tool for the Assessment of Regional Climate Change Impacts. <http://unfccc.int/resource/cd_roms/na1/v_and_a/Resoursce_materials/Climate/SDSM/SDSM.Manual.pdf>.
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 and Software 17, 147–159.
Wilby, R.L., Whitehead, P.G., Wade, A.J., Butterfield, D., Davis, R.J., Watts, G., 2006. Integrated modelling of climate change impacts on water resources and quality in a lowland catchment: River Kennet, UK. Journal of Hydrology 330, 204–220.
Yu, P.-S., Yang, T.-C., Wu, C.-K., 2002. Impact of climate change on water resources in southern Taiwan. Journal of Hydrology 260 (1–4), 161–175.