Environmental Modelling & Software 49 (2013) 141–151

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A stepwise cluster analysis approach for downscaled climate projection – A Canadian case study

Xiuquan Wang [a], Guohe Huang [a, b, *], Qianguo Lin [b, c], Xianghui Nie [a], Guanhui Cheng [a], Yurui Fan [a], Zhong Li [a], Yao Yao [a], Meiqin Suo [c]

[a] Faculty of Engineering and Applied Science, University of Regina, Regina, Saskatchewan, Canada S4S 0A2
[b] SC Institute for Energy, Environment and Sustainability Research, North China Electric Power University, Beijing 102206, China
[c] MOE Key Laboratory of Regional Energy and Environmental Systems Optimization, North China Electric Power University, Beijing 102206, China

ARTICLE INFO

Article history:
Received 1 May 2013
Received in revised form
19 August 2013
Accepted 20 August 2013
Available online 17 September 2013

Keywords:
Downscaling
Stepwise cluster analysis
Climate change
Scenario
Impacts studies

ABSTRACT

Downscaling techniques are used to obtain high-resolution climate projections for assessing the impacts of climate change at a regional scale. This study presents a statistical downscaling tool, SCADS, based on stepwise cluster analysis method. The SCADS uses a cluster tree to represent the complex relationship between large-scale atmospheric variables (namely predictors) and local surface variables (namely predictands). It can effectively deal with continuous and discrete variables, as well as nonlinear relations between predictors and predictands. By integrating ancillary functional modules of missing data detecting, correlation analysis, model calibration and graphing of cluster trees, the SCADS is capable of performing rapid development of downscaling scenarios for local weather variables under current and future climate forcing. An application of SCADS is demonstrated to obtain 10 km daily mean temperature and monthly precipitation projections for Toronto, Canada in 2070–2099. The contemporary reanalysis data derived from NARR is used for model calibration (1981–1990) and validation (1991–2000). The validated cluster trees are then applied for generating future climate projections.

Software availability

Name: SCADS (version 2.0)
Developed by: Xiuquan Wang, Guohe Huang, Qianguo Lin, Xianghui Nie, Guanhui Cheng, Yurui Fan, Zhong Li, Yao Yao, Meiqin Suo
Contact: Xiuquan Wang, Faculty of Engineering and Applied Science, University of Regina, Regina SK S4S 7H9 Canada. Tel: +1-306-585-4095; Fax: +1-306-585-4855; Email: xiuquan.wang@gmail.com.
Available since: 2009
Website: http://env.uregina.ca/sca
Cost: Freeware

1. Introduction

Future projections of climate change can be obtained from Global Climate Models (GCMs) based on multiple emission scenarios. However, for assessing the impacts of climate change at a regional scale, outputs of GCMs cannot be used directly due to the mismatch in the spatial resolution between GCMs and impacts assessment models (Hashmi et al., 2009; Willems and Vrac, 2011). Generally, GCMs have spatial resolutions in the order of hundreds of kilometers, while a much finer resolution (in the range of tens of kilometers, or even less) is required for impact analysis. Downscaling techniques are therefore developed in recent years to handle the spatial mismatch as an alternative to improve regional or local estimates of variables from GCM outputs (Hessami et al., 2008).

According to reviews of previous studies (Hewitson and Crane, 1996; Wilby and Wigley, 1997; Wilby et al., 1998, 2004; Murphy, 1999; Mearns et al., 2003), downscaling techniques can be classified into dynamical and statistical. As a typical dynamical downscaling approach, Regional Climate Models (RCMs) cannot only generate precipitation and temperature time series that contain temporal and spatial correlation consistent with physical mechanisms, but also help identify out-of-sample climate conditions and mechanisms previously not observed. However, it is difficult for RCMs to quickly generate a large set of possible outcomes and to cost-effectively provide high resolution station data. By contrast, statistical downscaling mainly involves developing quantitative relationships between large-scale atmospheric variables (or

* Corresponding author. Faculty of Engineering and Applied Science, University of Regina, Regina, Saskatchewan, Canada S4S 0A2. Tel.: +1-306-585-4095; fax: +1-306-585-4855.
E-mail address: huang@iseis.org (G. Huang).

1364-8152/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.08.006

predictors) and local surface variables (or predictands), which is easier to implement with much lower computation requirements (Wilby et al., 2004). Therefore statistical downscaling approach is widely used in studies of climate change impacts (Heyen et al., 1996; Maak and von Storch, 1997; Beckmann and Adri Buishand, 2002; Huth, 2002; Wood et al., 2004; Fowler et al., 2007; Timbal et al., 2009; Hashmi et al., 2011; Phatak et al., 2011; Mullan et al., 2012). In general, statistical downscaling methods can be classified into three categories: weather classification schemes (e.g. analog method, fuzzy classification, Monte Carlo methods), regression models (e.g. linear regression, stochastic models, spell length methods, mixture modeling) and weather generators (e.g. neural networks, canonical correlation analysis). Correspondingly, a number of downscaling tools were recently developed to facilitate climate change impact studies. For example, Wilby et al. (2002) developed a regression-based downscaling tool known as SDSM; Hessami et al. (2008) proposed an automated statistical downscaling (ASD) tool based on SDSM; Semenov and Barrow (1997) developed a weather generator model known as the Long Ashton Research Station Weather Generator (LARS-WG); Willems and Vrac (2011) developed an artificial intelligence data driven model using the Gene Expression Programming (GEP) to create symbolic downscaling functions. Among these downscaling approaches, most of them assume that each predictand of interest is a function of predictors. This is especially true for regression-based models. However, there is no guarantee that such a functional relationship must exist between predictand and predictors. Although we can establish a functional relationship constrainedly by reducing the number of variables or introducing more assumptions, it might not be able to improve significantly the projection quality compared to coarser outputs of GCMs. To this end, a stepwise-cluster-analysis-based downscaling tool (SCADS) will be proposed in this study, which expresses the complex interactions between predictors and predictands as a cluster tree, without requiring assumptions of functional relationships.

The proposed downscaling tool is inspired by a stepwise cluster analysis (SCA) method which was firstly introduced by Huang (1992). The SCA has been widely applied for environmental studies over the past years. For example, Huang et al. (2006) developed a forecasting system for supporting remediation design and process control based on SCA; Qin et al. (2007) applied SCA for establishing a linkage between remediation actions and system responses. The main purpose of this study is to develop a downscaling tool based on SCA and to test its capability of obtaining finer scenarios from coarser outputs of GCMs or RCMs. The following sections start with an overview of the SCA method on its basic principle, modeling process, and software implementation. An illustrative example is then presented to obtain 10 km high-resolution climate projections of Toronto, Canada in 2070–2099 by downscaling a 25 km scenario outputted from the PRECIS (Providing REgional Climates for Impacts Studies) model – a regional climate modeling system developed by the Met Office Hadley Centre. The last section states the main conclusions and recommendations in terms of SCADS application as well as its limitation.

2. Methodology

2.1. Basic principle of SCA

The fundamental algorithm of SCA is based on the theory of multivariate analysis of variance (Morrison, 1967; Cooley and Lohnes, 1971; Overall and Klett, 1972). In SCA, sample sets of dependent variables will be cut or merged into new sets (i.e. children clusters) based on given criteria, and the values of independent variables will be used as references to determine which new set a sample in the original set (i.e. parent cluster) will enter (Huang et al., 2006). The construction of a SCA cluster tree requires multiple cutting and merging operations, such a process is actually to divide the original set of dependent variables into many irrelevant subsets according to specific criteria which will be described later in this section. The generated cluster tree can express the complex relations between predictors and predictands, it will be used to predict future values of predictands based cutting or merging operations are based on the Wilks' Λ statistic (Wilks, 1962), defined as $\Lambda = |\mathbf{E}|/|\mathbf{E}+\mathbf{H}|$, where $\mathbf{E}$ and $\mathbf{H}$ are the within- and between-group sums of squares and cross products matrices, respectively. Let two sets of dependent variables $\mathbf{e}$ and $\mathbf{f}$ contain n_e and n_f samples, denoted as the following vectors: $\mathbf{e_i} = (e_{1i}, e_{2i}, \ldots, e_{di})', i = 1, 2, 3, \ldots, n_e$, and $\mathbf{f_j} = (f_{1j}, f_{2j}, f_{3j}, \ldots, f_{dj})', j = 1, 2, 3, \ldots, n_f$, where d is the dimension of $\mathbf{e}$ and $\mathbf{f}$. Then the $\mathbf{H}$ and $\mathbf{E}$ can be given by:

$$\mathbf{E} = \sum_{i=1}^{n_e} (\mathbf{e}_i - \bar{\mathbf{e}})'(\mathbf{e}_i - \bar{\mathbf{e}}) + \sum_{j=1}^{n_f} \left(\mathbf{f}_j - \bar{\mathbf{f}}\right)' \left(\mathbf{f}_j - \bar{\mathbf{f}}\right) \tag{1}$$

$$\mathbf{H} = \frac{n_e n_f}{n_e + n_f} \left(\bar{\mathbf{e}} - \bar{\mathbf{f}}\right)' \left(\bar{\mathbf{e}} - \bar{\mathbf{f}}\right) \tag{2}$$

where $\bar{\mathbf{e}}$ is the sample mean of set $\mathbf{e}$, $\bar{\mathbf{f}}$ is the sample mean of set $\mathbf{f}$, respectively. They can be defined as follows:

$$\bar{\mathbf{e}} = \frac{1}{n_e} \sum_{i=1}^{n_e} \mathbf{e}_i \tag{3}$$

$$\bar{\mathbf{f}} = \frac{1}{n_f} \sum_{j=1}^{n_f} \mathbf{f}_i \tag{4}$$

For example, let

$$\mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} = \begin{bmatrix} 36 & 9.98 \\ 48 & 12.96 \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_3 \end{bmatrix} = \begin{bmatrix} 50 & 9.84 \\ 31 & 8.84 \\ 29 & 8.9 \end{bmatrix}$$

where $n_e = 2$, $n_f = 3$, $d = 2$, $\bar{\mathbf{e}} = (42, 11.47)$, $\bar{\mathbf{f}} = (36.67, 9.19)$. According to Equations (1) and (2), $\mathbf{E}$ and $\mathbf{H}$ can be calculated as follows:

$$\mathbf{E} = \begin{bmatrix} 340.67 & 30.75 \\ 30.75 & 5.07 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 34.13 & 14.57 \\ 14.57 & 6.22 \end{bmatrix}.$$

According to Rao's F-approximation (Rao, 1952), the Wilk's Λ statistic under the above two groups of samples can be correlated to a F-variant as follows:

$$F\left(d,\ n_e + n_f - d - 1\right) = \frac{1-\Lambda}{\Lambda} \cdot \frac{n_e + n_f - d - 1}{d} \tag{5}$$

As described in Wilk's likehood-ratio criterion (Wilks, 1962), the smaller the Λ value, the larger the difference between the sample means of sets $\mathbf{e}$ and $\mathbf{f}$. Since the Λ value is directly related to the F statistics, we can compare the sample means of the two data sets for significant differences through F-test (Huang et al., 2006; Qin et al., 2008). The null hypothesis would be H_0: $\mu_e = \mu_f$ versus the alternative hypothesis H_1: $\mu_e \neq \mu_f$, where μ_e and μ_f are population means of sets $\mathbf{e}$ and $\mathbf{f}$. Let the significance level be α. The criterion for cutting would be: $F_{cal} \geq F_\alpha$ and H_0 is false, which implies that differences of means between two sets are significant; whereas, $F_{cal} < F_\alpha$ and H_0 is true would be the merging criterion which indicates these two sets have no significant variations.

2.2. Building a cluster tree

As described above, the principle of SCA is to divide a cluster (i.e. a sample set) containing a number of dependent and independent variables into indivisible sub-clusters, based on a series of cutting and merging operations. Generally, the initial operation will be cutting action by which samples in a parent cluster will be divided into two groups. After that, merging and cutting operations will be performed step by step until none of the sub-clusters can be further divided or merged with other sub-cluster. To better understand such a stepwise clustering procedure, the following cluster (namely **C**) will be used as an example for illustrational purpose, consisting of independent variable **X** and dependent variable **Y**:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 39.9 & 27 \\ 9.1 & 8 \\ 11.4 & 14 \\ 20.5 & 29 \\ 27.3 & 26 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 29 & 8.9 \\ 36 & 9.98 \\ 48 & 12.96 \\ 50 & 9.84 \\ 31 & 8.84 \end{bmatrix},$$

$$\mathbf{C} = [\mathbf{X}, \mathbf{Y}] = \left[\begin{pmatrix} 39.9 & 27 \\ 9.1 & 8 \\ 11.4 & 14 \\ 20.5 & 29 \\ 27.3 & 26 \end{pmatrix} \begin{pmatrix} 29 & 8.9 \\ 36 & 9.98 \\ 48 & 12.96 \\ 50 & 9.84 \\ 31 & 8.84 \end{pmatrix} \right].$$

The detailed cutting operation is described as follows (Huang, 1992; Huang et al., 2006; Qin et al., 2008):

(1) Sequence the cluster in an ascending order by the kth column of the independent variables, where $k = 1, 2, \ldots, m$, m is the total number of columns. The ordered cluster can be expressed as $\mathbf{C}_s(k, r)$, where s shows step flag of operations, k and r indicate column and row index of the independent variables respectively. For example, if $k = 1$, the ordered cluster, $\mathbf{C_0}(1)$ will be:

$$\mathbf{C}_0(1) = [\mathbf{X}_0(1), \mathbf{Y}_0(1)] = \left[\begin{pmatrix} 9.1 & 8 \\ 11.4 & 14 \\ 20.5 & 29 \\ 27.3 & 26 \\ 39.9 & 27 \end{pmatrix} \begin{pmatrix} 36 & 9.98 \\ 48 & 12.96 \\ 50 & 9.84 \\ 31 & 8.84 \\ 29 & 8.9 \end{pmatrix} \right]$$

(2) For each ordered cluster, divide it into two groups iteratively by the index of the rth row of independent variables, where $r = 1, 2, \ldots, n$, n is the total number of rows. For example, if $r = 2$ and $k = 1$, we have two groups $\mathbf{C}_{01}(1, 2)$ and $\mathbf{C}_{02}(1, 2)$ as follows:

$$\mathbf{C}_{01}(1,2) = [\mathbf{X}_{01}(1,2), \mathbf{Y}_{01}(1,2)] = \left[\begin{pmatrix} 9.1 & 8 \\ 11.4 & 14 \end{pmatrix} \begin{pmatrix} 36 & 9.98 \\ 48 & 12.96 \end{pmatrix} \right],$$

$$\mathbf{C}_{02}(1,2) = [\mathbf{X}_{02}(1,2), \mathbf{Y}_{02}(1,2)] = \left[\begin{pmatrix} 20.5 & 29 \\ 27.3 & 26 \\ 39.9 & 27 \end{pmatrix} \begin{pmatrix} 50 & 9.84 \\ 31 & 8.84 \\ 29 & 8.9 \end{pmatrix} \right].$$

(3) Calculate the Wilk's $\Lambda_s(k, r)$ statistic for the pair of $\mathbf{Y}_{s1}(k, r)$ and $\mathbf{Y}_{s2}(k, r)$. For example, if $k = 1$ and $r = 2$, we have $\Lambda_0(1, 2)$ statistic for the pair of $\mathbf{Y}_{01}(1, 2)$ and $\mathbf{Y}_{02}(1, 2)$ as follows:

$$\Lambda_0(1,2) = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \frac{\left| \begin{bmatrix} 340.67 & 30.75 \\ 30.75 & 5.07 \end{bmatrix} \right|}{\left| \begin{bmatrix} 340.67 & 30.75 \\ 30.75 & 5.07 \end{bmatrix} + \begin{bmatrix} 34.13 & 14.57 \\ 14.57 & 6.22 \end{bmatrix} \right|} = 0.36.$$

(4) Identify the minimal value among all possible results of $\Lambda_s(k, r)$, denoted as $\Lambda_s(k, r)^*$, which will be used to calculate the F-value (i.e. F_{cal}) according to Equation (5). For the given example (C_0), the minimal value of Λ is achieved while $k = 2$ and $r = 4$ (i.e. the 4th row while sorted by the 2nd column). Let the significance level $\alpha = 0.05$, we have $F_{cal} = 54.71 > F_{0.05}(2, 2) = 19$. Thus, cluster **C** can be cut into two sets ($\mathbf{C}_1$ and $\mathbf{C}_2$) as follows:

$$\mathbf{C}_1 = [\mathbf{X}_1, \mathbf{Y}_1] = \left[\begin{pmatrix} 9.1 & 8 \\ 11.4 & 14 \\ 27.3 & 26 \\ 39.9 & 27 \end{pmatrix} \begin{pmatrix} 36 & 9.98 \\ 48 & 12.96 \\ 31 & 8.84 \\ 29 & 8.9 \end{pmatrix} \right],$$

$$\mathbf{C}_2 = [\mathbf{X}_2, \mathbf{Y}_2] = [(20.5 \quad 29)(50 \quad 9.84)].$$

where the classifications of samples in $\mathbf{C}_1$ and $\mathbf{C}_2$ are determined based on $x_2 \leq 27$ and $x_2 > 27$ respectively.

The merging process is relatively simple compared to the cutting operation. The judging criterion is also based on the F-value of two sub-clusters. If $F_{cal} < F_\alpha$, it means these two sub-clusters have no significant difference and therefore can be merged into one cluster. The cutting and merging processes will be performed alternately by following such a rule: when no cluster can be cut, mergence of clusters will be performed; when no cluster can be merged with another cluster, cutting action will be carried out; step by step, when all hypotheses of further cutting or mergence are rejected, a cluster tree can then be derived for each dependent variable (Huang, 1992; Huang et al., 2006). A typical cluster tree may include intermediate clusters, cutting and merging rules, leaf clusters, as well as related statistical information. If an intermediate cluster can be cut, the cutting rule must be specified to help determine which sub-cluster a new sample should belong to in the prediction process. A leaf cluster is a sample set that can no longer be cut or merged, the mean value of the sample set or an interval bounded by the maximum and minimum values of the sample set can be used to estimate the predicting results. The prediction process for a given independent sample sets is in fact a searching process starting from the top of the tree and ending at a leaf cluster, following the flow path guided by the cutting and merging rules (Huang et al., 2006).

3. Development of SCADS

In general, the process of obtaining downscaled climate projections can be summarized as four steps: 1) screening a set of predictors for each predictand of interest, which generally requires some necessary correlation analyses for each pair of predictor and predictand; 2) establishing a quantitative relationship between predictors and predictands based on sample data, which is named as "training" in this study; 3) validating the established relationship against observation data to evaluate its performance in reproducing historical climate; and 4) generating local climate projections based on the established relationship, this step is called "prediction" in this study. The route chart for the development of SCADS is illustrated in Fig. 1. The outcome of the training process is a cluster tree which can deal with continuous and discrete variables, as well as

nonlinear relationships among the variables. The inputs for this prediction effort are primarily from large-scale climate projections outputted by GCMs or RCMs.

The SCADS is a web-based downscaling tool such that users from any countries around the world can access it freely through Internet. The users of SCADS are required to register an account to make use of all provided functions. Fig. 2 shows the main interface of SCADS, through which users can log into the system with their account information created by themselves. In order to avoid slowing down the hosting web server and to manage all downscaling requests effectively, we applied a queuing rule (namely, first come first served) for the development of SCADS to control user requirements on computing resources and time consumption. The main functions of SCADS will be introduced in the following sections.

3.1. Missing-data detection

Missing data arise in almost all serious statistical analyses (Gelman and Hill, 2006). Non-response or unreasonable results due to missing data may lead to a distraction to research goals. Therefore, it is necessary to check if the sample collection contains missing data as well as to make clear their distribution in terms of time series and data matrix structure, so that suitable approaches could be chosen to handle the missing data. The SCADS has integrated missing-data detection function which helps user understand how many elements are missing and how these missing values are distributed in a sample collection. As illustrated in Fig. 3, the original data collection contains four missing elements which are indicated by "NA". The missing overview panel generated by

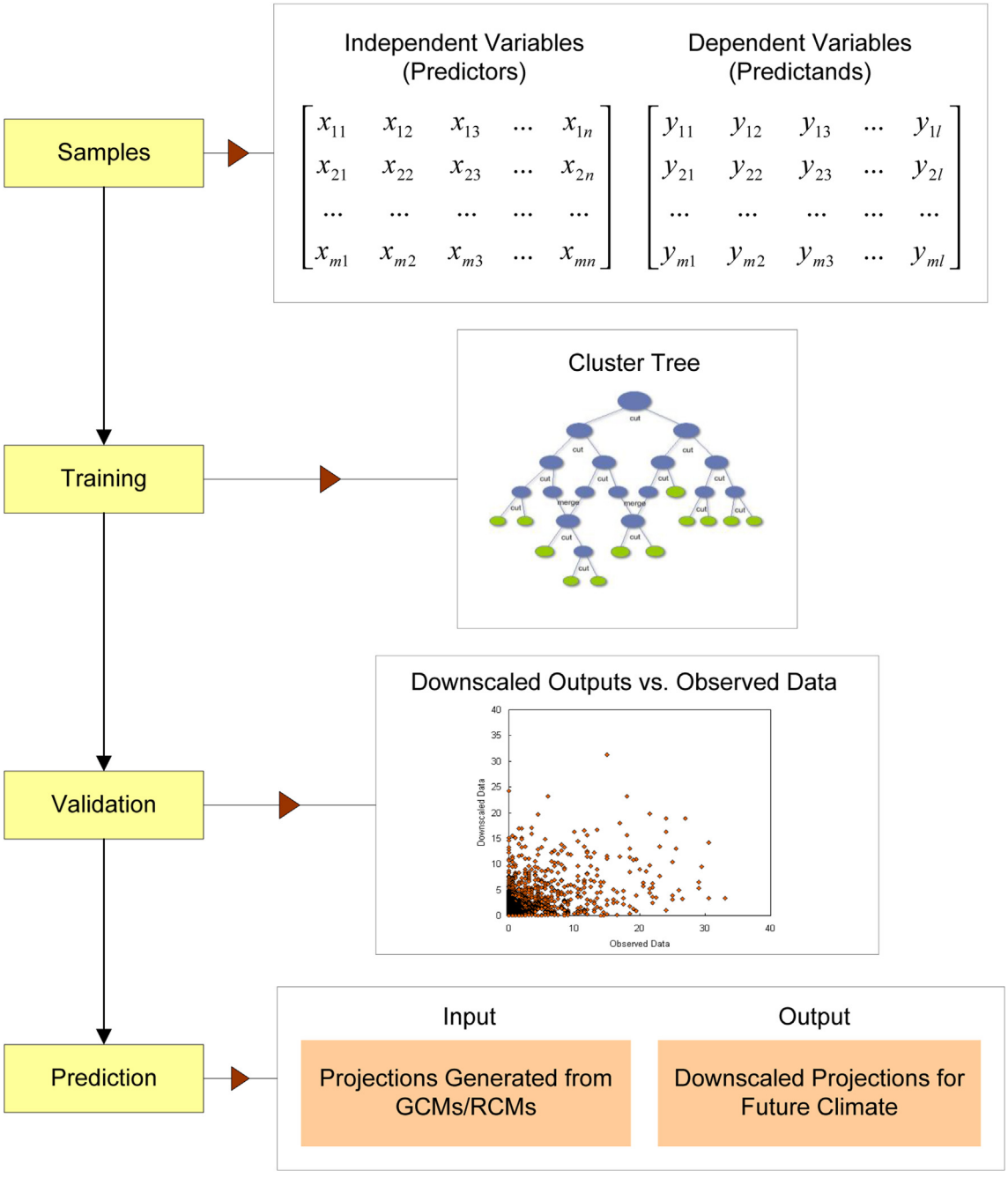


Fig. 1. Flow chart of SCADS.

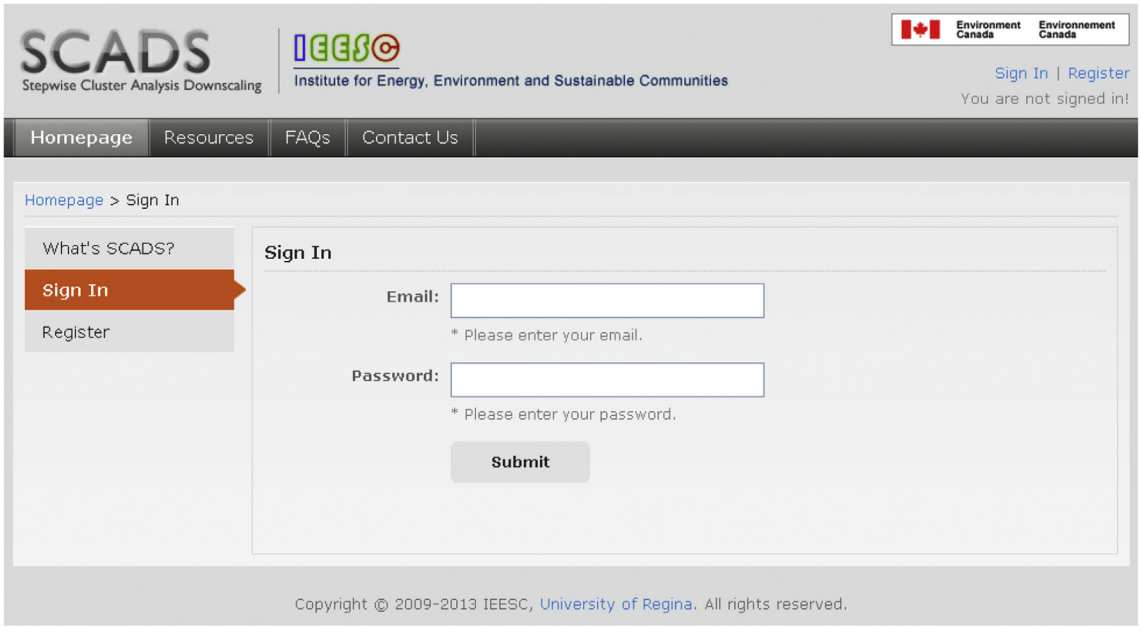
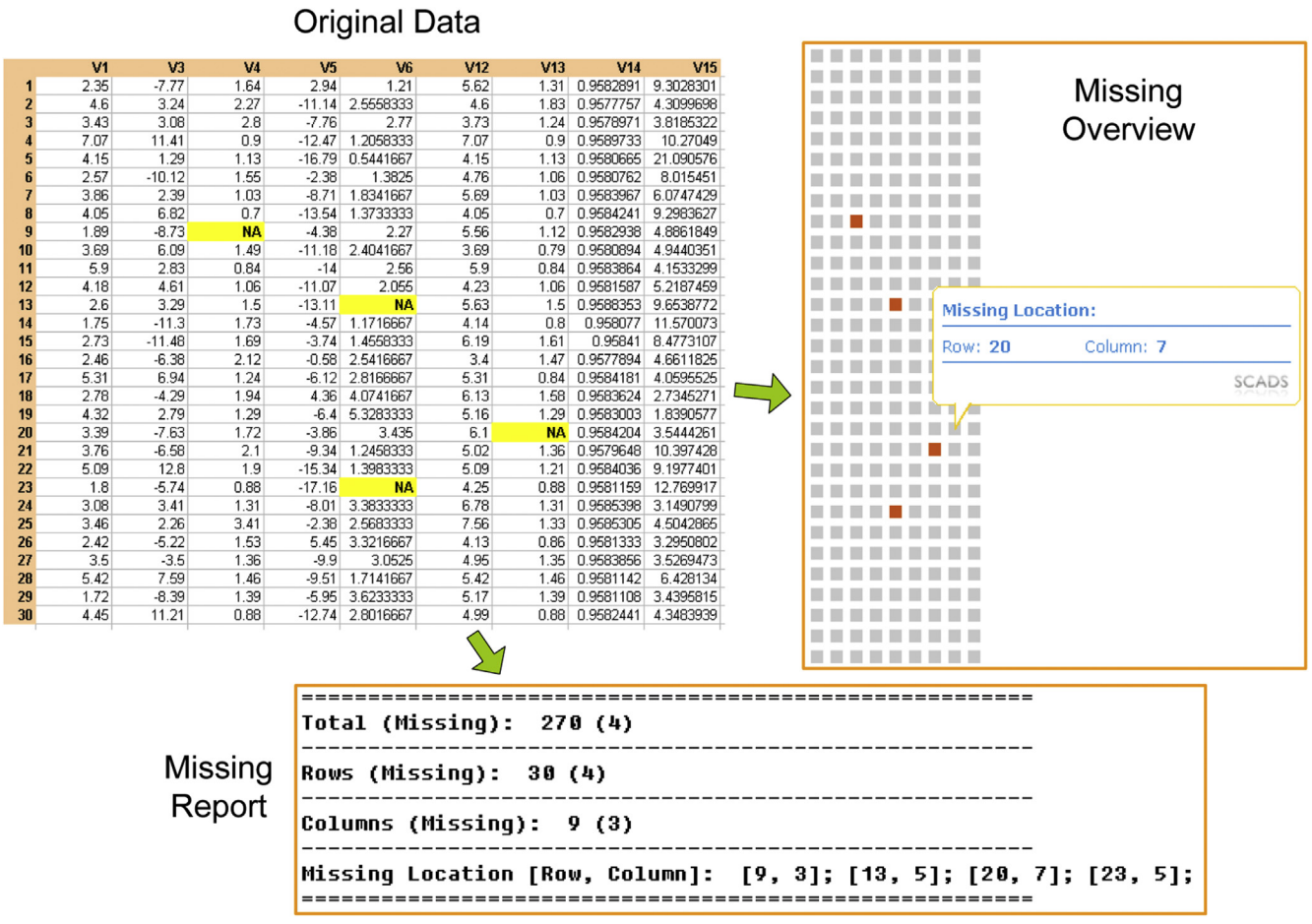


Fig. 2. Main interface of SCADS.



	V1	V3	V4	V5	V6	V12	V13	V14	V15
1	2.35	-7.77	1.64	2.94	1.21	5.62	1.31	0.9582891	9.3028301
2	4.6	3.24	2.27	-11.14	2.5558333	4.6	1.83	0.9577757	4.3099698
3	3.43	3.08	2.8	-7.76	2.77	3.73	1.24	0.9578971	3.8185322
4	7.07	11.41	0.9	-12.47	1.2058333	7.07	0.9	0.9589733	10.27049
5	4.15	1.29	1.13	-16.79	0.5441667	4.15	1.13	0.9580665	21.090576
6	2.57	-10.12	1.55	-2.38	1.3825	4.76	1.06	0.9580762	8.015451
7	3.86	2.39	1.03	-8.71	1.8341667	5.69	1.03	0.9583967	6.0747429
8	4.05	6.82	0.7	-13.54	1.3733333	4.05	0.7	0.9584241	9.2983627
9	1.89	-8.73	NA	-4.38	2.27	5.56	1.12	0.9582938	4.8861849
10	3.69	6.09	1.49	-11.18	2.4041667	3.69	0.79	0.9580894	4.9440351
11	5.9	2.83	0.84	-14	2.56	5.9	0.84	0.9583864	4.1533299
12	4.18	4.61	1.06	-11.07	2.055	4.23	1.06	0.9581587	5.2187459
13	2.6	3.29	1.5	-13.11	NA	5.63	1.5	0.9588353	9.6538772
14	1.75	-11.3	1.73	-4.57	1.1716667	4.14	0.8	0.958077	11.570073
15	2.73	-11.48	1.69	-3.74	1.4558333	6.19	1.61	0.95841	8.4773107
16	2.46	-6.38	2.12	-0.58	2.5416667	3.4	1.47	0.9577894	4.6611825
17	5.31	6.94	1.24	-6.12	2.8166667	5.31	0.84	0.9584181	4.0595525
18	2.78	-4.29	1.94	4.36	4.0741667	6.13	1.58	0.9583624	2.7345271
19	4.32	2.79	1.29	-6.4	5.3283333	5.16	1.29	0.9583003	1.8390577
20	3.39	-7.63	1.72	-3.86	3.435	6.1	NA	0.9584204	3.5444261
21	3.76	-6.58	2.1	-9.34	1.2458333	5.02	1.36	0.9579648	10.397428
22	5.09	12.8	1.9	-15.34	1.3983333	5.09	1.21	0.9584036	9.1977401
23	1.8	-5.74	0.88	-17.16	NA	4.25	0.88	0.9581159	12.769917
24	3.08	3.41	1.31	-8.01	3.3833333	6.78	1.31	0.9585398	3.1490799
25	3.46	2.26	3.41	-2.38	2.5683333	7.56	1.33	0.9585305	4.5042865
26	2.42	-5.22	1.53	5.45	3.3216667	4.13	0.86	0.9581333	3.2950802
27	3.5	-3.5	1.36	-9.9	3.0525	4.95	1.35	0.9583856	3.5269473
28	5.42	7.59	1.46	-9.51	1.7141667	5.42	1.46	0.9581142	6.428134
29	1.72	-8.39	1.39	-5.95	3.6233333	5.17	1.39	0.9581108	3.4395815
30	4.45	11.21	0.88	-12.74	2.8016667	4.99	0.88	0.9582441	4.3483939

Fig. 3. Illustration of missing-data detection.

SCADS gives a visual plot of the input data with missing elements highlighted in red color (in the web version). Meanwhile, a missing report is outputted by SCADS including detailed statistical information such as total number of missing elements, and missing location (row and column indices).

3.2. Correlation analysis

The correlation analysis function is intended to help screen a set of predictors effectively which is used to predict the value of corresponding predictand. In this study, we use the correlation coefficient (also known as R) as a criterion to measure the association between each pair of predictor and predictand. The correlation coefficient is a measure of the strength of the straight-line or linear relationship between two variables (Rodgers and Nicewander, 1988), by taking on values ranging between -1 and $+1$. The $+$ or $-$ sign denotes the direction of the correlation. The positive sign denotes direct correlation whereas the negative sign denotes inverse correlation. The closer the number moves towards 1, the stronger the correlation is. Zero signifies no correlation. Therefore, the correlation analysis module is designed based on correlation coefficient in two basic ways: to determine the predictive ability of a predictor and to determine the correlation between a predictor and its corresponding predictand. Fig. 4 shows an illustration of the correlation analysis module integrated in the current version of SCADS. The calculated correlation coefficient value is highlighted with different background colors (in the web version). A gradual color change from yellow (in the web version) to red (in the web version) indicates a gradual increase in the absolute value of correlation coefficient. By clicking on the correlation coefficient value for each pair of predictor and predictand, a pop-up window will be displayed containing a scatter plot which is helpful to understand the distribution pattern of all sample points as well as to identify extreme points effectively.

3.3. Training

Training process is to establish a relationship between predictors and predictands and to represent it using a quantified function or in other forms. In SCADS, the training-related transactions can be initialized and processed by creating a training job. The complicated relationship between predictors and predictands is expressed as a cluster tree. Three steps are required to create a training job in the current version of SCADS: 1) choose samples, 2) review samples, and 3) confirm and submit. It is recommended that users inspect the sample data before proceeding to create a new training job, with the aid of missing data checking and correlation analysis modules. If samples with missing data were inputted to a training job without any pre-processing, the SCADS would eliminate the entire data row as long as at least one element was missing. When creating a new training job, users will be asked to specify a friendly name to identify the job. The total time consumed by a training job often varies considerably. Generally, it depends on the sample size (i.e. total rows of sample collection), the numbers of predictors and predictands, and the correlation for each pair of predictor and predictand.

After a training job is submitted, the SCADS will decide whether it should be started immediately. If there are some jobs submitted before this job and at least one of them is waiting or running, the new job will be added into the waiting queue. Otherwise, it will begin to run right away. Once the training job is completed, the SCADS will output two plain-text files: tree file and map file, including cluster tree pathway and leaf nodes, respectively. A Windows-based utility, namely SCADS Cluster Tree Generator (CTG

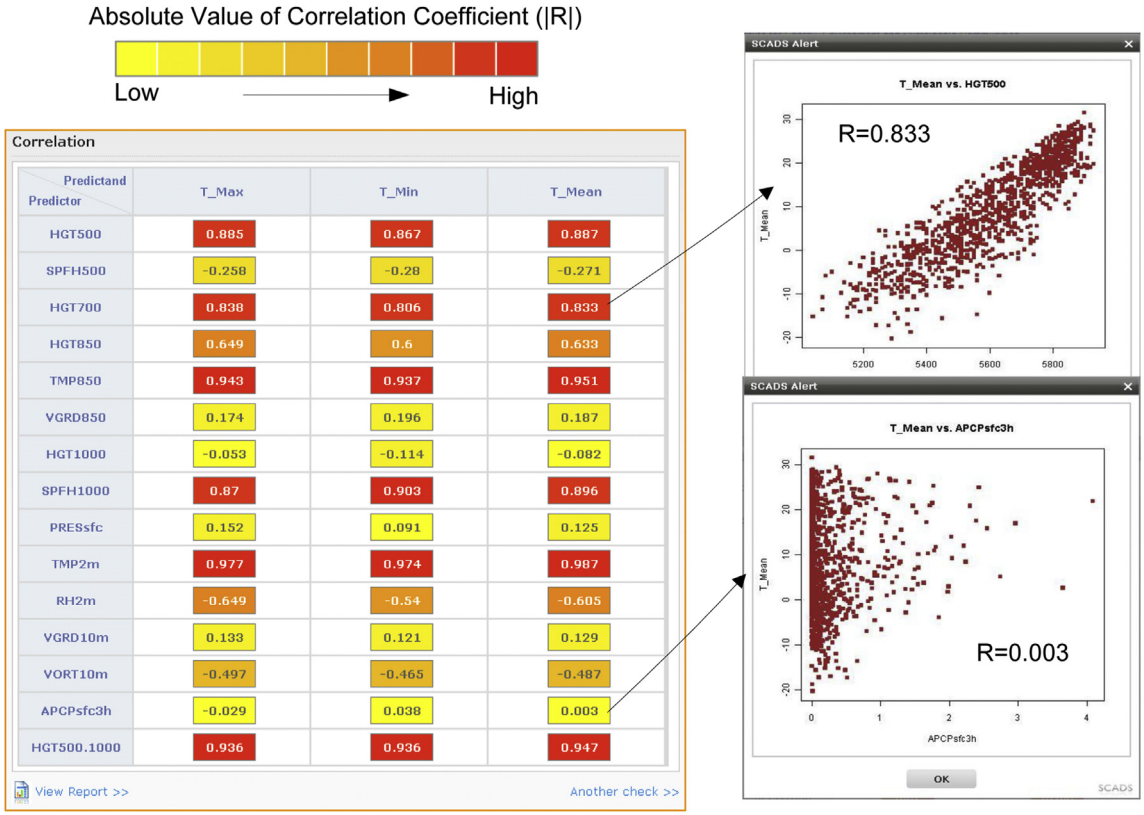


Fig. 4. Illustration of correlation analysis module.

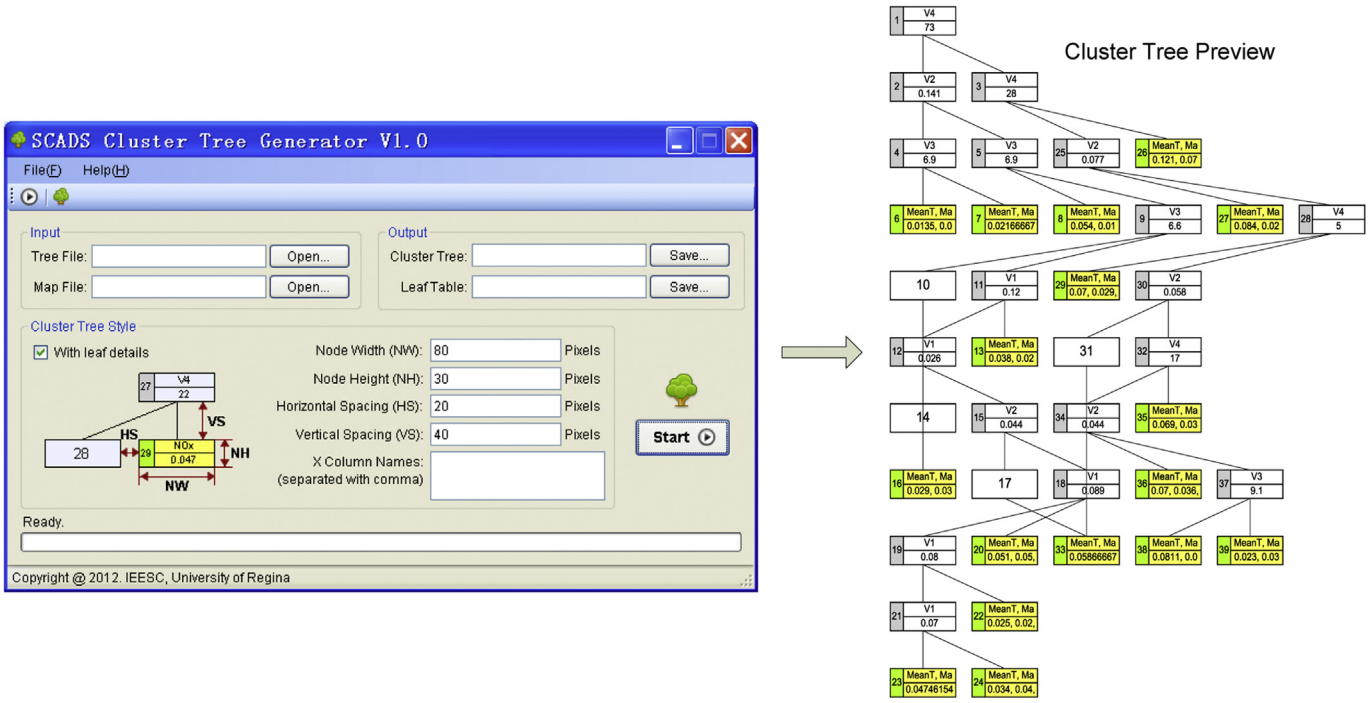


Fig. 5. Main interface of SCADS CTG V1.0.

V1.0), is developed together with SCADS to facilitate visualization of the cluster tree outputted by the training process. The main interface of CTG V1.0 is shown in Fig. 5.

3.4. Calibration

In the process of calibration, various parts of the model, including the input values, are changed so that the measured values (i.e. observations) are matched by the simulated values, with the purpose of accurately representing significant aspects of the actual system (Hill, 1998). In the SCADS, there is only one input parameter to be calibrated – significance level (α). Generally, its initial value may range from 0.01 to 0.05. The higher the significance level, the less sensitive the model will be, and the fewer leaf nodes for the resulting cluster tree. The default significance level in the SCADS is 0.05. However, users can adjust it repeatedly until the downscaled outputs match the sample predictands with an acceptable level of accuracy.

3.5. Validation

Validation is used to determine whether a model is an accurate representation of the real system, which is usually achieved through an iterative process of comparing the model to actual system behavior (Kleijnen, 1995; Bennett et al., 2013). The use of R-squared (R^2, the coefficient of determination) is well established in classical regression analysis (Rao, 1973). Its definition as the proportion of variance explained by the regression model makes it useful as a measure of success of predicting the dependent variable from the independent variables (Nagelkerke, 1991). Therefore, the SCADS employed the coefficient of determination as a key criterion to validate the model performance. As shown in Fig. 6, a number of *R*-squared values are calculated for each pair of predicted and observed predictands. By clicking on each value, a top-up window including the scatter plot of prediction versus observation will be displayed to help understand the model's performance visually.

3.6. Prediction

The prediction module is utilized to assistant users in developing high-resolution downscaled scenarios based on the validated cluster tree. The process of creating a prediction job is straightforward. The users specify the training job name as well as input data for predictor variables. In return, the SCADS predicts the corresponding predictands, according to the output tree and map files from the training process.

4. Application of SCADS

The SCADS was applied to the City of Toronto, Canada. Large-scale predictor variables for the period of 1981–2000 were derived from North American Regional Reanalysis Dataset (NARR), which were originally produced at the National Centers for Environmental Prediction (NCEP). The NARR project was an extension of the NCEP Global Reanalysis over the North America. The NARR model used the very high resolution NCEP Eta Model (32 km/45 layer) together with the Regional Data Assimilation System (RDAS) which, significantly, assimilated precipitation along with other variables (Mesinger and Coauthors, 2006; Saha and Coauthors, 2010). The collected data were then re-gridded to the 25 km grids of the PRECIS model for prediction purpose. Regional-scale predictand variables such as daily mean temperature (°C) and monthly precipitation (mm) for the same period were extracted from the 10 km gridded climate dataset as provided by the National Land and Water Information Service, Agriculture and Agri-Food, Canada. The gridded data were interpolated from daily Environment Canada climate station observations through a thin plate smoothing spline surface fitting method as implemented by ANUSPLIN V4.3 (NLWIS, 2008). The first ten-year data (i.e. 1981–1990) were used for model calibration, and the remaining ten-year ones (i.e. 1991–2000) for model validation.

Secondly, a limited set of large-scale predictor variables were screened out from a large suite of candidate variables with the aid of correlation analysis of SCADS. Table 1 lists correlation coefficients of candidate predictors versus daily mean temperature

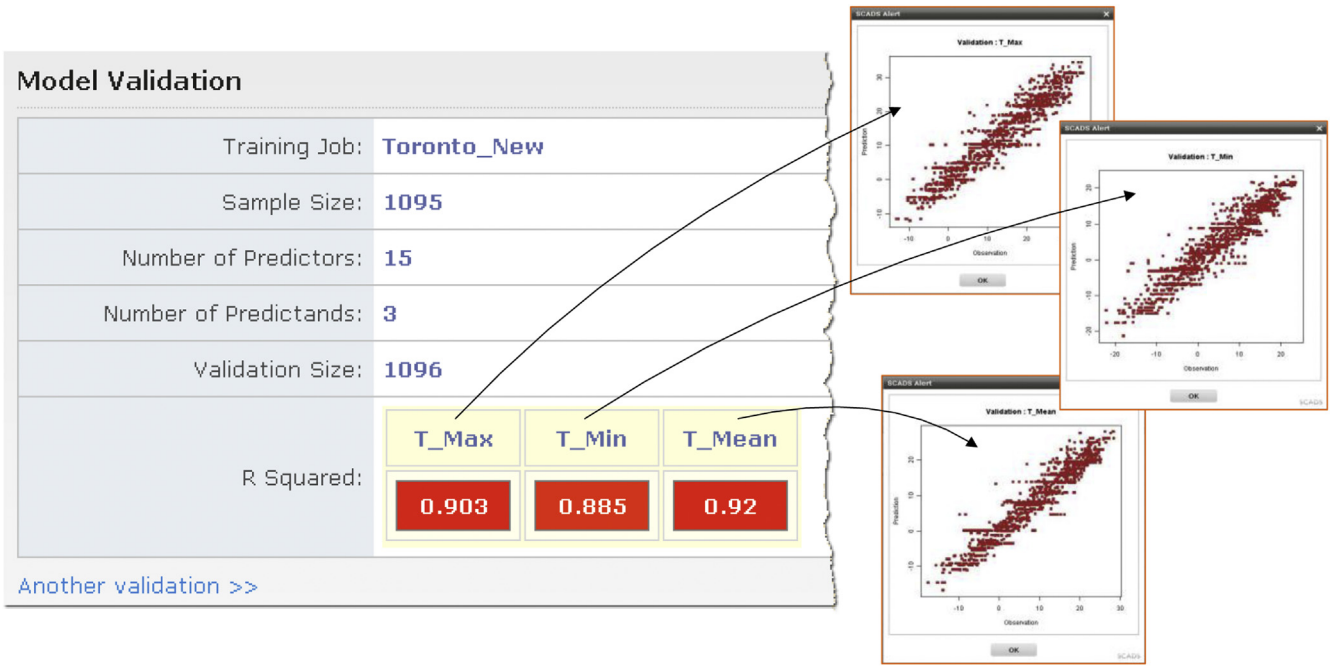


Fig. 6. Illustration of validation module.

and monthly precipitation. The bold values, denoting predictor variables selected for the training and prediction processes, indicate the most promising combinations of predictor variables and the corresponding predictands.

The screened sets of predictors were inputted to the training process to generate cluster trees at different significance levels (i.e. 0.01, 0.02, 0.03, 0.04, and 0.05). Results of daily mean temperature and monthly precipitation for the period of 1981–1990 were reproduced based on the generated cluster trees. Table 2 shows the calibration results at different significance levels with consideration of *R*-squared and root mean square errors (RMSE). It indicates that there is a slight improvement in estimating monthly precipitation while α value increases from 0.01 to 0.05. The best fitting results for precipitation could be obtained when $\alpha = 0.04$, with *R*-squared being 0.5279 and RMSE 16.1787 mm. However, choosing different α levels would not significantly affect the fitting results of daily mean temperature (with R-squared being a constant value of 0.9710), except a very slight improvement in RMSE when $\alpha = 0.05$. Therefore, the calibrated significance levels for temperature and precipitation would be different from each other in this example; the projected value for daily mean temperature would be 0.05, while that for monthly precipitation be 0.04.

To validate the performance of SCADS in hindcasting recent climate, values of daily mean temperature and monthly precipitation in the period of 1991–2000 were reproduced through calibrated cluster trees. The SCADS outputs were compared with the observed data retrieved from Environment Canada. Fig. 7 shows the validation results for daily mean temperature during the validation period of 1991–2000. The high *R*-squared value of 0.9705 indicates the SCADS can better reproduce the observed daily temperature, with RSME being as low as 1.6689 °C for the 10-year period. Fig. 8 shows the validation result for monthly precipitation. The overall performance of SCADS was satisfactory in hindcasting the monthly total precipitation for the validation period (with a *R*-squared value of 0.5156 and a RSME of 16.8004 mm), revealing the capability of SCADA in estimating extreme precipitation values which are usually related to extreme weather events such as floods in summer and snow storms in winter.

Table 1
Correlation coefficients of candidate predictors at Toronto, 1981–1990.

Predictor	Predictand		Predictor	Predictand	
	Tmean (°C)	Precip (mm)		Tmean (°C)	Precip (mm)
HGT500	**0.856**	0.07	PRESsfc	−0.231	−0.188
TMP500	**0.868**	0.135	TMPsfc	**0.971**	0.049
SPFH500	−0.166	−0.047	TMP2m	**0.978**	0.064
VVEL500	−0.039	**−0.286**	SPFH2m	**0.916**	0.146
UGRD500	−0.311	0.012	RH2m	−0.278	**0.286**
VGRD500	0.043	**0.255**	PRES2m	−0.237	−0.19
HGT700	0.796	0.02	UGRD10m	−0.228	−0.196
TMP700	0.873	0.128	VGRD10m	0.09	−0.005
SPFH700	−0.089	−0.026	TMP10m	**0.978**	0.063
VVEL700	0.045	**−0.243**	PRES10m	−0.232	−0.189
UGRD700	−0.224	0.033	SPFH10m	**0.916**	0.146
VGRD700	0.085	**0.276**	UGRD30m	−0.218	−0.197
HGT850	0.584	−0.075	VGRD30m	0.084	−0.005
TMP850	0.92	0.134	TMP30m	**0.979**	0.062
SPFH850	−0.029	−0.005	PRES30m	−0.221	−0.189
VVEL850	−0.115	**−0.239**	SPFH30m	**0.916**	0.146
UGRD850	−0.125	0.012	APCPsfc3h	0.095	**0.447**
VGRD850	0.12	**0.268**	VORT500	0.028	0
HGT1000	−0.192	−0.189	VORT700	−0.067	−0.028
TMP1000	**0.979**	0.07	VORT850	−0.075	−0.081
SPFH1000	**0.911**	0.151	VORT1000	−0.156	**−0.225**
VVEL1000	−0.097	−0.095	VORT10m	−0.135	−0.128
UGRD1000	−0.235	**−0.202**	VORT30m	−0.174	−0.175
VGRD1000	0.095	−0.028			

Bold signifies predictors that were selected in the modeling process.

Table 2
Calibration results at different significance levels, 1981–1990.

Significance level (α)	Tmean		Precip	
	R^2	RMSE (°C)	R^2	RMSE (mm)
0.01	0.9710	1.6916	0.4775	16.8632
0.02	0.9710	1.6912	0.5071	16.4723
0.03	0.9710	1.6913	0.5090	16.3432
0.04	0.9710	1.6913	**0.5279**	**16.1787**
0.05	**0.9710**	**1.6908**	0.5117	16.5368

Bold represents significance levels that were chosen in the modeling process.

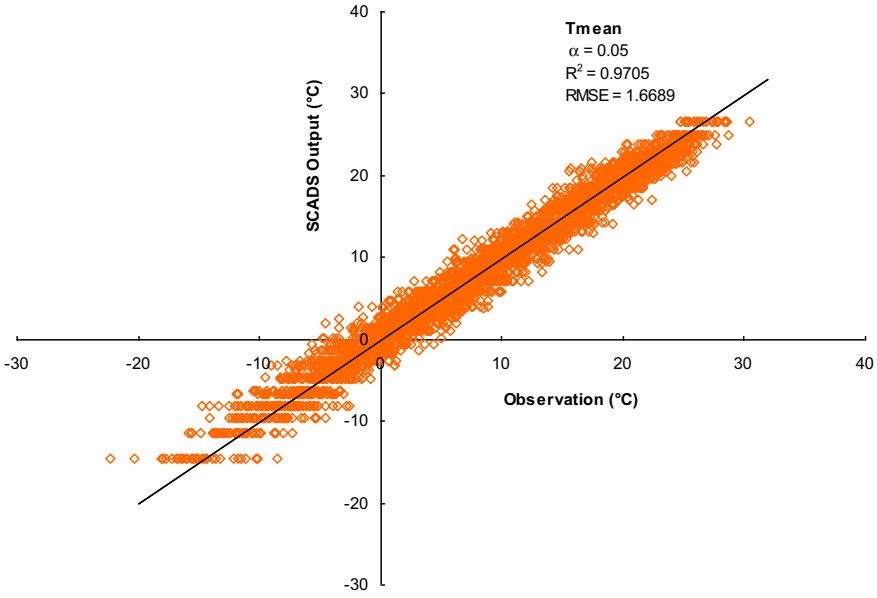


Fig. 7. Validation for daily mean temperature at Toronto, 1991–2000.

Containing the complex relationships between predictors and daily mean temperature and monthly precipitation, the validated cluster trees were next used to downscale equivalent regional predictor variables from the PRECIS model. The results of PRECIS model for Toronto were provided by the Center for Studies in Energy and Environment from the University of Regina (CSEE, 2010). Two 30-year time-slices were considered in this illustration to project the future climate scenarios of 1961–1990 and 2070–2099. The former period is generally defined as a baseline indicating the current climatic forcing, while the latter is indicative of future climate. Changes of future climate relative to the baseline period were then analyzed to help understand plausible future variations in daily mean temperature and monthly precipitation.

Fig. 9 shows changes in monthly mean temperature at Toronto for the period of 2070–2099 relative to the mean values in the baseline period. It reports a consistent increasing trend of mean temperature in all months, with an average change around +4 °C. Warming phenomena in January, February and March are considerably apparent with the change as high as +5 to +6 °C, while the remaining months shows relatively low increases (equal to or less than +4.5 °C). Overall, the projected warming trends would substantially lift the annual mean temperature to a great extent in 2070–2099. For example, the winter average temperature (i.e. December, January, and February) would be above 0 °C, while the summer one (i.e. June, July, and August) would be as high as 23 °C.

Fig. 10 shows changes in monthly total precipitation levels in Toronto for the period of 2070–2099 relative to those for the baseline period. There is a large variation in the changes of total precipitation from January to December. The total precipitation levels in winter (i.e. December, January, and February) and spring (i.e. March, April, and May) show significant increases within the range of +26 to +46 mm, except that in May (only +5 mm). However, the monthly precipitation totals in July–September are

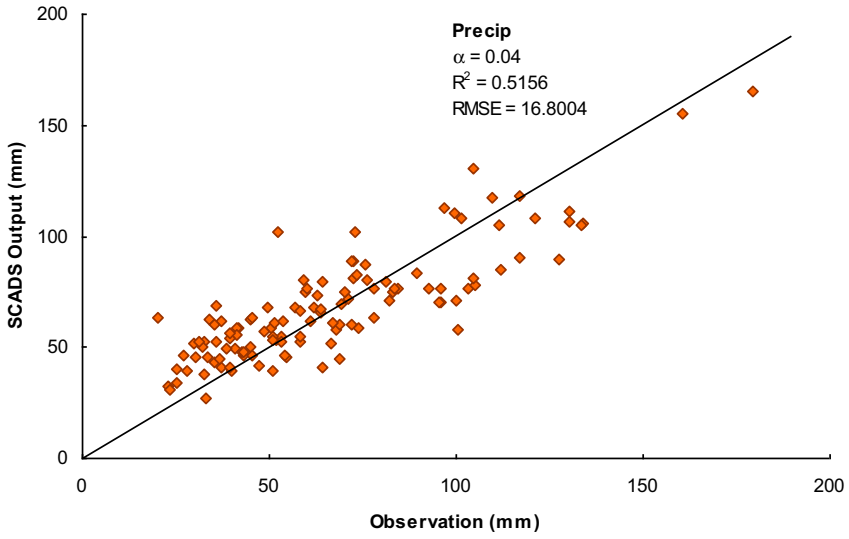


Fig. 8. Validation for monthly precipitation at Toronto, 1991–2000.

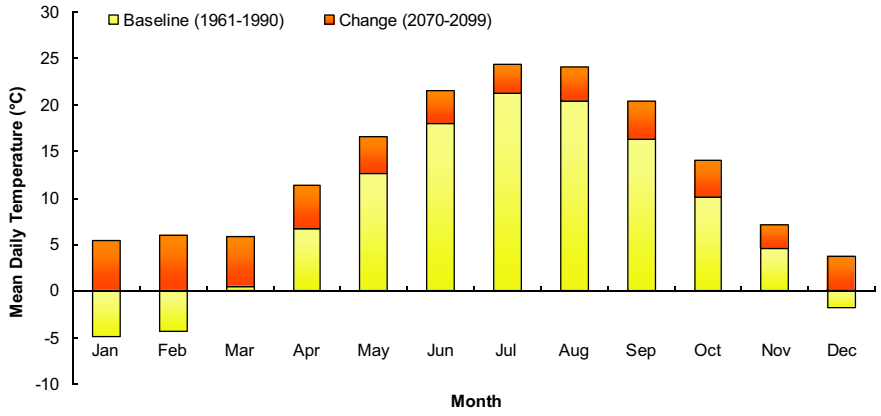


Fig. 9. Changes in monthly mean temperature at Toronto between 2070–2099 and 1961–1990.

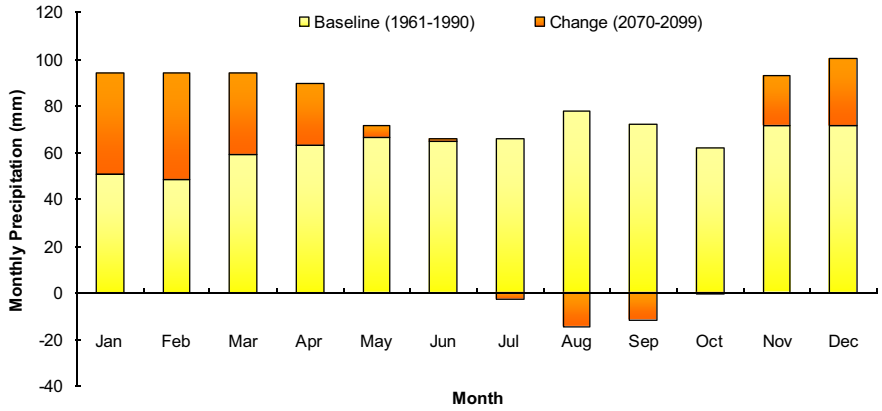


Fig. 10. Changes in monthly total precipitation at Toronto between 2070–2099 and 1961–1990.

projected to decrease by −3 to −15 mm while no obvious changes are observed in June and October. In general, the annual precipitation in Toronto would increase since the magnitudes of expected increases are much higher than those of projected decreases in terms of monthly total precipitation.

5. Caveats

The developed downscaling tool depends on a number of assumptions that may impose severe caveats on its future applications. First, the SCADS is still a statistical downscaling tool. In other words, all fundamental assumptions of statistical downscaling still hold for the SCADS. For example, the basic assumption is that the statistical relationships developed for the present climate also hold for the future – a limitation that also applies to dynamical models (Wilby et al., 2004); the predictor set is supposed to sufficiently represent future climate change signals, which requires that the predictors be screened based on their relevance to the target predictands as well as their accurate representation by climate models (Wilby and Wigley, 2000; Giorgi et al., 2001). Secondly, the SCADS assumes that the local variables are normally distributed so that the cutting and merging process can be effectively handled based on the Wilk's statistic. Therefore, poor results would be obtained for daily precipitation because most days in a year might have no precipitation (i.e. 0 mm) which leads to a gamma distribution. To deal with this weakness, further improvements regarding downscaling daily precipitation could be carried out by introducing precipitation occurrence model and precipitation amount model proposed by Fealy and Sweeney (2007). Thirdly, the cluster tree is trained and calibrated based on the previous climate of the target region. It means that the future projections of target predictands will not lie outside the range of the previous climatology; consequently, new extreme values cannot be captured. In addition, future efforts will also focus on improving the software's capability of projecting extremes as caused by local non-stationary process in the context of climate change, such as hydro-meteorological extremes (Khaliq et al., 2006).

6. Conclusions

A statistical downscaling tool (SCADS) has been developed to assist obtaining high-resolution climate change scenarios, based on the stepwise cluster analysis method. The SCADS uses a cluster tree to represent the complex relationship between large-scale atmospheric variables and local surface variables. It can effectively deal with continuous and discrete variables, as well as nonlinear relations between predictors and predictands. By integrating ancillary functional modules of missing data detecting, correlation analysis, model calibration and graphing of cluster trees, the SCADS can perform rapid development of downscaling scenarios for local weather variables under current and future climate forcing. SCADS was used to generate 10 km daily mean temperature and monthly precipitation projections for Toronto, Canada. Two cluster trees were built based on the observed data of 1981–1990, and were then used to reproduce the historical climatology of 1991–2000 for validation purpose. The results showed that the observed temperature and precipitation in the validation period were well hindcasted by SCADS. The validated models were then applied to obtain

temperature and precipitation projections for the period of 2070–2099.

Acknowledgments

This research was supported by the Major Project Program of the Natural Sciences Foundation (51190095), the Program for Innovative Research Team in University (IRT1127), Environment Canada and the Natural Science and Engineering Research Council of Canada.

References

Beckmann, B.R., Adri Buishand, T., 2002. Statistical downscaling relationships for precipitation in the Netherlands and North Germany. Int. J. Climatol. 22, 15–32.
Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.
Cooley, W.W., Lohnes, P.R., 1971. Multivariate Data Analysis. John Wiley & Sons, Inc.
CSEE, 2010. Regional Climate Modelling over Ontario Using UK PRECIS. Center for Studies in Energy and Environment, University of Regina.
Fealy, R., Sweeney, J., 2007. Statistical downscaling of precipitation for a selection of sites in Ireland employing a generalised linear modelling approach. Int. J. Climatol. 27, 2083–2094.
Fowler, H.J., Blenkinsop, S., Tebaldi, C., 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int. J. Climatol. 27, 1547–1578.
Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
Giorgi, F., Christensen, J., Hulme, M., Von Storch, H., Whetton, P., Jones, R., Mearns, L., Fu, C., Arritt, R., Bates, B., 2001. Regional climate information – evaluation and projections. In: Houghton, J.T., et al. (Eds.), Climate Change 2001: the Scientific Basis Contribution of Working Group to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, US.
Hashmi, M.Z., Shamseldin, A.Y., Melville, B.W., 2009. Statistical downscaling of precipitation: state-of-the-art and application of Bayesian multi-model approach for uncertainty assessment. Hydrol. Earth Syst. Sci. Discuss. 6, 6535–6579.
Hashmi, M.Z., Shamseldin, A.Y., Melville, B.W., 2011. Statistical downscaling of watershed precipitation using Gene Expression Programming (GEP). Environ. Model. Softw. 26, 1639–1646.
Hessami, M., Gachon, P., Ouarda, T.B.M.J., St-Hilaire, A., 2008. Automated regression-based statistical downscaling tool. Environ. Model. Softw. 23, 813–834.
Hewitson, B., Crane, R., 1996. Climate downscaling: techniques and application. Clim. Res. 7, 85–95.
Heyen, H., Zorita, E., Von Storch, H., 1996. Statistical downscaling of monthly mean North Atlantic air-pressure to sea level anomalies in the Baltic Sea. Tellus A 48, 312–323.
Hill, M.C., 1998. Water-resources Investigation Report: Methods and Guidelines for Effective Model Calibration.
Huang, G., 1992. A stepwise cluster analysis method for predicting air quality in an urban environment. Atmos. Environ. B Urban Atmos. 26, 349–357.
Huang, G.H., Huang, Y.F., Wang, G.Q., Xiao, H.N., 2006. Development of a forecasting system for supporting remediation design and process control based on NAPL-biodegradation simulation and stepwise-cluster analysis. Water Resour. Res. 42.
Huth, R., 2002. Statistical downscaling of daily temperature in central Europe. J. Clim. 15, 1731–1742.
Khaliq, M.N., Ouarda, T.B.M.J., Ondo, J.C., Gachon, P., Bobée, B., 2006. Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: a review. J. Hydrol. 329, 534–552.
Kleijnen, J.P.C., 1995. Verification and validation of simulation models. Eur. J. Oper. Res. 82, 145–162.
Maak, K., von Storch, H., 1997. Statistical downscaling of monthly mean air temperature to the beginning of flowering of *Galanthus nivalis* L. in Northern Germany. Int. J. Biometeorol. 41, 5–12.
Mearns, L., Giorgi, F., Whetton, P., Pabon, D., Hulme, M., Lal, M., 2003. Guidelines for Use of Climate Scenarios Developed from Regional Climate Model Experiments. Data Distribution Centre of the Intergovernmental Panel on Climate Change.
Mesinger, F., Coauthors, 2006. North American regional reanalysis. Bull. Am. Meteor. Soc. 87, 343–360.
Morrison, D.F., 1967. Multivariate Statistical Methods. McGraw-Hill, New York.
Mullan, D., Fealy, R., Favis-Mortlock, D., 2012. Developing site-specific future temperature scenarios for Northern Ireland: addressing key issues employing a statistical downscaling approach. Int. J. Climatol. 32, 2007–2019.
Murphy, J., 1999. An evaluation of statistical and dynamical techniques for downscaling local climate. J. Clim. 12, 2256–2284.
Nagelkerke, N.J.D., 1991. A note on a general definition of the coefficient of determination. Biometrika 78, 691–692.
NLWIS, 2008. Daily Maximum Termperatures (Celsius) for Canada. National Land and Water Information Service, Agriculture and Agri-Food, Canada.
Overall, J.E., Klett, C.J., 1972. Applied Multivariate Analysis. McGraw-Hill, New York.
Phatak, A., Bates, B.C., Charles, S.P., 2011. Statistical downscaling of rainfall data using sparse variable selection methods. Environ. Model. Softw. 26, 1363–1371.
Qin, X., Huang, G., Chakma, A., 2007. A stepwise-inference-based optimization system for supporting remediation of petroleum-contaminated sites. Water Air Soil Pollut. 185, 349–368.
Qin, X.S., Huang, G.H., Zeng, G.M., Chakma, A., 2008. Simulation-based optimization of dual-phase vacuum extraction to remove nonaqueous phase liquids in subsurface. Water Resour. Res. 44, W04422.
Rao, C.R., 1952. Advanced Statistical Methods in Biometric Research. Wiley, New York.
Rao, C.R., 1973. Linear Statistical Inference and its Application, second ed. Wiley, New York.
Rodgers, J.L., Nicewander, W.A., 1988. Thirteen ways to look at the correlation coefficient. Am. Stat. 42, 59–66.
Saha, S., Coauthors, 2010. The NCEP climate forecast system reanalysis. Bull. Am. Meteor. Soc. 91, 1015–1057.
Semenov, M.A., Barrow, E.M., 1997. Use of a stochastic weather generator in the development of climate change scenarios. Clim. Change 35, 397–414.
Timbal, B., Fernandez, E., Li, Z., 2009. Generalization of a statistical downscaling model to provide local climate change projections for Australia. Environ. Model. Softw. 24, 341–358.
Wilby, R., Charles, S., Zorita, E., Timbal, B., Whetton, P., Mearns, L., 2004. Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods.
Wilby, R., Wigley, T., 2000. Precipitation predictors for downscaling: observed and general circulation model relationships. Int. J. Climatol. 20, 641–661.
Wilby, R.L., Dawson, C.W., Barrow, E.M., 2002. SDSM—a decision support tool for the assessment of regional climate change impacts. Environ. Model. Softw. 17, 145–157.
Wilby, R.L., Wigley, T., 1997. Downscaling general circulation model output: a review of methods and limitations. Prog. Phys. Geogr. 21, 530–548.
Wilby, R.L., Wigley, T.M.L., Conway, D., Jones, P.D., Hewitson, B.C., Main, J., Wilks, D.S., 1998. Statistical downscaling of general circulation model output: a comparison of methods. Water Resour. Res. 34, 2995–3008.
Wilks, S.S., 1962. Mathematical Statistics. Wiley, New York.
Willems, P., Vrac, M., 2011. Statistical precipitation downscaling for small-scale hydrological impact investigations of climate change. J. Hydrol. 402, 193–205.
Wood, A.W., Leung, L.R., Sridhar, V., Lettenmaier, D., 2004. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Clim. Change 62, 189–216.