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.