Journal of Hydrology 525 (2015) 286–301

Contents lists available at ScienceDirect

Journal of Hydrology

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

Dynamic coupling of support vector machine and K-nearest neighbour for downscaling daily rainfall

Manjula Devak, C.T. Dhanya *, A.K. Gosain

Department of Civil Engineering, Indian Institute of Technology Delhi, New Delhi 110016, India

ARTICLE INFO

Article history:
Received 30 October 2014
Received in revised form 7 January 2015
Accepted 24 March 2015
Available online 2 April 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Fi-John Chang, Associate Editor

Keywords:
Climate change impact
Statistical downscaling
Uncertainty
Ensemble approach

SUMMARY

Climate change impact assessment studies in water resources section demand the simulations of climatic variables at coarser scales from dynamic General Circulation Models (GCMs) to be mapped to even finer scales. Related studies in this area have mostly been relying on statistical techniques for downscaling variables to finer resolution. This demands a careful selection of a suitable downscaling model, to alleviate the downscaling uncertainty. In this study, it is proposed to develop a dynamic framework for downscaling purpose by integrating the frequently used techniques, K-Nearest Neighbour (KNN) and Support Vector Machine (SVM). In order to give flexibility in future predictors–predictand relationships and to account the sensitivity in model parameters, it is also proposed to generate an ensemble of outputs by identifying various plausible model parameter combinations. The performance of this framework for downscaling daily precipitation values at different locations is compared with simple KNN and SVM models. The proposed hybrid model is found to be better in capturing various characteristics of daily precipitation than individual models, especially in simulating the extremes, both in magnitude and duration. The mean ensemble is found to be efficient than single best simulation with optimum parameter combinations. The efficacy of hybrid SVM–KNN ensemble downscaling model is established through detailed investigations. The future downscaled projection for mid-century and late century employing this hybrid model indicates an increased variability in future precipitation, though the intensity varies for various locations. The developed methodology hence ensures lesser downscaling uncertainty and also eliminates the inherent assumption of relationship stationarity considered in many downscaling models.

1. Introduction

Climate change and its impact on water resources has been an intriguing topic to hydrologic research community since a few past decades. Global Climate Models (GCMs), principal tools and sole means to identify and assess the climate change, are evolving gradually but steadily, to a more complex formulation and relatively realistic representation of global climate system. Development of more comprehensive coupled models and availability of more number of runs have undoubtedly aided in obtaining a more vivid representation of future climate. The origin, history and evolution of climate modelling have been addressed in great detail by various literatures (Edwards, 2011; IPCC, 2013; IPCC, 2007a; Randall, 2010; Randall, 2000; Stute et al., 2001). Even so, with the availability of many more GCMs and their accountability of complex feedbacks in earth's climate, a clear consensus on the extent of global warming and climate change is still lacking. Efforts by PCMDI (Program for Climate Model Diagnosis and Inter-comparison) and IPCC (Intergovernmental Panel on Climate Change) through various coupled model inter-comparison projects like CMIP3 and CMIP5 (Taylor et al., 2012), provide improved future projections by evaluation and diagnosis of various climate models. Nevertheless, errors and uncertainty in model simulations remain substantial (IPCC, 2013). The applicability of these climate projections in any water resources management decision making studies, hence, ultimately relies to a great extent on research aiming to reduce these uncertainties.

Three distinct sources contributing to uncertainties in future projections are model uncertainty, scenario uncertainty, and internal variability. IPCC's Fifth Assessment Report (AR5) has evaluated projections from around 40 Atmosphere–Ocean General Circulation Models (AOGCMs), four future scenarios (Representative Concentration Pathways, RCPs) and a few number of ensemble runs for each model (Flato et al., 2013). Studies of climate change impact assessment of water resources, be it on the changes in surface water and ground water, or on water quality or on extreme events etc., are mostly focussed at a regional scale

* Corresponding author. Tel.: +91 11 26597328; fax: +91 11 26581117.
E-mail addresses: manjuladevak26@gmail.com (M. Devak), dhanya@civil.iitd.ac.in (C.T. Dhanya), gosain@civil.iitd.ac.in (A.K. Gosain).

http://dx.doi.org/10.1016/j.jhydrol.2015.03.051
0022-1694/

or at a basin scale. Since the global climate model projections are generally at a coarser resolution, apart from the three major uncertainty sources mentioned before, climate change impact assessments in water resources need to confront another main source of uncertainty i.e., downscaling uncertainty. Downscaling techniques to obtain finer grid data from coarser grid data are of two types, viz., dynamic downscaling and statistical downscaling. While dynamic downscaling involves numerical meteorological modelling to reflect how the global patterns affect the local weather conditions (Castro, 2005; Kunstmann et al., 2004), statistical downscaling, on the other hand, involves setting up statistical relationship between global scale variables and local scale hydrologic variables (Frías et al., 2006; Schoof et al., 2007). Various studies in the past had compared and analysed the efficacy of statistical downscaling over dynamical downscaling (Gutmann et al., 2012; Diez et al., 2005; Yoon et al., 2012; Kidson and Thompson, 1998; Murphy, 1999; Wilby et al., 2004). It is clear that, even though dynamic downscaling ensures an accurate physical interpretation of the regional climate system, due to the complexity in modelling and computations, many studies assessing climate change impact on water resources have relied on statistical downscaling techniques due to its simplicity and ease in implementation. Nevertheless, there is still ambiguity in the selection of proper statistical downscaling method; whilst choosing a befitting downscaling method will reduce the downscaling uncertainty in climate change studies.

Statistical downscaling techniques are categorised into three methods viz. weather typing (e.g. analogue method, hybrid approaches, fuzzy classification, self-organising maps, Monte Carlo methods), weather generators (e.g. Markov chains, stochastic models, spell length methods, storm arrival times, mixture modelling) and regression based methods (e.g. Linear regression, neural networks, canonical correlation analysis, kriging). In the past, many researchers have compared various methods of statistical downscaling. Wilby and Wigley (1997) and Wilby et al. (1998) compared the performance of six statistical downscaling approaches (two neural networks, two weather generators and two regression based methods) and have concluded that, despite using the common sets of GCM predictors, the methods used in the study generate radically different future projections and also added that regression based methods give better estimates. Zorita and von Storch (1999) demonstrated that analogue method is efficient as well as simpler to implement when compared to linear regression method, classification methods and neural networks. Sunyer et al. (2012) compared five statistical downscaling methods, including two statistical correction methods (change in mean and change in mean and variance methods) which belong to the group of regression models and three weather generators (WG) and highlighted the importance of acknowledging the limitations and advantages of different statistical downscaling methods as well as the uncertainties in downscaling climate change projections for use in hydrological models. Duan and Mei (2013) compared three frequently applied statistical downscaling tools such as Statistical Downscaling Model (SDSM), Support Vector Machine (SVM), and Long Ashton Research Station Weather Generator (LARS-WG) and demonstrated the usefulness and weakness of different downscaling methods in simulating various precipitation characteristics under different circumstances. Whilst, the aforementioned comparison studies proved the efficiency of one method over the other, it is also to be noted that none of the downscaling technique can assure an accurate estimate of precipitation under different situations. This establishes the need for a hybrid downscaling technique which can enhance/nullify the advantages/disadvantages of downscaling approaches. Hence, in this study we propose to integrate two popular downscaling approaches, Support Vector Machine (SVM) (Cortes and Vapnik, 1995), a transfer function method and K-Nearest Neighbour (KNN), an analogue approach.

Despite many advantages of both these methods, various studies have pointed out noteworthy disadvantages of these methods. Burges (1998) applied SVM in pattern recognition and showed that although SVM model captures non-linear regression relationships between the predictors and predictands, the biggest limitation lies in the choice of kernel. SVM provide a promising alternative to conventional artificial neural networks for statistical downscaling and are suitable for performing climate change impact studies. However, it is also reported that SVM is unable to capture extreme rainfall events (Anandhi et al., 2008; Tripathi et al., 2006). The reason for the same is that as SVM is a regression based model, it generally does not simulate entire variance of the downscaled variable. SVM also faces the problem of overtraining i.e., large difference between the performances with training and testing data, which can be overcome by appropriate selection of SVM parameters (Ghosh, 2010). SVM demands a laborious approach in determining its parameters (kernel width and penalty term). While SVM model overestimates the mean and underestimates the standard deviation and shows worse fit to the observed data because of over fitting (Raje and Mujumdar, 2011), KNN method underestimates the standard deviation and autocorrelation coefficient and is very much dependant to the number of nearest neighbours (Buishand and Brandsma, 2001; Lall and Sharma, 1996). Although, KNN approach is computationally efficient and can be easily implemented, it is unable to generate the events that are not present in the past record (Gangopadhyay et al., 2005). Hence, while SVM directly operates on kernel, it involves time consuming training on whole data set. And, while KNN is an easy and candid approach, its certainty in the prediction depends upon the number of nearest neighbours. Hybrid of SVM with KNN could couple the advantages of improved decision boundaries of SVM and smooth distance function of KNN (Zhang et al., 2006). In this study, SVM and KNN is integrated by selecting the neighbours which are near to the query sample and then training the SVM model for each time step with the corresponding nearest neighbours only. The calibration and validation of SVM–KNN model at each time step provides a dynamic framework downscaling approach, unlike the simple SVM approach where the parameters are fixed a priori. The hybrid also nullifies the major limitation of simple KNN approach i.e., restricting to the arithmetic mean or weighted average of the neighbours identified from the sample data; though KNN approach offers a dynamic framework otherwise. In addition, since calibration is done at each time step separately, it is clear that SVM–KNN hybrid do not assume any stationarity in relationships between predictand and predictors or fixed parameter values, which is the main disadvantage with most of the statistical downscaling models.

In addition to the facts discussed above, since the outputs from any downscaling model depends on the set of calibrated model parameters, it is better to adopt en ensemble approach with a set of plausible model parameters, which will capture the uncertainties in the parameterization, so as to account for all possible relationships. The aim of this paper is to integrate the popular statistical downscaling approaches, SVM and KNN in order to enhance the downscaling ability and hence reduce the downscaling uncertainty in climate change impact studies. An ensemble of future projections is obtained for a set of plausible model parameters to capture the modelling uncertainty. The various methods employed and data used are described first, followed by delineating the methodology used in this study. The methodology developed is demonstrated on Mahanadi basin and nearby region and finally the results are discussed. The results from hybrid SVM–KNN are then compared with both SVM and KNN models.