Atmospheric Research 172–173 (2016) 37–47

Contents lists available at ScienceDirect

Atmospheric Research

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

Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction

A. Belayneh, J. Adamowski *, B. Khalil, J. Quilty

[a] Department of Bioresource Engineering, Faculty of Agricultural and Environmental Sciences, McGill University, 21 111 Lakeshore, Ste Anne de Bellevue, Quebec H9X 3V9, Canada.

ARTICLE INFO

Article history:
Received 19 August 2015
Received in revised form 19 November 2015
Accepted 30 December 2015
Available online 7 January 2016

Keywords:
Support vector regression
SPI
Drought forecasting
Wavelet transforms
Bootstrap
Boosting

ABSTRACT

This study explored the ability of coupled machine learning models and ensemble techniques to predict drought conditions in the Awash River Basin of Ethiopia. The potential of wavelet transforms coupled with the bootstrap and boosting ensemble techniques to develop reliable artificial neural network (ANN) and support vector regression (SVR) models was explored in this study for drought prediction. Wavelet analysis was used as a pre-processing tool and was shown to improve drought predictions. The Standardized Precipitation Index (SPI) (in this case SPI 3, SPI 12 and SPI 24) is a meteorological drought index that was forecasted using the aforementioned models and these SPI values represent short and long-term drought conditions. The performances of all models were compared using RMSE, MAE, and R^2. The prediction results indicated that the use of the boosting ensemble technique consistently improved the correlation between observed and predicted SPIs. In addition, the use of wavelet analysis improved the prediction results of all models. Overall, the wavelet boosting ANN (WBS-ANN) and wavelet boosting SVR (WBS-SVR) models provided better prediction results compared to the other model types evaluated.

1. Introduction

Meteorological droughts are defined as a deficit in precipitation compared to the long-term average, over a period of time (Belayneh and Adamowski, 2013). Globally, 22% of the economic damage caused by natural disasters and 33% of the damage in terms of the number of affected persons can be attributed to drought (Keshavarz et al., 2013). The impacts of drought are more severe in sub-Saharan Africa, where rain-fed agriculture comprises 95% of all agriculture in the region. Effective drought predictions can be a tool to help reduce and mitigate some of the impacts of drought.

One approach to hydrological prediction, including drought prediction, is the use of machine learning techniques such as artificial neural networks (ANN) and support vector regression (SVR) (Campisi et al., 2012; Tiwari and Adamowski, 2014; Tiwari and Adamowski, 2015; Rathinasamy et al., 2015; Nourani et al., 2014; Belayneh et al., 2014; Karran et al., 2014; Rathinasamy et al., 2014; Adamowski et al., 2012a, 2012b). ANNs have been used in several studies as drought prediction tools (Mishra and Desai, 2006; Morid et al., 2007; Bacanli et al., 2008; Barros and Bowden, 2008; Cutore et al., 2009; Karamouz et al., 2009; Marj and Meijerink, 2011; Mishra and Nagarajan, 2012; Belayneh and Adamowski, 2012, 2013). There are also a number of studies where SVR was used for hydrological predictions. Khan and Coulibaly (2006) found that an SVR model performed better than ANNs in 3–12 month predictions of lake water levels. Kisi and Cimen (2009) used SVRs to estimate daily evaporation. More recently, SVR models have begun to be explored for drought prediction purposes. Belayneh et al. (2014), Belayneh and Adamowski (2013), and Belayneh and Adamowski (2012) explored the use of SVR models to predict the SPI in the Awash River Basin in Ethiopia.

In addition to the use of these machine learning techniques, researchers have started to couple wavelet transforms with the aforementioned model types. In coupled models, wavelet transforms are used as a pre-processing tool for the data before inputting the data into the data driven models. For example, coupled wavelet-ANN (WANN) models have been recently used in drought prediction studies (Kim and Valdes, 2003; Ozger et al., 2012; Mishra and Singh, 2012; Belayneh and Adamowski, 2012; Belayneh and Adamowski, 2013; Belayneh et al., 2014). Even more recently, coupled wavelet-SVR models have been explored for predicting drought (Belayneh et al., 2014). These studies have found that the most accurate method is the WANN approach (for both short and long term predictions). However, it should be noted that no studies have been completed to date that have explored the use of the bootstrap or boosting ensemble approaches in either short or long term drought prediction. This issue was explored in this study.

There are several strong theoretical and practical justifications to pursue an ensemble approach for drought prediction. An ensemble model may provide an efficient approach for certain applications in

* Corrresponding author. Tel.:+1 514 398 7786; fax: +1 514 398 8387.
E-mail address: jan.adamowski@mcgill.ca (J. Adamowski).

http://dx.doi.org/10.1016/j.atmosres.2015.12.017
0169-8095/

which the number of data is too large for building a single model (Helmy et al., 2013). Ensemble models have also been shown to be effective in the absence of adequate training data by building different models using re-sampling techniques (Erdal and Karakurt, 2013).

In this study, the bootstrap and boosting ensemble approaches were combined with machine learning techniques to predict drought. Although the boosting algorithm has a better generalization ability than the bootstrap algorithm in a number of applications (e.g., Shu and Burn, 2004), the latter algorithm has the advantage of training the member networks in an ensemble independently, hence in parallel. In light of this, comparing these two approaches (i.e., bootstrap and boosting) was deemed to be useful to determine which approach is better at assessing uncertainty in drought predictions. This has not been explored to date in the drought prediction literature. In addition to creating bootstrap and boosting ensemble models, this study also coupled these ensemble models with wavelet transforms, which has also not been explored to date in the drought prediction literature. The wavelet transforms were used to pre-process the data before the use of the ensemble models.

The bootstrap-ANN (BANN) ensemble modeling approach has been the focus of several hydrological forecasting applications over the past decade. Shu and Bum (2004) used BANN ensembles to study flood frequency, and Tiwari and Chatterjee (2010) used BANN ensembles to forecast floods. Zaier et al. (2010) compared the effectiveness of BANN models with boosting-ANN (BS-ANN) ensembles for the estimation of ice thickness on lakes. Tiwari and Chatterjee (2011) coupled BANN models with wavelet transforms to forecast the uncertainty of flood forecasts. Li et al. (2010) used bootstrap SVR (BSVR) models for the purposes of streamflow prediction. Shu and Burn (2004) compared the ability of BANN and BS-ANN models in estimating the index flood and the 10-year flood quantile, and Erdal and Karakurt (2013) built boosting and bootstrap ensembles using a benchmark SVR model for the purposes of streamflow forecasting.

This study predicted the Standardized Precipitation Index (SPI) using a combination of the bootstrap and boosting techniques with ANN and SVR models. In addition, wavelet analysis was used to pre-process the SPI data series. The SPI was chosen because it is a standardized index that enables the comparison of drought on multiple time scales. This study forecasted SPI 3, SPI 12 and SPI 24, which are representative of short and long-term drought conditions. To the best knowledge of the authors, this study is the first to couple wavelet transforms and ANN and SVR models with bootstrap and boosting ensembles for the purposes of drought predictions.

2. Study area: The Awash River Basin, Ethiopia

In this study, the SPI was forecasted for the Awash River Basin in Ethiopia (Fig. 1a). Drought is a common occurrence in the basin (Edossa et al., 2010) and the heavy dependence of the population on rain-fed agriculture has made the people and the country's economy extremely vulnerable to the impacts of droughts. The mean annual rainfall in the basin varies from about 1600 mm in the highlands to 160 mm in the northern point of the basin. The total amount of rainfall also varies greatly from year to year, resulting in severe droughts in some years and flooding in others. The total annual surface runoff in the Awash Basin amounts to $4900 \times 106\ m^3$ (Edossa et al., 2010). Effective predictions of the SPI can be used to mitigate the impacts of drought that manifest as a result of rainfall shortages in the area. The climate of the Awash River Basin varies between a mild temperate climate in the Upper Awash sub-basin and a hot semi-arid climate in both the Middle and the Lower sub-basins.

The Upper Awash Basin (Fig. 1b) has a temperate climate with annual mean temperatures between 15 and 22 °C and an annual rainfall of between 500 and 2000 mm (Edossa et al., 2010). The Middle Awash Basin (Fig. 1c) has a semiarid climate with an annual precipitation of between 200 and 1500 mm (Edossa et al., 2010). The Lower Awash Basin (Fig. 1d) has a hot semi-arid climate with an annual precipitation between 200 and 700 mm and an average annual temperature between 22 and 32 °C. Rainfall records from 1970 to 2005 were used to generate SPI 3, SPI 12 and SPI 24 time series from each of the three stations used. The three stations chosen (Bati, Dubti and Debre Zeit) were from the lower, middle and upper Awash River sub-basins, respectively. Table 1 shows the precipitation statistics at each of the stations from 1970 to 2005. At each station, monthly precipitation records were used to generate the SPI time series.

3. Model development

The following sections describe the methodology used to generate the prediction models. A description of the Standardized Precipitation Index (SPI) and its computation is described below. The development of the ANN and SVR models is described, along with the theory and implementation of the bootstrap and boosting techniques. The selection of the appropriate wavelet transform is also described.

3.1. The Standardized Precipitation Index (SPI)

The Standardized Precipitation Index (SPI) was developed by McKee et al. (1993) and it is based on precipitation alone making its evaluation relatively easy compared to other drought indices, namely the Palmer Index and the Crop Moisture Index (Cacciamani et al., 2007). A major advantage of the SPI index is that it allows for the description of drought on multiple time scales (Tsakiris and Vangelis, 2004; Mishra and Desai, 2006). Given the fact that this study will explore forecasts of both short and long-term SPI values, this characteristic is especially useful. In another study, the SPI was also determined to be the best drought index for representing the variability in East African droughts (Ntale and Gan, 2003). Given the location of the Awash River Basin in East Africa, the choice of the SPI was deemed to be appropriate.

SPI values can be categorized according to classes (Mishra and Desai, 2006). Normal conditions are established from the aggregation of two classes: $-1 < SPI < 0$ (mild drought) and $0 \le SPI \le 1$ (slightly wet). SPI values are positive or negative for greater or less than mean precipitation, respectively. The variance from the mean precipitation is a probability indication of the severity of the flood or drought, which can be used for risk assessment (Morid et al., 2007). The more negative the SPI value for a given location, the more severe the drought. In this study, an SPI_SL_6 program developed by the National Drought Mitigation Center, University of Nebraska-Lincoln, was used to compute the time series of drought indices (SPI) for each station in the basin and for each month of the year at different time scales.

3.2. ANN models

Artificial neural networks (ANNs) are non-linear data driven models that can provide powerful solutions to many complex modeling problems. An ANN model is based on a connectionist approach to computation involving several transformation elements (neurons), interconnected and distributed over different layers (Cutore et al., 2009). Some of the characteristics of ANNs that make them attractive for hydrologic modeling are that they are able to recognize the relation between the input and the output variables without explicit considerations. ANNs work well even when the training sets contain noise and measurement errors (Mishra and Desai, 2006), and they are able to adapt to solutions over time to compensate for changing circumstances. ANNs were used in this study because they have parsimonious data requirements, they have the advantage of producing models without a complete understanding of catchment processes, and they have rapid execution times.

The ANN models used in this study have a feed forward multi-layer perceptron (MLP) architecture that was trained with the Levenberg–Marquardt (LM) back propagation algorithm. MLPs have often been