Journal of Hydrology 545 (2017) 100–108

Contents lists available at ScienceDirect

Journal of Hydrology

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

Technical Note

Choice of rainfall inputs for event-based rainfall-runoff modeling in a catchment with multiple rainfall stations using data-driven techniques

Tak Kwin Chang [a], Amin Talei [a,*], Sina Alaghmand [a], Melanie Po-Leen Ooi [b,a]

[a] School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia
[b] School of Engineering and Physical Sciences, Heriot-Watt University Malaysia, No. 1 Jalan Venna P5/2, Precinct 5, 62200 Putrajaya, Malaysia

ARTICLE INFO

Article history:
Received 22 June 2016
Received in revised form 1 November 2016
Accepted 12 December 2016
Available online 20 December 2016
This manuscript was handled by G. Syme, Editor-in-Chief, with the assistance of Bellie Sivakumar, Associate Editor

Keywords:
Input selection
Rainfall-runoff modeling
ANFIS
HEC-HMS

ABSTRACT

Input selection for data-driven rainfall-runoff models is an important task as these models find the relationship between rainfall and runoff by direct mapping of inputs to output. In this study, two different input selection methods were used: cross-correlation analysis (CCA), and a combination of mutual information and cross-correlation analyses (MICCA). Selected inputs were used to develop adaptive network-based fuzzy inference system (ANFIS) in Sungai Kayu Ara basin, Selangor, Malaysia. The study catchment has 10 rainfall stations and one discharge station located at the outlet of the catchment. A total of 24 rainfall-runoff events (10-min interval) from 1996 to 2004 were selected from which 18 events were used for training and the remaining 6 were reserved for validating (testing) the models. The results of ANFIS models then were compared against the ones obtained by conceptual model HEC-HMS. The CCA and MICCA methods selected the rainfall inputs only from 2 (stations 1 and 5) and 3 (stations 1, 3, and 5) rainfall stations, respectively. ANFIS model developed based on MICCA inputs (ANFIS-MICCA) performed slightly better than the one developed based on CCA inputs (ANFIS-CCA). ANFIS-CCA and ANFIS-MICCA were able to perform comparably to HEC-HMS model where rainfall data of all 10 stations had been used; however, in peak estimation, ANFIS-MICCA was the best model. The sensitivity analysis on HEC-HMS was conducted by recalibrating the model by using the same selected rainfall stations for ANFIS. It was concluded that HEC-HMS model performance deteriorates if the number of rainfall stations reduces. In general, ANFIS was found to be a reliable alternative for HEC-HMS in cases whereby not all rainfall stations are functioning. This study showed that the selected stations have received the highest total rain and rainfall intensity (stations 3 and 5). Moreover, the contributing rainfall stations selected by CCA and MICCA were found to be located near the outlet of contributing sub-catchments. This provides valuable information towards identifying the more contributing sub-catchments in catchments such as Sungai Kayu Ara where no flow measurement is available for sub-catchments.

1. Introduction

Rainfall-runoff (R-R) modeling is an important topic in hydrology research. It aims to capture the rainfall-runoff association and understand its process. R-R modeling contributes in resolving many hydrological problems such as flood forecasting, water resources management, and urban water planning. To date, several R-R modeling techniques have been developed and employed which are widely available in literature. A well-versed group of techniques is the system theoretic model which does not require the physical process to be considered. Instead, it focuses on the direct relationship between the rainfall and runoff data. In the 1990s and early 2000s, several well-known system theoretic models were adopted in R-R modeling such as regression models, Artificial Neural Networks (ANN), and Neuro-Fuzzy Systems (NFS). However, many of the models inherit a black-box nature, hence modeling approaches shifted completely from black-box models to semantic-based fuzzy systems in recent years (Ang and Quek, 2005). NFS is a fuzzy system that is derived from a hybrid of fuzzy theory and neural networks. It can capture the non-linear association between the input and output through fuzzy logic by using low level learning capabilities of neural networks (Cho et al., 2009).

Fuzzy models that assume local model presentations with local function dynamics at the consequent or rule-layer of the models are known as Takagi-Sugeno-Kang (TSK) models (Takagi and Sugeno, 1985). TSK models inherit the ability to perform estimations for non-linear systems (Quah and Quek, 2006). Adaptive

* Corresponding author.
E-mail address: amin.talei@monash.edu (A. Talei).

http://dx.doi.org/10.1016/j.jhydrol.2016.12.024
0022-1694/

Network-based Fuzzy Inference System (ANFIS) (Jang, 1993) is one example of a TSK model which conducts learning through the minimization of global error within the model. In hydrological modeling and water resources application, ANFIS has been widely used in a number of applications including R-R modeling (Mukerji et al., 2009; Nayak et al., 2004, 2005b; Remesan et al., 2009). Several studies have concluded the superiority of ANFIS to other data-driven models such as ANN, Auto-Regressive Moving Average (ARMA), and Auto Regressive with exogenous inputs (ARX) models (Mukerji et al., 2009; Nayak et al., 2004, 2005b; Remesan et al., 2009). ANFIS also has been compared with different physically-based and conceptual R-R models such as Storm Water Management Model (SWMM) (Talei et al., 2010a) and HEC-HMS (Ji et al., 2012) in which ANFIS results are found comparable to the ones obtained by those models.

Unlike physically-based R-R models, where model inputs are defined based on physical parameters of the catchment, selecting proper type and number of inputs to be used in data-driven models is more challenging as they are not known a priori (Govindaraju, 2000). The type of the input is mainly depending on the problem and availability of data. In some studies, rainfall antecedents are considered as the only inputs of the data-driven model (Chua et al., 2008; Sajikumar and Thandaveswara, 1999; Talei et al., 2010a) while in many studies a combination of rainfall and discharge antecedents are used (Aqil et al., 2007; Dawson and Wilby, 1998; Nayak et al., 2005a; Riad et al., 2004; Talei et al., 2013). There are very few studies also in which other parameters such as temperature (Talei et al., 2013; Tokar and Markus, 2000), soil moisture deficit (Cheng and Noguchi, 1996), and infiltration rate (Tayfur and Singh, 2006) have been used in addition to rainfall and discharge inputs. Physical understanding of the problem can lead to better choice of input variable for a proper capturing of the R-R relationship; however, the appropriate inputs to be used need to be determined explicitly in order to achieve reasonable results.

Knowing the type of inputs, the next challenge in input selection procedure for a data-driven model will be the proper number of inputs. To date, several approaches have been used for input selection in data-driven models. In order to select the proper antecedents of rainfall (as the main input in R-R models), some studies have used a sequence of rainfall time series in a time window which starts from present time to a specific time (Firat and Güngör, 2008; Tokar and Markus, 2000). This time window can be determined by sensitivity analysis (Tokar and Johnson, 1999), correlation analyses (Sohail et al., 2008), or by assuming the catchment time of concentration as the window threshold (Jain and Prasad Indurthy, 2003). Some studies have suggested a narrower time window around the most correlated rainfall antecedent with runoff for which pruning of the unnecessary inputs is required (Nayak et al., 2007, 2005a). In some studies where runoff antecedents were supposed to be considered as input, auto-correlation and partial auto-correlation analyses have been adopted (Jain et al., 2004; Mutlu et al., 2008; Senthil Kumar et al., 2012; Sudheer et al., 2002). Cross-correlation analysis has been used widely in input selection of many of data-driven R-R modeling studies (Lekkas et al., 2001; Lohani et al., 2014; Maier and Dandy, 1997). Bowden et al. (2005) presented two input selection methodologies namely partial mutual information (PMI) algorithm and self-organizing map (SOM) for using ANN in water resources applications. Authors concluded that both approaches could be recommended when predictive performance is the primary aim. PMI has been also successfully employed in another study by He et al. (2011). de Vos and Rientjes (2007) used correlation coefficient and nonlinear average mutual information between output and potential inputs to identify the best input combination. In some studies, a pre-processing mechanism on rainfall inputs has been conducted by a linear transformation function. The weights of such function can be found in parametric form by using two-parameter Gama distribution (Jacquin and Shamseldin, 2006; Noori et al., 2011; Shamseldin, 2010). Principal component analysis (PCA) is another input selection method which is based on shrinkage feature selection and has been used in very few R-R modeling applications including Noori et al. (2011).

Despite successful usage of all afore-mentioned methods in input selection, applicability of them for event-based R-R modeling is not evaluated sufficiently. In event-based R-R modeling by data-driven models, input selection would be a challenging task since individual events' characteristics are important and influential. Talei et al. (2010a) studied the impact of hydrograph shape on ANFIS development for an event-based R-R modeling. In a separate study, Talei and Chua (2012), studied the effect of lag time on event-based R-R modeling by ANFIS. Talei et al. (2010b) investigated the effect of inputs on event-based runoff forecasting using ANFIS on an experimental catchment in Singapore. Authors concluded that non-sequential rainfall antecedents can produce better results compared to sequential rainfall inputs. This finding was also validated in a separate study by Talei and Chua (2012) on small partially urbanized catchment, in Kranji, Singapore. Authors, suggested that an input selection method based on correlation and mutual information analyses is able to identify optimum set of rainfall inputs for event-based R-R modeling by ANFIS.

From the reviews above, there are several input selection approaches for data-driven models such as ANFIS. However, there is lack of study involving catchments with multiple rainfall stations for which many potential rainfall antecedents can be chosen as inputs. This makes the task of selecting the optimum set of rainfall inputs (from different stations by different lead time) very challenging specially when dealing with event-based modeling. This is an important task as the performance of ANFIS could be affected due to unnecessary complexity when so many inputs are involved (Talei and Chua, 2012). Developing a robust model with fewer number of inputs would also be beneficial to reduce the computational time. Moreover, such studies can contribute in identifying the redundant rainfall stations in a catchment which in turn could be useful in moderating the maintenance costs of current stations, thus resulting in a more effective catchment data collection. Hence, the objective of this study is focused on selecting optimal number of rainfall inputs to develop an ANFIS model for event-based rainfall-runoff simulation. For this, the input selection method proposed by Talei and Chua (2012) is improved from single station rainfall problem to multiple rainfall stations while the potential presence of discharge antecedents is also considered. The results of developed ANFIS model are then compared with the ones obtained by conceptual R-R model, HEC-HMS.

2. Adaptive network-based fuzzy inference system (ANFIS)

Fuzzy Inference system (FIS) is a system that uses fuzzy set theory to formulate a mapping from an input to an output. A typical FIS comprises of four stages (Jang, 1993): (1) Fuzzification of inputs, (2) Application of fuzzy operator for each rule, (3) Aggregation of all output rules, (4) Defuzzification using different approaches like Center of Area (COA), Mean of Maximums (MOMs), etc. ANFIS is a Takagi-Sugeno FIS which is suitable for modeling ill-defined and uncertain systems through function approximations. ANFIS employs both, the low level learning style of neural networks and the high reasoning style of fuzzy systems. ANFIS can be constructed as a five layer multilayer perceptron (MLP) network. Further detail of each layer and their specific operations can be found in Talei et al. (2010b). ANFIS is implemented using the Fuzzy Logic Toolbox (MATLAB, 2013a).