Choice of Rainfall Inputs for Event-based Rainfall-Runoff Modeling in a Catchment with Multiple Rainfall Stations Using Data-driven Techniques

Citation for published version:
Chang, TK, Talei, A, Alaghmand, S & Ooi, M 2017, 'Choice of Rainfall Inputs for Event-based Rainfall-Runoff Modeling in a Catchment with Multiple Rainfall Stations Using Data-driven Techniques', *Journal of Hydrology*, vol. 545, pp. 100–108. https://doi.org/10.1016/j.jhydrol.2016.12.024

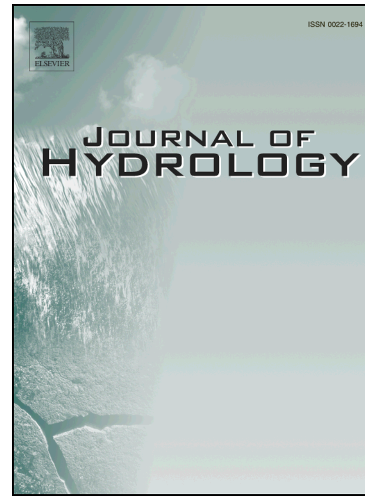
Digital Object Identifier (DOI):
10.1016/j.jhydrol.2016.12.024

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Peer reviewed version

Published In:
Journal of Hydrology

Accepted Manuscript



Research papers

Choice of Rainfall Inputs for Event-based Rainfall-Runoff Modeling in a Catchment with Multiple Rainfall Stations Using Data-driven Techniques

Tak Kwin Chang, Amin Talei, Sina Alaghmand, Melanie Po-Leen Ooi

PII: S0022-1694(16)30813-7
DOI: http://dx.doi.org/10.1016/j.jhydrol.2016.12.024
Reference: HYDROL 21703

To appear in: *Journal of Hydrology*

Received Date: 22 June 2016
Revised Date: 1 November 2016
Accepted Date: 12 December 2016

Please cite this article as: Kwin Chang, T., Talei, A., Alaghmand, S., Po-Leen Ooi, M., Choice of Rainfall Inputs for Event-based Rainfall-Runoff Modeling in a Catchment with Multiple Rainfall Stations Using Data-driven Techniques, *Journal of Hydrology* (2016), doi: http://dx.doi.org/10.1016/j.jhydrol.2016.12.024

Choice of Rainfall Inputs for Event-based Rainfall-Runoff Modeling in a Catchment with Multiple Rainfall Stations Using Data-driven Techniques

Tak Kwin Chang[1], Amin Talei[2*], Sina Alaghmand[3], Melanie Po-Leen Ooi[4,5]

[1]PhD student, School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia.

[2]Senior Lecturer, School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia.

[3]Lecturer, School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia.

[4]Associate Professor, School of Engineering and Physical Sciences, Heriot-Watt University Malaysia, No. 1 Jalan Venna P5/2, Precinct 5, 62200 Putrajaya, Malaysia.

[5]Adjunct Senior Research Fellow, School of Engineering, Monash University Malaysia.

[*]Corresponding Author

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.

Keywords: input selection; rainfall-runoff modeling; ANFIS; HEC-HMS.

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 1990's and early 2000's, 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 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; Nayak et al., 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 modelling 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).

3. Methodology

3.1. Study Catchment and Data Used

Sungai Kayu Ara river basin is located in southeast of Kuala Lumpur, Malaysia, and covers an area of 23.22 km² as shown in Fig. 1. The main river of this basin originates from the reserved highland area of Penchala and Segambut. Sungai Kayu Ara river basin lies in equatorial zone and it subjects to northeast monsoon (December to March) and southwest monsoon (June to September). The inter-monsoon seasons normally starts from April to May and from October to November (Desa and Niemczynowicz, 1996). Annual mean rainfall in this river basin is more than 2000mm as stated by Desa et al. (2005). Average daily temperature ranges from 25°C to 33°C. The mean monthly relative humidity falls within 70% to 90% depending on the location and rainfall season. Most of the area in the study catchment has been flattened for development. The annual average evaporation rate for the Sungai Kayu Ara river basin is estimated to be 4 to 5 mm/day. In this study, 24 major 10-minutes interval rainfall-runoff events between March-1996 to July-2004 are considered from which the first 18 (chronological order) were used for training while the remaining 6 events were used as testing data (Alaghmand et al., 2012).

3.2. Evaluation Criteria

The following error statistics and goodness of fit measures were adopted in this study:

1. Nash-Sutcliffe coefficient of efficiency or CE:

$$CE = 1 - \frac{\sum_{i=1}^{n}\left(Q_i - \hat{Q}_i\right)^2}{\sum_{i=1}^{n}\left(Q_i - \overline{Q}\right)^2} \qquad (1)$$

where Q_i and $\hat{Q}_i$ are the observed and simulated flow rate of the ith observation respectively, $\bar{Q}$ is the average value of the observed flow rate and n is the total number of observations. This ranges from -∞ to 1 and indicates how well a model explain the variance in the observed discharge. The ideal value for both metrics is one. CE was used to compare the goodness-of-fit between the measured flow and the simulated flow.

2. Coefficient of determination or r²:

$$r^2 = \left[\frac{\sum_{i=1}^{n} \left(Q_i - \overline{Q}\right)\left(\hat{Q}_i - \tilde{Q}\right)}{\sqrt{\sum_{i=1}^{n} \left(Q_i - \overline{Q}\right)^2} \times \sqrt{\sum_{i=1}^{n} \left(\hat{Q}_i - \tilde{Q}\right)^2}} \right]^2 \quad (2)$$

where $\tilde{Q}$ is the average simulated discharge. r^2 shows the degree of co-linearity between the observed and simulated time series and has a range of 0.0-1.0, with higher values indicating a higher degree of co-linearity.

3. Root mean square error or RMSE (m^3s^{-1}):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(Q_i - \hat{Q}_i\right)^2}{n}} \quad (3)$$

RMSE accords extra importance on the outliers in the data set and is therefore biased towards errors in the simulation of high flow rates (Dawson et al., 2006).

4. Mean Absolute Error or MAE (m^3s^{-1}):

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left|\left(Q_i - \hat{Q}_i\right)\right|}{n} \quad (4)$$

MAE computes all deviations from the original data regardless of sign and is not weighted towards high flow values (Abrahart et al., 2004).

5. Relative Peak Error (RPE):

In addition to the overall goodness-of-fit, accurate prediction of peak flow is also important. Thus, RPE has been included in this study to evaluate the ability of the proposed models to accurately predict peak flows. RPE is defined as:

$$\mathrm{RPE} = \frac{\left|\left(Q_p\right) - \left(\hat{Q}_p\right)\right|}{Q_p} \tag{5}$$

where Q_p and $\hat{Q}_p$ are the observed and simulated peak discharge. Values of RPE closer to zero indicate better estimation of peak flows.

3.3. Input selection Methods

Input selection analysis was required to identify the most informative rainfall antecedents from the most influential rainfall stations. One of the popular input selection methods is cross-correlation analysis (CCA) which has been widely used in R-R modeling by data-driven models (Brion et al., 2001; Coulibaly et al., 2000; Golob et al., 1998; Maier and Dandy, 1997). Mutual Information (MI) analysis is another approach to identify inputs which can be used for R-R modeling (Elshorbagy et al., 2010; He et al., 2011). Talei and Chua (2012) proposed an input selection method in which MI and CC analyses (MICCA) are employed. In this approach, the potential input data should be highly correlated with desired output while possessing low mutual information with other inputs. The correlation coefficient is expressed by:

$$\mathrm{C}C(x,y) = \frac{COV(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6}$$

where COV is the covariance between variables x and y; σ_x and σ_y are standard deviations of x and y, respectively; $\bar{x}$ and $\bar{y}$ are average values of x and y, respectively and n is the number of data points calculated between the input lead times and the runoff Q(t). In addition, mutual information is expressed by da Costa Couto (2009):

$$MI(x,y)=\frac{1}{2}\log\left(\frac{|C_{XX}||C_{YY}|}{|COV(x,y)|}\right) \tag{7}$$

where C_{xx} and C_{yy} are the variance of variables x and y respectively. In the present study, a ranking coefficient has been defined to select the optimal combination between inputs based on calculated MI and CC values:

$$R_k=\sum_{i=1}^{n}\hat{C}_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\hat{I}_{ij} \tag{8}$$

where $\hat{C}_i$ is the normalized correlation coefficient between Q(t) and a rainfall lead time in kth input combination; n is the number of inputs in this specific input combination; and $\hat{I}_{ij}$ is the normalized value for 1 – MI calculated for ith and jth rainfall lead times in kth input combination. Since MI is mutual information, 1 – MI represents independence. For an input combination with n inputs the second term of Eq. (8) includes $C_n^2=n!/2!(n-2)!$ number of mutual information terms. For example, in an input combination with 4 inputs, 6 terms of mutual information are expected. The ideal normalized value of $\hat{C}_i$ (for the maximum correlation) and 1 – MI (for the lowest mutual information) would be 1. Therefore, R can be 3, 6, 10, … for input combinations with 2, 3, 4, … rainfall inputs, respectively. In this study, the top five input combinations with 2, 3, 4, … inputs selected by CCA and MICCA approaches were used to develop ANFIS models. Finally, the results obtained by each set of models were compared to identify the best set of inputs for CCA and MICCA.

3.4. Data preprocessing and model development

Data standardization were implemented before model training and testing. Standardization concentrates the dispersed data to a defined interval. All input and output data were standardized over an interval between 0.1 and 0.9 using a standardization method proposed by Rajurkar et al. (2002). The method is expressed by:

$$x_n = 0.1 + 0.8\left(\frac{x_i}{x_{\max}}\right) \tag{9}$$

where x_n is the normalized measurement; x_i is the observed measurement; and $x_{\max}$ is the maximum observed measurement.

ANFIS model performance was analyzed through different model parameters such as number of membership functions, membership function type and epoch number. The optimum number of membership functions was determined to be 2 where increasing the number yields diminishing results. This was also consistent with the study by Nayak et al. (2004). The type of membership function was narrowed down to either a triangular or Gaussian membership function through a trial-and-error procedure. Since ANFIS performance was almost similar for both triangular and Gaussian membership functions, the triangular membership function was adopted to maintain simplicity. Moreover, a sensitivity analysis was also conducted for number of epochs to avoid over-fitting. Epoch number of 60 was resulted and adopted for ANFIS model.

4. Results and Discussion

Input selection based on CCA was carried out on the training dataset to identify the most correlated rainfall antecedent with runoff for each of stations. The correlation coefficient (CC) values between discharge at present time Q(t) and rainfall antecedents are presented in Fig. 2 for all 10 rainfall stations of this study. The most correlated rainfall antecedent for each station and its corresponding CC value is presented in Table 1. As can be seen, the top five CC values are produced by stations 1, 5, 8, 3, and 2, respectively. The gradual increase in number of rainfall inputs in ANFIS from 2 to 5 showed that using more inputs not only does not improve the model performance but also deteriorates it. This was attributed to the unnecessary complexity added into the ANFIS architecture due to the redundant inputs. Moreover, using rainfall inputs alone was found insufficient to produce precise estimation of Q(t). Therefore, an autro-correlation analysis was conducted (See Fig. 3) to assess the potential usage of discharge antecedents as inputs. As can be seen in Fig. 3, the first and second antecedents of discharge have quite high correlation coefficient of 0.93 and 0.77, respectively; however, CC is reduced drastically for the remaining lead time as it goes beyond (t-2). Therefore, the first and second discharge antecedent Q(t-1) and Q(t-2) were assessed to be added to the input combinations. Initial results showed that using Q(t-1) alone is sufficient as using Q(t-2) didn't improve the results. The best ANFIS model developed based on CCA was resulted when ANFIS used R1(t-4), R5(t-5), and Q(t-1) as inputs. This model is denoted as ANFIS-CCA in this study.

Similar approach of gradual increase of rainfall inputs from 2 to 5 was also carried out to identify the proper input set by MICCA. The top 5 combinations for each case (i.e. having 2, 3, 4, and 5 rainfall inputs) were selected for which Q(t-1) was also considered as the additional required input. Similar to CCA results, a sensitivity analysis showed that adding Q(t-2) does not improve ANFIS results. Table 2 summarizes the top 5 selected rainfall input sets by MICCA, their ranking, and their corresponding ANFIS performance in terms of average CE, r^2, RMSE, MAE and RPE values on testing events. As can be seen in Table 2, the best ANFIS performance was obtained when R1(t-8), R3(t-3), R5(t-6), Q(t-1) were used as inputs. This model is denoted as ANFIS-MICCA in this study.

The results of the two ANFIS models namely ANFIS-CCA and ANFIS-MICCA were compared against the ones obtained by HEC-HMS (Alaghmand et al., 2010) and are presented in

Table 3. In this table, the CE, r^2, RMSE, MAE and RPE values are average values over the 6 testing events. As can be seen, ANFIS-MICCA was able to consistently outperform the other two models. The improvement in terms of CE, r^2, RMSE, MAE were modest; however, RPE value obtained by ANFIS-MICCA was about 32% and 22% lower than the ones obtained by HEC-HMS and ANFIS-CCA models, respectively. For better comparison, the boxplots of RPE values obtained by ANFIS-CCA, ANFIS-MICCA, and HEC-HMS models for the six testing events are presented for in Fig. 4. The box plots indicate the 25th percentile (Q_1), median, and 75th percentile (Q_3) statistics. The lower and upper whiskers indicate the values Q_1 – 1.5(IQR) and Q_3 + 1.5(IQR), respectively, while IQR is the interquartile range. Values out of the range of lower and upper whiskers are considered as outliers. As can be seen in Fig. 4, ANFIS-MICCA results are comparable with the ones obtained by other two models; however, ANFIS-MICCA shows better consistency in peak estimation as the values for the six testing events are less scattered when compared to the other two models. Therefore, among the two ANFIS models, ANFIS-MICCA was considered for further comparison with HEC-HMS.

The observed versus simulated hydrographs by ANFIS-MICCA and HEC-HMS models for the 6 testing events are shown in Fig. 5a-f. As can be seen, both models were able to simulate discharge comparably; however, HEC-HMS showed significant underestimation of peaks in events with low flows (Fig. 5c and 5e). Among the six testing events of this study, Events 1 and 6 were high flow events. For Event 1, both models underestimated the peak; however, HEC-HMS showed slightly better performance. For Event 6, both models overestimated the peak where HEC-HMS was again slightly better than ANFIS-MICCA. In general, results obtained by ANFIS-MICCA for different ranges of testing events of this study were found to be comparable to the ones obtained by HEC-HMS.

Although the proposed ANFIS model in this study showed marginally similar performance in simulating discharge compared to HEC-HMS, there was a major difference in their data usage. In HEC-HMS model, the rainfall data of all 10 stations were used while in the two ANFIS models ANFIS-CCA and ANFIS-MICCA only 2 and 3 rainfall stations were used, respectively. In order to have a better understanding of HEC-HMS sensitivity to rainfall data, two new HEC-HMS models were calibrate using the same selected two (stations 1 and 5) and three (stations 1, 3, and 5) rainfall stations by CCA and MICCA approaches. These two models are denoted as HEC-HMS-2 and HEC-HMS-3, respectively. This analysis was considered to assess the performance of HEC-HMS model for the case that catchment has less number of active rainfall stations. The average values of CE, r^2, RMSE, MAE, and RPE for the six validation events are presented and compared with other models in Table 3. As can be seen, the performances of HEC-HMS-2 and HEC-HMS-3 are deteriorated in terms of all statistics compared to HEC-HMS model where the data of all ten rainfall stations were used. It is worth mentioning that HEC-HMS-3 performed slightly better than HEC-HMS-2 and it was concluded that HEC-HMS model performance could consistently improve by increasing the number of rainfall stations used in calibration and validation. In general, it was inferred that HEC-HMS is more sensitive to the number of rainfall stations used in its calibration compared to ANFIS models. Therefore, the proposed ANFIS model can be advantageous where for any reason (malfunctioning, periodic maintenance, etc.) one or some of the stations temporarily or permanently are declared as out of service. In such occasions, ANFIS can be a reliable discharge estimator which can operate with very few number of rainfall stations.

In order to investigate the selected rainfall stations (R1, R3, and R5), rainfall distribution on the 10 rainfall stations of this study was assessed. The statistics of recorded rainfall data for

24 events of this study (training and testing) are provided in Table 4. As can be seen, stations 3 and 5 have got the highest total and average rainfall over the 24 recorded events. Moreover, station 5 shows the highest diversity in rainfall events followed by station 1 where standard deviation (STDEV) was high compared to other stations. For further comparison, the boxplots of rainfall intensities of the 24 rainfall events recorded in the 10 stations are presented in Fig. 6. Station 3 has got the highest median rainfall intensity of 12.5 mm/h (average value of 14.6 mm/h) followed by stations 9 and 8. However, looking at the average intensity values, station 5 stands in the second position by 13.8 mm/h. It is worth mentioning that the outliers in each station in Fig. 6 are related to one single event in which all stations have recorded very high intensities. Moreover, by referring to Fig. 1 and checking the location of the selected stations by CCA (stations 1 and 5) and MICCA (stations 1, 3, and 5), it can be seen that they are located near the main branches of the river. It can be concluded that these branches are the most contributing ones to the main stream and catchment outlet. On the other hand, the distribution of the selected stations (stations 1, 3, and 5) over the catchment is in a way that they can represent the dominating down, middle, and upstream catchment rainfall. This could be valuable information in this study catchment as there is no other flow measurement along the stream to evaluate the contribution of individual river branches to the catchment outlet. Therefore, selecting the most informative rainfall stations can partially compensate the lack of such data in developing a reliable discharge estimator model.

5. Conclusion

The following can be concluded from this study:

i. Two input selection methods were used to identify the proper rainfall inputs for ANFIS model for event-based rainfall-runoff modeling in a tropical catchment with 10 rainfall stations: (1) Cross-correlation analysis (CCA); and (2) mutual-information and cross-correlation analyses (MICCA). ANFIS model developed by the inputs selected by MICCA was slightly superior to the one developed by CCA.

ii. The two ANFIS models ANFIS-CCA and ANFIS-MICCA were able to perform comparably to HEC-HMS by using rainfall data of only 2 and 3 stations, respectively while HEC-HMS used data from all 10 rainfall stations. It was also found that HEC-HMS models developed using the same 2 and 3 selected rainfall stations used in ANFIS models performed drastically worse compared to the original HEC-HMS model.

iii. The selected rainfall stations by CCA and MICCA have had the highest total rain and intensity (stations 3 and 5) and event diversity (stations 1 and 5). Moreover, these stations are located near some of the branches of the river catchment. It was concluded that those branches could be the most contributing ones to the discharge at catchment outlet. This information could be valuable in a catchment such as Sungai Kayu Ara where no other flow stations are available to evaluate the contribution of the river branches.

iv. The proposed ANFIS model was found to be advantageous as an alternative approach for HEC-HMS when one or more rainfall stations are not operating efficiently or are out of service.

6. Acknowledgement

Financial support from Monash University Malaysia is gratefully acknowledged for the PhD scholarship scheme. This project was funded by a grant FRGS/1/2014/TK02/MUSM/03/1 from the Ministry of Higher Education, Malaysia.

7. References

Abrahart, R., Kneale, P. E., and See, L. M., 2004, Neural networks for hydrological modeling, CRC Press.

Alaghmand, S., bin Abdullah, R., Abustan, I., and Eslamian, S., 2012, Comparison between capabilities of HEC-RAS and MIKE11 hydraulic models in river flood risk modelling (a case study of Sungai Kayu Ara River basin, Malaysia): International Journal of Hydrology Science and Technology, v. 2, no. 3, p. 270-291.

Alaghmand, S., Bin Abdullah, R., Abustan, I., and Vosoogh, B., 2010, GIS-based river flood hazard mapping in urban area (a case study in Kayu Ara River Basin, Malaysia): Int. J. Eng. Technol, v. 2, no. 6, p. 488-500.

Ang, K. K., and Quek, C., 2005, RSPOP: Rough Set–Based Pseudo Outer-Product Fuzzy Rule Identification Algorithm: Neural Computation, v. 17, no. 1, p. 205-243.

Aqil, M., Kita, I., Yano, A., and Nishiyama, S., 2007, A comparative study of artificial neural networks and neuro-fuzzy in continuous modeling of the daily and hourly behaviour of runoff: Journal of Hydrology, v. 337, no. 1-2, p. 22-34.

Bowden, G. J., Dandy, G. C., and Maier, H. R., 2005, Input determination for neural network models in water resources applications. Part 1 - Background and methodology: Journal of Hydrology, v. 301, no. 1-4, p. 75-92.

Brion, G. M., Neelakantan, T. R., and Lingireddy, S., 2001, Using neural networks to predict peak Cryptosporidium concentrations: Journal / American Water Works Association, v. 93, no. 1, p. 99-105.

Cheng, X., and Noguchi, M., 1996, Rainfall-raunoff modelling by neural network approach, Proc. of Int. Conf. on Water Resour. & Environ. Res., Volume Vol. (II): Kyoto, Japan, p. 143-150.

Cho, S.-Y., Quek, C., Seah, S.-X., and Chong, C.-H., 2009, HebbR 2-Taffic: A novel application of neuro-fuzzy network for visual based traffic monitoring system: Expert Systems with Applications, v. 36, no. 3, p. 6343-6356.

Chua, L. H. C., Wong, T. S. W., and Sriramula, L. K., 2008, Comparison between kinematic wave and artificial neural network models in event-based runoff simulation for an overland plane: Journal of Hydrology, v. 357, no. 3-4, p. 337-348.

Coulibaly, P., Anctil, F., and Bobée, B., 2000, Daily reservoir inflow forecasting using artificial neural networks with stopped training approach: Journal of Hydrology, v. 230, no. 3-4, p. 244-257.

da Costa Couto, M. P., 2009, Review of input determination techniques for neural network models based on mutual information and genetic algorithms: Neural Computing and Applications, v. 18, no. 8, p. 891-901.

Dawson, C. W., See, L. M., Abrahart, R. J., and Heppenstall, A. J., 2006, Symbiotic adaptive neuro-evolution applied to rainfall–runoff modelling in northern England: Neural Networks, v. 19, no. 2, p. 236-247.

Dawson, C. W., and Wilby, R., 1998, An artificial neural network approach to rainfall-runoff modelling: Hydrological Sciences Journal, v. 43, no. 1, p. 47-66.

de Vos, N. J., and Rientjes, T. H. M., 2007, Correction of timing errors of artificial neural network rainfall-runoff models: Hydroinformatics in Practice Computational Intelligence and Technological Developments in Water Applications.

Desa, M., Munira, M., Akhmal, H., and Kamsiah, A., Capturing extreme rainfall events in Kerayong catchment, *in* Proceedings 10th Int. Conf. on Urban Drainage, Copenhagen Denmark2005, p. 21-26.

Desa, M., and Niemczynowicz, J., 1996, Spatial variability of rainfall in Kuala Lumpur, Malaysia: long and short term characteristics: Hydrological sciences journal, v. 41, no. 3, p. 345-362.

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P., 2010, Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology: Hydrology and Earth System Sciences, v. 14, no. 10, p. 1931-1941.

Firat, M., and Güngör, M., 2008, Hydrological time-series modelling using an adaptive neuro-fuzzy inference system: Hydrological Processes, v. 22, no. 13, p. 2122-2132.

Golob, R., Štokelj, T., and Grgič, D., 1998, Neural-network-based water inflow forecasting: Control Engineering Practice, v. 6, no. 5, p. 593-600.

Govindaraju, R. S., 2000, Artificial neural networks in hydrology. II: Hydrologic applications: Journal of Hydrologic Engineering, v. 5, no. 2, p. 124-137.

He, J., Valeo, C., Chu, A., and Neumann, N. F., 2011, Prediction of event-based stormwater runoff quantity and quality by ANNs developed using PMI-based input selection: Journal of Hydrology, v. 400, no. 1-2, p. 10-23.

Jacquin, A. P., and Shamseldin, A. Y., 2006, Development of rainfall-runoff models using Takagi-Sugeno fuzzy inference systems: Journal of Hydrology, v. 329, no. 1-2, p. 154-173.

Jain, A., and Prasad Indurthy, S. K. V., 2003, Comparative analysis of event-based rainfall-runoff modeling techniques - Deterministic, statistical, and artificial neural networks: Journal of Hydrologic Engineering, v. 8, no. 2, p. 93-98.

Jain, A., Sudheer, K. P., and Srinivasulu, S., 2004, Identification of physical processes inherent in artificial neural network rainfall runoff models: Hydrological Processes, v. 18, no. 3, p. 571-581.

Jang, J.-S. R., 1993, ANFIS: adaptive-network-based fuzzy inference system: Systems, Man and Cybernetics, IEEE Transactions on, v. 23, no. 3, p. 665-685.

Ji, J., Choi, C., Yu, M., and Yi, J., 2012, Comparison of a data-driven model and a physical model for flood forecasting: 3rd International Conference on Flood Recovery, Innovation and Response, FRIAR 2012, v. 159, p. 133-142.

Lekkas, D. F., Imrie, C. E., and Lees, M. J., 2001, Improved non-linear transfer function and neural network methods of flow routing for real-time forecasting: J. Hydroinformatics, v. 3, no. 3, p. 153-164.

Lohani, A. K., Goel, N. K., and Bhatia, K. K. S., 2014, Improving real time flood forecasting using fuzzy inference system: Journal of Hydrology, v. 509, p. 25-41.

Maier, H. R., and Dandy, G. C., 1997, Determining inputs for neural network models of multivariate time series: Microcomputers in Civil Engineering, v. 12, no. 5, p. 353-368.

MATLAB, 2013a, MATLAB Fuzzy Logic Toolbox Reference Manual, Version 8.1.0. The Math Works Inc. Natick, Mass.

Mukerji, A., Chatterjee, C., and Singh Raghuwanshi, N., 2009, Flood forecasting using ANN, neuro-fuzzy, and neuro-GA models: Journal of Hydrologic Engineering, v. 14, no. 6, p. 647-652.

Mutlu, E., Chaubey, I., Hexmoor, H., and Bajwa, S. G., 2008, Comparison of artificial neural network models for hydrologic predictions at multiple gauging stations in an agricultural watershed: Hydrological Processes, v. 22, no. 26, p. 5097-5106.

Nayak, P. C., Sudheer, K. P., and Jain, S. K., 2007, Rainfall-runoff modeling through hybrid intelligent system: Water Resources Research, v. 43, no. 7.

Nayak, P. C., Sudheer, K. P., and Ramasastri, K. S., 2005a, Fuzzy computing based rainfall-runoff model for real time flood forecasting: Hydrological Processes, v. 19, no. 4, p. 955-968.

Nayak, P. C., Sudheer, K. P., Rangan, D. M., and Ramasastri, K. S., 2004, A neuro-fuzzy computing technique for modeling hydrological time series: Journal of Hydrology, v. 291, no. 1-2, p. 52-66.

-, 2005b, Short-term flood forecasting with a neurofuzzy model: Water Resources Research, v. 41, no. 4, p. 1-16.

Noori, R., Karbassi, A. R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M. H., Farokhnia, A., and Gousheh, M. G., 2011, Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction: Journal of Hydrology, v. 401, no. 3-4, p. 177-189.

Quah, K. H., and Quek, C., 2006, FITSK: online local learning with generic fuzzy input Takagi-Sugeno-Kang fuzzy framework for nonlinear system estimation: Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, v. 36, no. 1, p. 166-178.

Rajurkar, M., Kothyari, U., and Chaube, U., 2002, Artificial neural networks for daily rainfall—runoff modelling: Hydrological Sciences Journal, v. 47, no. 6, p. 865-877.

Remesan, R., Shamim, M. A., Han, D., and Mathew, J., 2009, Runoff prediction using an integrated hybrid modelling scheme: Journal of Hydrology, v. 372, no. 1-4, p. 48-60.

Riad, S., Mania, J., Bouchaou, L., and Najjar, Y., 2004, Predicting catchment flow in a semi-arid region via an artificial neural network technique: Hydrological Processes, v. 18, no. 13, p. 2387-2393.

Sajikumar, N., and Thandaveswara, B. S., 1999, A non-linear rainfall-runoff model using an artificial neural network: Journal of Hydrology, v. 216, no. 1-2, p. 32-55.

Senthil Kumar, A. R., Ojha, C. S. P., Goyal, M. K., Singh, R. D., and Swamee, P. K., 2012, Modeling of Suspended Sediment Concentration at Kasol in India Using ANN, Fuzzy Logic, and Decision Tree Algorithms: Journal of Hydrologic Engineering, v. 17, no. 3, p. 394-404.

Shamseldin, A. Y., 2010, Artificial neural network model for river flow forecasting in a developing country: Journal of Hydroinformatics, v. 12, no. 1, p. 22-35.

Sohail, A., Watanabe, K., and Takeuchi, S., 2008, Runoff analysis for a small watershed of tono area Japan by back propagation artificial neural network with seasonal data: Water Resources Management, v. 22, no. 1, p. 1-22.

Sudheer, K. P., Gosain, A. K., and Ramasastri, K. S., 2002, A data-driven algorithm for constructing artificial neural network rainfall-runoff models: Hydrological Processes, v. 16, no. 6, p. 1325-1330.

Takagi, T., and Sugeno, M., 1985, Fuzzy identification of systems and its applications to modeling and control: IEEE Transactions on Systems, Man and Cybernetics, v. 15, no. 1, p. 116-132.

Talei, A., and Chua, L. H. C., 2012, Influence of lag time on event-based rainfall–runoff modeling using the data driven approach: Journal of Hydrology, v. 438–439, no. 0, p. 223-233.

Talei, A., Chua, L. H. C., and Quek, C., 2010a, A novel application of a neuro-fuzzy computational technique in event-based rainfall-runoff modeling: Expert Systems with Applications, v. 37, no. 12, p. 7456-7468.

Talei, A., Chua, L. H. C., Quek, C., and Jansson, P.-E., 2013, Runoff forecasting using a Takagi–Sugeno neuro-fuzzy model with online learning: Journal of Hydrology, v. 488, no. 0, p. 17-32.

Talei, A., Chua, L. H. C., and Wong, T. S. W., 2010b, Evaluation of rainfall and discharge inputs used by Adaptive Network-based Fuzzy Inference Systems (ANFIS) in rainfall-runoff modeling: Journal of Hydrology, v. 391, no. 3-4, p. 248-262.

Tayfur, G., and Singh, V. P., 2006, ANN and fuzzy logic models for simulating event-based rainfall-runoff: Journal of Hydraulic Engineering, v. 132, no. 12, p. 1321-1330.

Tokar, A. S., and Johnson, P. A., 1999, Rainfall-runoff modeling using Artificial Neural Networks: Journal of Hydrologic Engineering, v. 4, no. 3, p. 232-239.

Tokar, A. S., and Markus, M., 2000, Precipitation-runoff modeling using artificial neural networks and conceptual models: Journal of Hydrologic Engineering, v. 5, no. 2, p. 156-161.

List of Figures:

List of Tables:

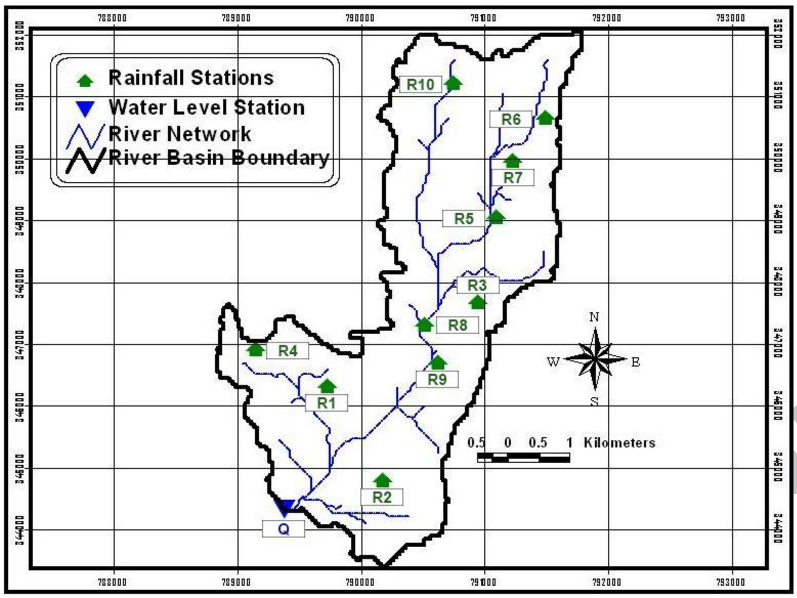
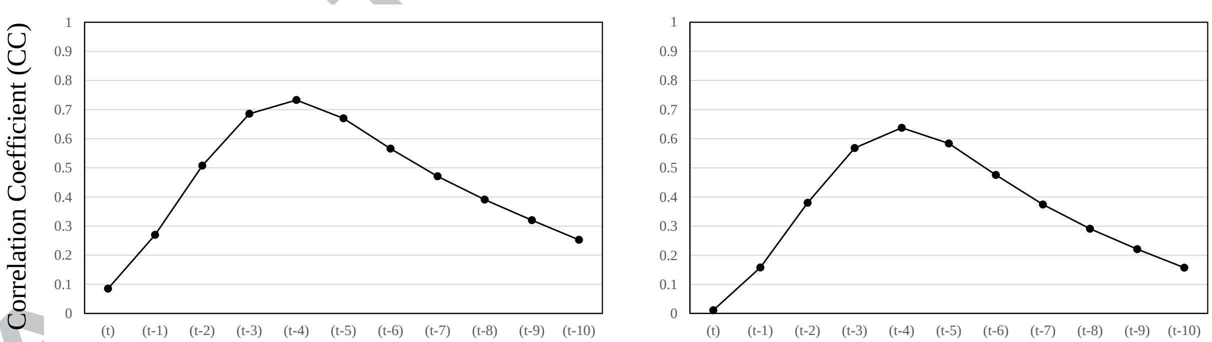


Figure 1: Schematic map of Sungai Kayu Ara catchment, Selangor, Malaysia.

Station R1

Station R2



Station R3

Station R4

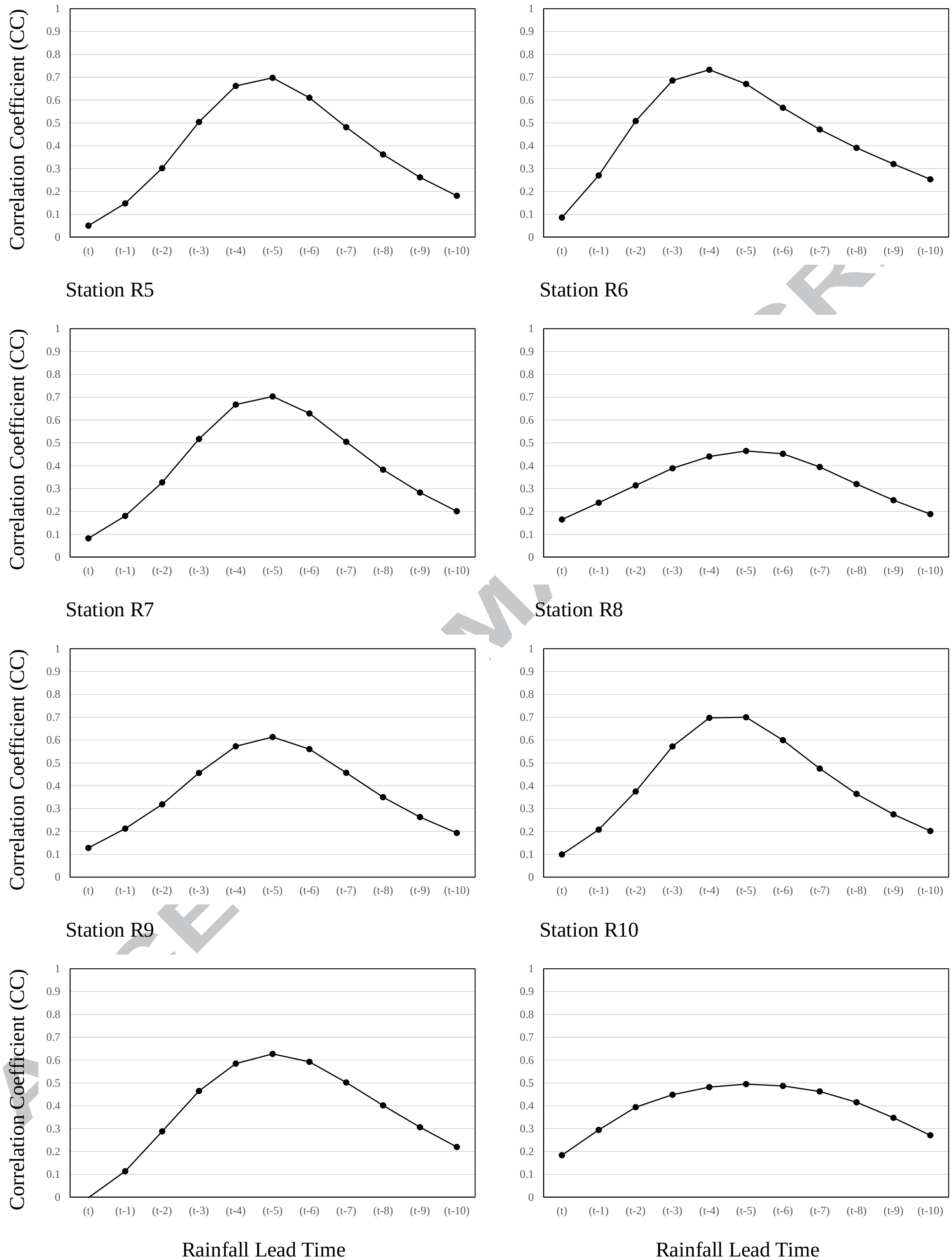

Correlation Coefficient (CC)
(t) (t-1) (t-2) (t-3) (t-4) (t-5) (t-6) (t-7) (t-8) (t-9) (t-10)
Station R5
Station R6
Station R7
Station R8
Station R9
Station R10
Rainfall Lead Time
Rainfall Lead Time

Figure 2: Cross-correlation results between discharge and rainfall antecedents for the 10 rainfall stations of Sungai Kayu Ara catchment.

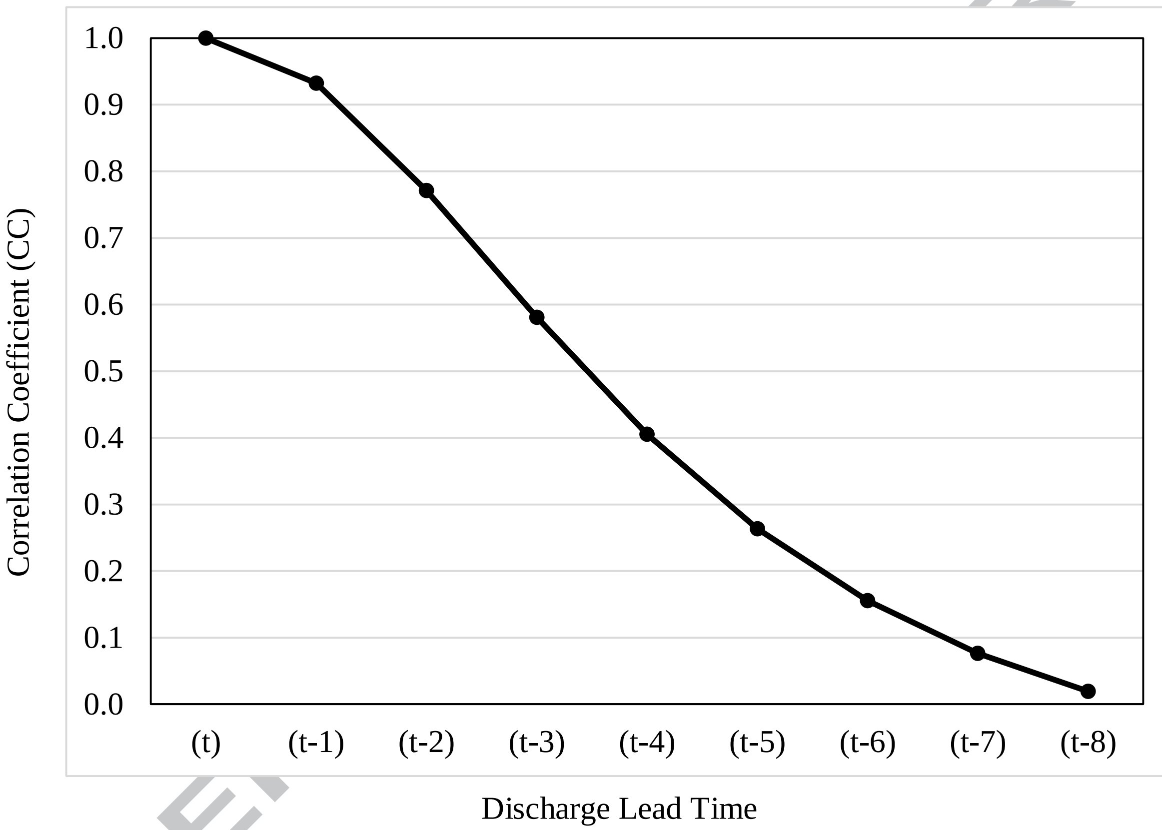


Figure 3: Auto-correlation analysis for discharge time series of Sungai Kayu Ara catchment

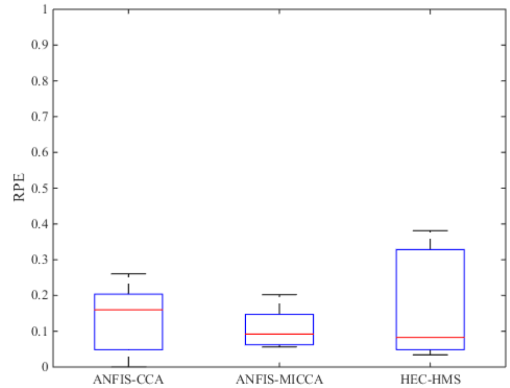
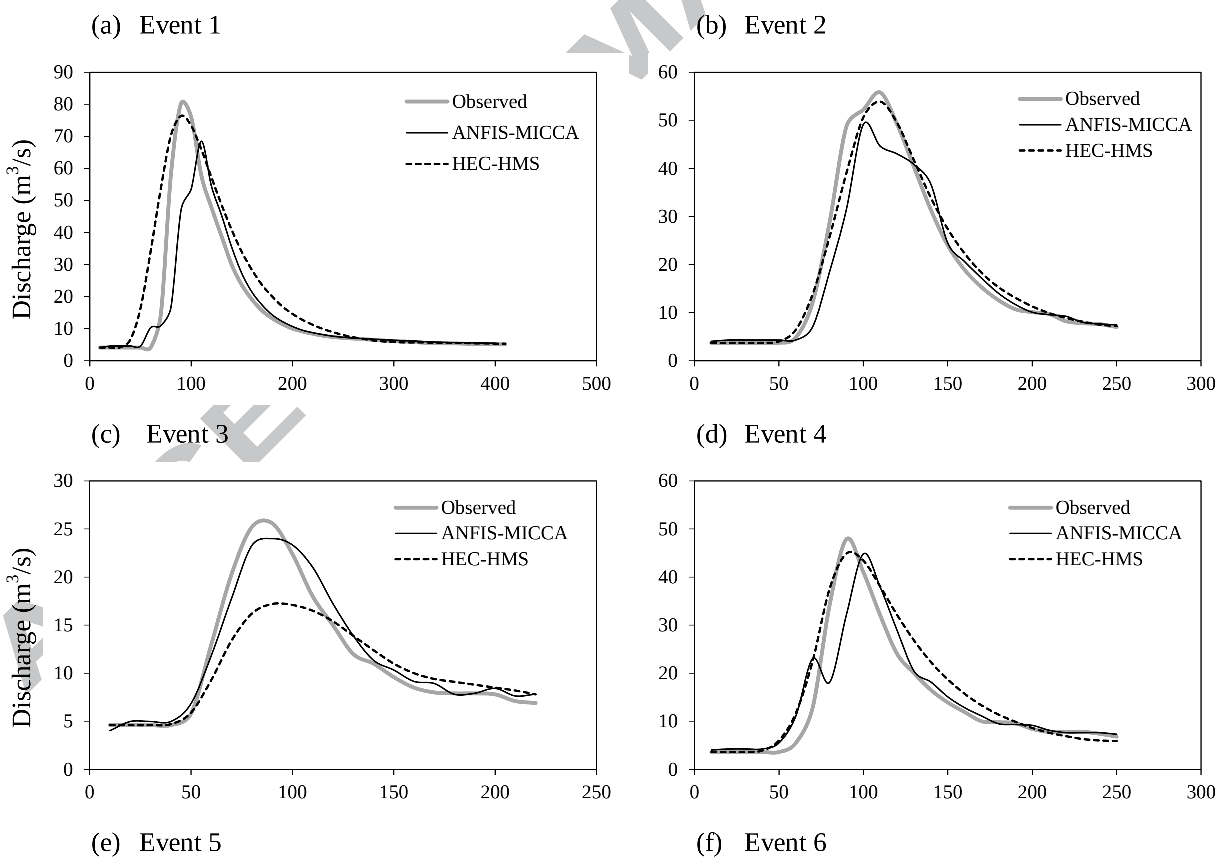


Figure 4: Box-plots of RPE values for the six testing events obtained by ANFIS-CCA, ANFIS-MICCA, and HEC-HMS models.



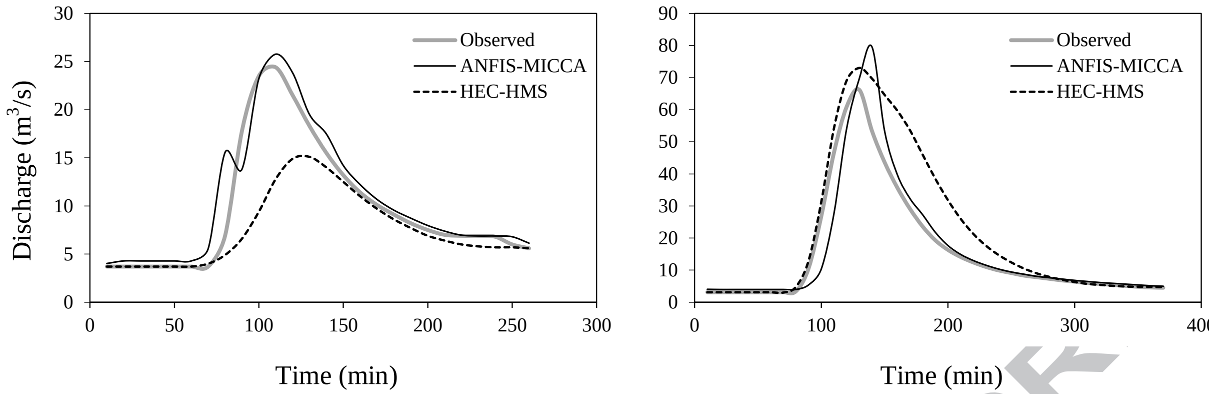


Figure 5: Observed versus simulated hydrographs by ANFIS-MICCA and HEC-HMS model for the six testing events: (a) Event 1, (b) Event 2, (c) Event 3, (d) Event 4, (e) Event 5, (f) Event 6.

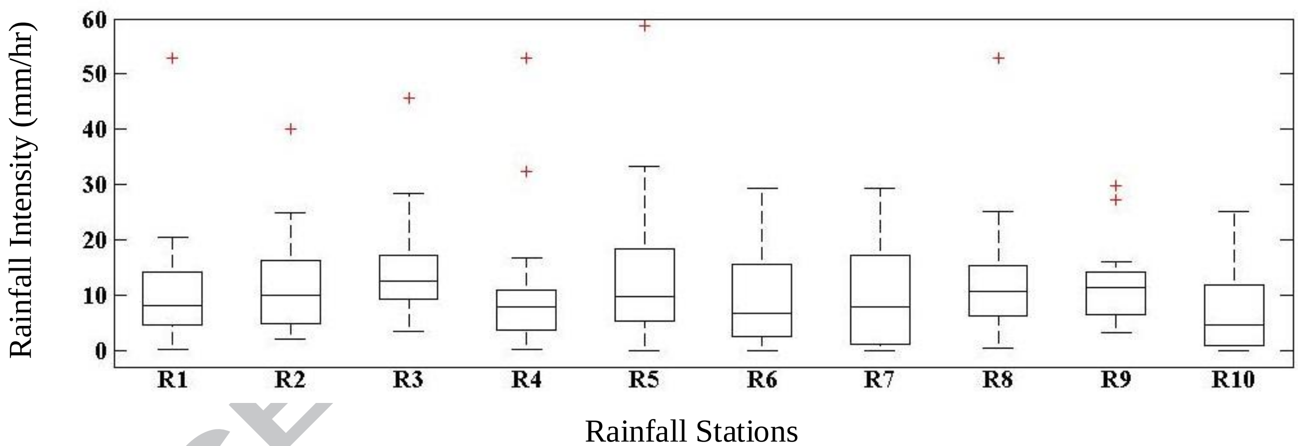


Figure 6: Distribution of rainfall intensities of 24 rainfall events in each rainfall station of the study catchment.

Table 1: The most correlated rainfall antecedent with discharge Q(t) for the 10 rainfall stations.

Station	The most correlated rainfall antecedent with Q(t)	Lead Time (min)	Correlation Coefficient (CC)
R1	R1(t-4)	40	0.732
R2	R2(t-4)	40	0.637
R3	R3(t-5)	50	0.697
R4	R4(t-5)	50	0.598
R5	R5(t-5)	50	0.703
R6	R6(t-5)	50	0.464
R7	R7(t-5)	50	0.613
R8	R8(t-5)	50	0.699
R9	R9(t-5)	50	0.627
R10	R10(t-5)	50	0.495

Table 2: Input combinations of 3, 4, 5, and 6 inputs chosen by MICCA method and their corresponding average ANFIS performance on testing events.

No. of Inputs	**Input Combination**	**Rank**	**CE**	**r^2**	**RMSE (m^3/s)**	**MAE (m^3/s)**	**RPE**
3	R2(t-3), R2(t-5), Q(t-1)	1	0.765	0.857	5.698	3.055	0.138
	R2(t-3), R3(t-5), Q(t-1)	2	0.770	0.871	5.541	2.998	0.149
	R2(t-3), R9(t-5), Q(t-1)	3	0.742	0.854	5.740	3.094	0.139
	R2(t-5), R5(t-5), Q(t-1)	4	0.770	0.797	6.640	3.340	0.077
	R2(t-5), R5(t-4), Q(t-1)	5	0.784	0.833	6.150	3.111	0.105
4	R1(t-4), R2(t-3), R2(t-6), Q(t-1)	1	0.697	0.827	7.290	3.386	0.258
	R1(t-8), R3(t-3), R5(t-6), Q(t-1)	***2***	***0.867***	***0.880***	***4.957***	***2.486***	***0.109***
	R2(t-4), R3(t-6), R5(t-4), Q(t-1)	3	0.823	0.849	5.588	2.670	0.116
	R2(t-3), R2(t-6), R3(t-5), Q(t-1)	4	0.703	0.818	7.525	3.458	0.194
	R2(t-3), R3(t-5), R5(t-7), Q(t-1)	5	0.783	0.847	6.505	2.973	0.146
5	R4(t-4), R9(t-3), R9(t-6), R10(t-6), Q(t-1)	1	0.817	0.870	5.580	2.596	0.185
	R2(t-6), R4(t-4), R9(t-3), R10(t-6), Q(t-1)	2	0.718	0.842	7.073	3.233	0.304
	R3(t-6), R4(t-4), R9(t-3), R10(t-6), Q(t-1)	3	0.774	0.846	6.276	2.826	0.249
	R2(t-4), R4(t-3), R9(t-6), R10(t-5),	4	0.793	0.829	5.921	2.972	0.166

	Q(t-1)						
	R4(t-4), R9(t-3), R9(t-7), R10(t-6), Q(t-1)	5	0.747	0.841	6.698	3.096	0.338
6	R2(t-6), R4(t-4), R8(t-5), R9(t-3), R10(t-6), Q(t-1)	1	0.613	0.781	8.465	3.690	0.354
	R2(t-3), R3(t-5), R4(t-6), R10(t-4), R10(t-8), Q(t-1)	2	0.431	0.728	9.684	4.058	0.368
	R4(t-4), R8(t-5), R9(t-3), R9(t-7), R10(t-6), Q(t-1)	3	0.507	0.836	8.943	3.843	0.444
	R2(t-3), R4(t-5), R8(t-5), R10(t-4), R10(t-7), Q(t-1)	4	0.833	0.866	5.396	2.497	0.184
	R2(t-3), R3(t-5), R4(t-6), R10(t-5), R10(t-8), Q(t-1)	5	0.775	0.878	5.434	2.563	0.1696

Table 3: The average CE, r^2, RMSE, MAE, and RPE over the 6 testing events obtained by different models.

Model	CE	r^2	RMSE (m^3/s)	MAE (m^3/s)	RPE
ANFIS-CCA	0.827	0.839	5.321	2.700	0.139
ANFIS-MICCA	0.867	0.880	4.957	2.486	0.109
HEC-HMS	0.761	0.871	5.593	3.277	0.160
HEC-HMS-2	0.234	0.653	10.543	6.519	0.394
HEC-HMS-3	0.380	0.707	9.038	5.608	0.286

Table 4: Statistical data of 10 rainfall stations for 24 rainfall events of this study.

	Rainfall Stations									
Statistics	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**
Total Rain (mm)	640	681	807	590	722	563	514	685	666	439
Min (mm)	0.5	4.2	14.5	0.5	0.2	0.2	0.2	1.6	11.8	0.2
Max (mm)	88	66.8	76	88	98	61	65	88	89.2	72.5
Average (mm)	26.7	28.4	33.6	24.6	30.1	23.4	21.4	28.5	27.8	18.3
Median (mm)	25.1	28	28.5	19.9	25.1	25.4	23.6	24.2	21.5	10.3
STDEV (mm)	21.2	16.6	16.3	21.0	22.1	17.3	18.2	20.1	17.2	20.2

- This study is on input-selection for a catchment with multiple rainfall stations.
- Study catchment has 10 rainfall and one discharge stations.
- Combined Cross-correlation and mutual information analysis is proposed.
- Selected inputs were used to develop an event-based rainfall-runoff ANFIS model.
- Proposed method chose rainfall inputs only from 2-3 stations to be used in ANFIS.
- Selected stations had high total rain and were located near sub-catchments outlet.
- ANFIS was able to perform comparable to HEC-HMS by using 2 and 3 stations.
- HEC-HMS performance deteriorated when less number of rainfall stations were used.