Jun 26th, 5:00 PM - 7:00 PM

Variable selection for improving predictions of hydrological events

Marina G. Erechtchoukova
York University, marina@yorku.ca

Marina Zistler
York University, marina.zistler2@gmail.com

iEMSs International Congress on Environmental Modelling and Software

9th International Congress on Environmental Modelling and Software
Fort Collins, Colorado, USA, Mazdak Arabi, Olaf David, Jack Carlson, Daniel P. Ames (Eds.)
https://scholarsarchive.byu.edu/iemssconference/2018/

Variable Selection for Improving Predictions of Hydrological Events

Marina Zistler[a], Marina G. Erechtchoukova[a]
[a]*School of Information Technology, Faculty of Liberal Arts and Professional Studies, York University, Canada* (*marina.zistler2@gmail.com*, *marina@yorku.ca*)

Abstract Application of supervised classification to short-term predictions of hydrological events relies on data routinely collected by Conservation Authorities with high frequencies on streams and their watersheds. This implies that prediction quality depends on the location of monitoring stations and the representativeness of data sets used in the analysis. Given that the application of classification algorithms requires data transformation, an attempt was made to improve the performance of these algorithms by extending the set of variables of black-box models, which are supplied by stream and rain gauges, and to include their derivatives as they may carry very important information as well. The original computational scheme was based on the application of time-delay embedding to data from all observation sites of a watershed. The variable selection was implemented using both hydrological knowledge and computational procedures. The computational experiments were conducted on data of various granularity and years with different hydrological characteristics. The results of the study are presented in the paper.

Keywords: Flood event, classification algorithms, time-delay embedding, variable selection.

1 INTRODUCTION

Floods are among the most severe natural catastrophes, leading to both the loss of human life and economic damage globally. These natural disasters are ranked the fourth highest cause of death and the highest cause for affecting human life (CRED, 2016). Therefore, early warning systems are part of effective Flood Risk Management, helping improve the preparedness for floods and as a result, reduce their risk and impact. Warning systems must accurately forecast floods and provide estimates early enough to effectively mitigate the risk (Plate, 2002; United Nations, 2004). This becomes especially challenging when facing flash floods that are caused by rapidly rising water level because of intense rainfall over a small area or of moderate to intense rainfall over highly saturated or impervious land surfaces and generally occur within minutes to several hours of the rainfall event (AMS, 2000). Over the past 60 years, the nonlinear relationship between rainfall and streamflow has received considerable attention from the academic world, leading to computational techniques that range from complex physically-based models to black-box representations (Young, 2003).

Black-box models have been expanded with the application of stochastic and data mining algorithms. To describe the relationships between historical inputs of rainfall, air temperature and outputs such as watershed runoff or water level, these models operate as a transfer function which can be applied on the input data in order to receive the desired information on hydrological events. Artificial neural networks (ANNs) are a mature black-box model approach to hydrological problems (Aqil et al., 2007; Hu et al., 2001; ASCE 2000a; ASCE 2000b). Since then ANNs evolved into various hybrid schemes such as Bayesian artificial

neural networks (Kingston et al., 2005), neuro-fuzzy systems (Nayak et al., 2005), or deep learning tools (e.g., Bai et al., 2016). Other common techniques used for rainfall-runoff modeling are regression methods, including the M5 algorithm used, for example, by Solomatine and Dulal (2003) and genetic programming (Savic et al., 1999, Whigham & Crapper, 1999). Black-box models are especially promising for application on watersheds where little is known about morphometric and hydrological characteristics of the waterbody, noise is present, or the input is incomplete or ambiguous (Tokar and Johnson, 1999). While these models require a large amount of data, the data is often easy to acquire using hydrological sensors (i.e., stream and rain gauges) or is available through meteorological agencies. For effective flood management, what is required to make timely decisions is not the magnitude of hydrological characteristics but rather whether a flood event will occur at a location of interest. This allows for utilization of classification algorithms that can predict hydrological events as class labels. For early flood warning systems, the two classes of 'flood' and 'no-flood' are especially important. Existing research that explores the application of classification algorithms to prediction of hydrological events covers the use of categorical attributes as input variables for classification algorithms instead of using the actual measured values. McCulloch et al. (2008) employed this approach in the form of updateable linguistic decision trees for real-time forecasting of daily average water levels. Han et al. (2002) modeled the daily streamflow using fuzzy decision trees. Segretier et. al. (2012) aimed to address varying characteristics of watersheds by applying an evolutionary algorithm for variable selection and to create the best classifier juries using the majority vote (Segretier et al., 2012, 2013). Furquim et al. combined wireless sensor networks and machine learning for flash flood nowcasting (Furquim, et al., 2014).

It is necessary to preprocess hydrological data to apply machine learning algorithms. Given that a flood event is the result of hydrological processes and their recent dynamics, the observation data should be reconstructed into a phase space based on the time-delay embedding technique. This technique has been applied, amongst others, by Erechtchoukova et al. (2016) and Damle and Yalcin (2007).

This paper presents the results of exploratory computational experiments to extend the methodology for the short-term prediction of hydrological events by incorporating derivatives of water level and precipitation along with time series of these characteristics into the set of independent variables. The extension of the methodology was tested for different lead time intervals and classification algorithms. The investigation was conducted based on a case study of a small stream – Spring Creek, Ontario, Canada. The results showed that the derivatives carry important information, can improve model's performance and have the potential to extend the lead time for hydrological predictions.

2 PROBLEM FORMULATION

The most important information that operational flood early warning systems need to provide is whether a flood will occur at a specific location of interest. This enables officials to respond to hydrological emergencies promptly. Rather than predicting the actual magnitude of hydrological characteristics, each event can be assigned a class label indicating whether it was a high-flow ('flood') or low-flow ('no-flood') event. This calls for the application of supervised classification algorithms. For the flood early warning system to be an efficient tool, predictions must occur in a timely manner, allowing enough time for officials to take appropriate actions. Additionally, missing an actual flood event can have serious consequences, but inversely, a high number of false alarms will reduce the effectiveness of the system and result in a loss of public trust. Therefore, predictions should be made as early as possible while ensuring at least 80% accuracy. The model must produce reliable predictions with input from data generated by routine monitoring networks, which are usually equipped with stream and rain gauges.

2.1 Methodology

This study follows the methodology on the application of classification algorithms to short-term predictions of hydrological events using time-delay embedding and data from multiple observation sites. The methodology outlined the development of a black-box model as a heterogeneous ensemble of classifiers (Erechtchoukova et al., 2016). It is reasonable to assume that along with magnitudes of water level and

precipitation, their rates of change over time also carry important information about current and future hydrological conditions at an investigated watershed. Mathematically, the rate of change of a process is described as the derivative of a variable. Since the data about continuous processes are available in the discrete form of time series, the derivatives were replaced by finite difference approximations using the smallest possible timestep corresponding to the observation time interval. These approximations are further referred as deltas.

Including deltas for all independent variables into datasets doubles the size of a phase space which makes standard procedures for variable selection computationally expensive. Different classification algorithms disagree on the set of important variables due to variations in the goal functions applied by individual inducers. To overcome this obstacle, the hydrological analysis of observation data was undertaken to uncover the scale and patterns in durations and frequencies of intensive precipitation and high-flow events. Notable variations of these characteristics did not allow reduction of the subset of finite difference elements which would suit all inducers for extended lead time of predictions. Therefore, the exploratory computations to investigate the usefulness of additional information carried by deltas and to identify the reduced subset of these variables were conducted. Further refinement of the set of variables can follow standard procedures.

2.2 Data transformation

According to the time-delay embedding approach, time series of observation data are transformed into a phase space, in which tuples contain measurements from the recent past along with the current values of the observed variables. For each field of the tuple, its finite difference, delta, was added to the tuple as well. Deltas were computed as the difference between the water level or precipitation values of the current moment of time and the previous one. An element of the phase space z_t is then described by (1):

$$z_t = (Y_{1,t-(R-1)\tau}, (Y_{1,t-(R-1)\tau} - Y_{1,t-(R)\tau}), \dots, Y_{1,t}, (Y_{1,t} - Y_{1,t-\tau}), Y_{2,t-(R-1)\tau}, (Y_{2,t-(R-1)\tau} - Y_{2,t-(R)\tau}), \dots, Y_{2,t}, (Y_{2,t} - Y_{2,t-\tau}), \dots, Y_{M,t-(R-1)\tau}, (Y_{M,t-(R-1)\tau} - Y_{M,t-(R)\tau}), \dots, Y_{M,t}, (Y_{M,t} - Y_{M,t-\tau})) \quad (1)$$

where $Y_{k,t}$ is the value measured at observation site k at time t, k=1,…,M, the M^{th} observation site is the cross section of interest, $R\tau$ is the total window interval, and τ is the time interval between measurements.

Creating the phase space over a large window size (several consecutive observations at different points in time) can add redundancy, noise and can degrade the performance of the model and increase the computation time. Selecting a window size that is too small could leave out important information for the prediction, also decreasing the performance (Galka, 2000). Based on a previous analysis of the flood events at the investigated watershed showing that extensive rainfall takes around 2 to 3 hours to induce a rise in water level, the window size was set to 3 hours. To assign a class label to each tuple from the phase space, an event characterization function was introduced as (2):

$$f_k(z_t) = \begin{cases} 'flood', Y_{k,t} \geq H_{thresh} \\ 'no\text{-}flood', Y_{k,t} < H_{thresh} \end{cases}, \quad (2)$$

where H_{thresh} is the water level threshold to delineate between flood and no-flood conditions used by the Toronto and Regional Conservation Authority (TRCA).

To obtain more realistic estimates of the model's generalization error, the entire phase space was split into two subsets, allowing for training the model on one subset and testing its performance on unseen tuples from the other subset. The split was implemented by random selection of tuples without repetition, with 70% of data allocated to the training set and 30% of tuples forming the testing set. The selection was implemented using the Caret R package (Kuhn, 2017) in the way that the training set contained 70% of all tuples corresponding to the flood events and 70% of all 'no-flood' tuples. The rest of the tuples were put into the testing set.

3 EXPERIMENT DESIGN

Experiments were conducted using the Weka Experimenter (Hall et al., 2009). Seven inducers including five decision trees (J48, NBTree, RandomForest, SimpleCart, REPTree) and two rule-based algorithms (JRip, Ridor) were used. Each experiment was conducted with a lead time of 15 mins, 30 mins, 45 mins and 60 mins. As misclassifying a 'no-flood' event is not as dangerous as misclassifying a 'flood' event the Precision and Recall for 'flood' events were selected as the main indicators of model performance.
They are defined as (3) and (4):

$$Precision = \frac{TP}{TP+FP}, \tag{3}$$

$$Recall = \frac{TP}{TP+FN}, \tag{4}$$

where *TP* is the number of correctly identified tuples corresponding to flood events, *FP* is the number of misclassified tuples corresponding to no-flood conditions, and *FN* represent the number of incorrectly identified tuples corresponding to flood events. One round of baseline experiments was conducted without additional variables, representing finite differences of water level and precipitation. During the second round of experiments, delta attributes were added to the training and testing sets over the full window size once for all attributes, just the rainfall attributes and just the water level attributes. These experiments assessed how the different types of deltas affect the prediction results. The third round of experiments attempted to identify the most important period from the recent past with dynamics that affect the results. In this round, delta variables were added to the tuples for variables corresponding to the most recent one- and two-hour intervals.

4 CASE STUDY

The experiments were performed based on a dataset collected from the Spring Creek watershed in the Greater Toronto Area, Ontario, Canada. The Spring Creek is one of two tributaries of the Etobicoke Creek. The Spring Creek is over 23 km long and stretches through a watershed area of approximately 50 km^2. The main branch of the Etobicoke Creek and the Spring Creek are relatively steep. Most of its watershed is completely urbanized which creates a highly impervious surface. As a result, heavy rainfalls generate quick runoff responses in very short periods of time, often even less than an hour. This leads to water inundation and flash flood events starting from April and lasting until December (TRCA 2006). The data used for the experiments consisted of time series collected by the TRCA during the 2013 and 2014 years.

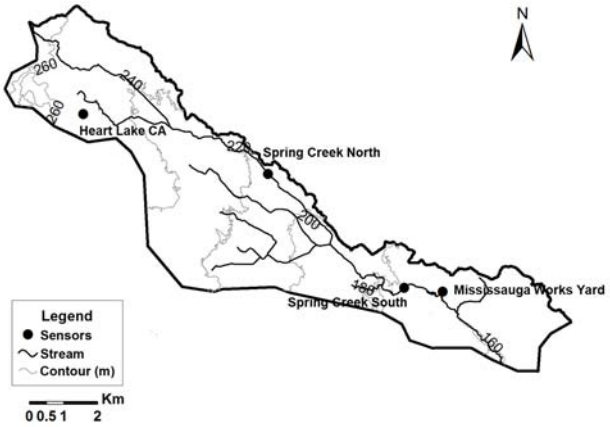

Figure 1. Location of observation sites

Each dataset includes measurements from two observations sites at the Spring Creek watershed that supply data on water level (Spring Creek North and Spring Creek South). TRCA also collects data on precipitation from two observation sites equipped with rain gauges (Heart Lake CA and Mississauga Works Yard) (Figure 1). Yearly rainfall ranges from around 500 mm in 2014 to 800 mm in 2013. Monthly rainfall varies from 30 mm to 190 mm. Monthly rainfall does not follow a periodical recurrence and usually spikes between the months of June and September. Instantaneous water discharge at Spring Creek South cross section varies from 0.05 m^3/s to 9 m^3/s in 2014 and from 0.1 m^3/s to 29 m^3/s in 2013. Therefore, the year 2013 can be described as wet and the year 2014 as hydrologically dry.

Water level data was measured every 15 minutes while the rain gauges generated data on a 5-minute interval. The precipitation data was upscaled to a 15-minute interval according to the hydrological characteristics of the tipping bucket rain gauges that were used for data collection. Time series from all four sensors were then consolidated. After the event characterization function was applied to the reconstructed tuples, the dataset for the 2013 year showed 388 tuples corresponding to flood events while the 2014 dataset had only 171 tuples representing flood events, which constituted approximately 44% of the previous year. Datasets for both years were transformed into corresponding phase spaces following the methodology for the short-term prediction of hydrological events and its extension according to (1) and (2). The phase spaces were combined together to ensure that hydrological dynamics from wet and dry periods are reflected and split into training and testing sets as it was described above.

5 RESULTS

The baseline experiments produced the results consistent with previous studies. The performance of classifiers declined with the increasing lead time of prediction. Individual classifiers performed differently on investigated data sets preventing the identification of a clear winner which outperforms other classifiers. After the baseline round of experiments was completed, the computations were performed on data sets with finite differences added to each independent variable in the set. Overall, the trained models demonstrated improved performance (Figure 2).

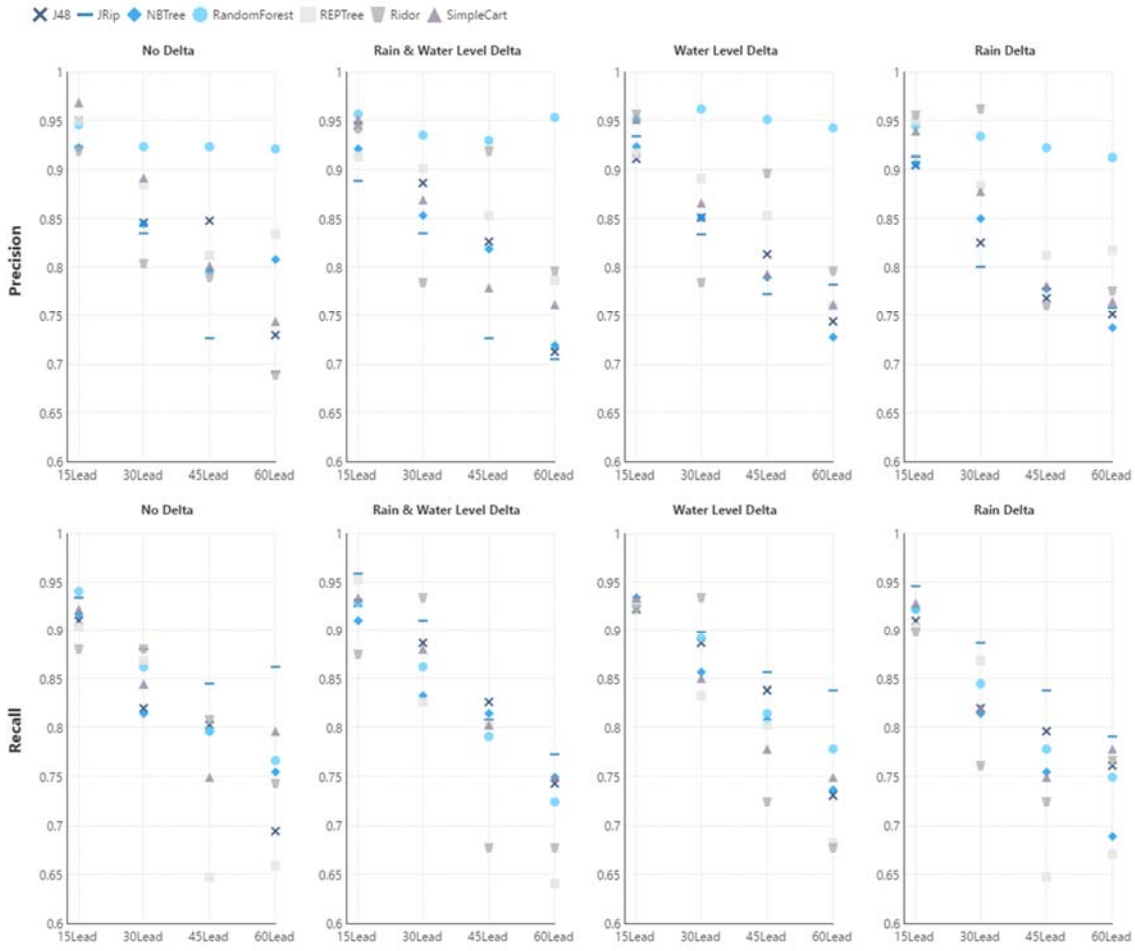

Figure 2. Comparison of models with different set of independent variables

In addition, the potential contribution of each type of variables to the model improvement was evaluated. Experiments were performed to obtain estimates of the Recall and Precision of models with deltas added to only variables representing water level magnitudes. Another set of experiments investigated the importance of changes in precipitation data. Figure 2 demonstrates that adding finite differences of water level variables to the phase space produces models with the best results. They outperform all other models developed in the study. Adding finite differences to all independent variables resulted in models which performed almost at the same level as the previous one. Estimates of their generalization errors exhibited steeper declines with increasing prediction lead time. The additional variables for rainfall data did not improve predictions. The performance of these classifiers was below the baseline models. These results suggested that some additional variables introduced more noise than valuable information. Therefore, a third round of experiments was performed to select the range of variables which carry more information than the others.

The finite differences of independent variables were added over the preceding one and two hours prior to the moment of prediction generation. The constructed models were tested and the results were compared with both the baseline experiments and the full delta experiments described above. This scheme was applied to all variables, only to variables representing precipitation and only for those corresponding to water level magnitudes. The results of this round of experiments are presented in Figure 3.

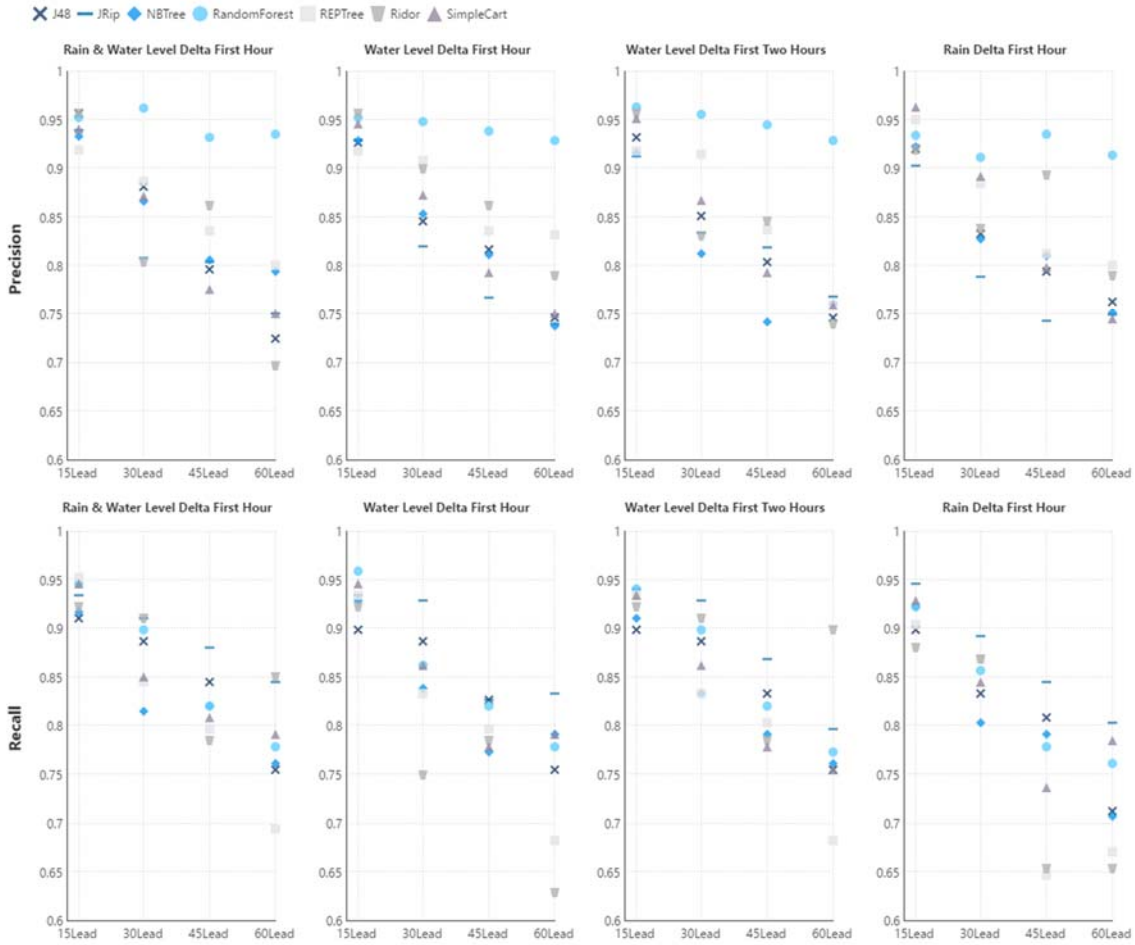

Figure 3. Comparison of classifiers developed with the finite differences of variables corresponding to one- and two-hour windows

All experiments with additional variables corresponding to water level showed that adding the most recent finite differences improves the performance of the developed models across all lead times. The Precision

results for one-hour and two-hour window were very similar, while for Recall, only the first hour of delta achieved the better results for all lead times. Precipitation finite differences did not make a notable impact on the model performance.

6 DISCUSSION AND CONCLUSIONS

The performed preliminary analysis presented in the paper confirms the hypothesis that adding information about independent variables to the model can improve model performance. To achieve better performance, however, it is necessary to choose the most informative variables to reduce the effect of noise. It has been shown that even though rainfalls are the major contributor to the development of flash flood events in the Spring Creek watershed, the changes in precipitation amounts do not provide additional information on a future event. Adding such variables resulted in deterioration of prediction results. Further investigation of the intensive precipitation dynamics, including other meteorological parameters may be required to provide justification. One of the possible reasons can be in the scale of involved processes.

The experiments demonstrated that additional information on water level dynamics is important when it reflects the most recent changes in water level magnitudes. Further extension of the set of such variables does not improve model performance. Given that extending the set of model variables increases the size of datasets by 15% - 100%, the preliminary analysis is necessary to exclude non-informative variables.

The results obtained for individual classifiers indicated that the Recall and Precision of some of them increased by 5% - 9%. The ensembles of baseline models performed better than their individual members and supported reliable predictions with 20% uncertainty for a 45-minute lead time. Both facts suggest further investigation of ensembles of classifiers developed following extended methodology to estimate the lead time of reliable predictions which can be achieved in the investigated watershed.

The conducted analysis was not aimed to substitute the formal process of variable selection. It was implemented to outline and possibly reduce the search space of important variables by excluding those which do not improve model performance. As such, it is recommended to use finite differences of water level variables corresponding to the first immediate hour. The selection of deltas suggested by the outcomes of the computational experiments is consistent with available knowledge about hydrological processes in the investigated watershed. The formal process for variable selection is a subject of future investigation.

ACKNOWLEDGMENTS

The analysis was implemented on data received from Toronto and Region Conservation Authority. The authors are thankful to TRCA for providing this invaluable resource and necessary clarifications. The authors are grateful to editors and anonymous reviewers for their thoughtful suggestions and helpful comments on the manuscript which led to its improvement. The advices on future research are very much appreciated.

REFERENCES

American Meteorological Society (AMS), 2000. http://glossary.ametsoc.org/wiki/Flash_flood (last accessed 11.11.2017).

Aqil, M., Kita, I., Yano, A., Nishiyama, S., 2007. Neural networks for real time catchment flow modelling and prediction. Water Resources Management, 21(10), 1781-1796.

ASCE, 2000a. Artificial neural networks in hydrology. I: Preliminary concepts. Journal of hydrologic Engineering, 5(2), 115-123.

ASCE, 2000b. Artificial neural networks in hydrology. II: Hydrologic Applications. Journal of hydrologic Engineering, 5(2), 124-137.

Bai, Y., Zeng, B., Li, C., 2016. Daily reservoir inflow forecasting using multiscale deep feature learning with hybrid models. Journal of Hydrology, 532, 193-206.

Crawford, N. H., R. S. Linsley,

CRED, 2016. CRED Crunch. Disaster Data: A Balanced Perspective, 41. cred.be/sites/default/files/CredCrunch41.pdf (last accessed 15.10.2017).

Damle, C., Yalcin, A. 2007. Flood prediction using Time Series Data Mining. Journal of Hydrology, 333 (2-4), 305-316.

Erechtchoukova, M. G., Khaiter, P. A., Saffarpour, S., 2016. Short-Term Predictions of Hydrological Events on an Urbanized Watershed Using Supervised Classification. Water Resource Management, 30, 4329-4343.

Furquim, G. et al., 2014. Combining Wireless Sensor Networks and Machine Learning for Flash Flood Nowcasting. 28th International Conference on Advanced Information Networking and Applications Workshops. Victoria, BC, 13-16 May 2014. IEEE. pp. 67-72.

Galka, A., 2000. Topics in nonlinear time series analysis: with implications for EEG analysis. World Scientific. 14. World Scientific.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explor. 11 (1).

Han, D., Cluckie, I. D., Karbassioun, D. Lawry, J., Krauskopf, B. 2002. River flow modelling using fuzzy decision trees. Water Resource Management. 16. 431-445.

Hu, T. S., Lam, K. C., Ng, S. T., 2001. River flow time series prediction with a range-dependent neural network. Hydrological Sciences Journal, 46 (5), 729-745.

Kingston, G.B., Lambert, M.F., Maier, H.R., 2005. Bayesian training of artificial neural networks used for water resources modeling. Water Resources Research, 41, W12409, doi:10.1029/2005WR004152

Kuhn, M., 2017. Data Splitting. http://topepo.github.io/caret/data-splitting.html (last accessed 18.03.2018).

McCulloch, D. R., Lawry, J., Cluckie, I.D., 2008. Real-time flood forecasting using updateable linguistic decision trees. IEEE International Conference on Fuzzy Systems 2008, Hong Kong, China, IEEE: pp. 1935-1942.

Nayak, P.C., Sudheer, K.P., Rangan, D.P., Ramasastri, K.S., 2005. Short-term flood forecasting with a neurofuzzy model. Water Resources Research 41, W04004. http://dx.doi.org/10.1029/2004WR003562.

Plate, E. J., 2002. Flood risk and flood management. Journal of Hydrology, 267 (1-2), pp. 2-11.

Savic, D. A., Walters G. A., Davidson J.W., 1999. A genetic programming approach to rainfall-runoff modelling. Water Resource Management, 13, 219-231.

Segretier, W., Clergue, M., Collard, M., Izquierdo, L., 2012. An evolutionary data mining approach on hydrological data with classifier juries. IEEE World Congress on Computational Intelligence. Brisbane, Australia, June 10-15, 2012. IEEE. pp. 844-851.

Segretier, W., Collard, M., Clergue, M., 2013. Evolutionary predictive modelling for flash floods. Congress on Evolutionary Computation. Cancun, Mexico, 20-23 June 2013. IEEE. pp. 844-851.

Solomatine, D. P., Dulal, K. N., 2003. Model trees as an alternative to neural networks in rainfall—runoff modelling. Hydrological Sciences Journal, 48 (3), 399-411.

Tokar, A.S., Johnson, P.A., 1999. Rainfall-runoff modelling using artificial neural networks. Journal of Hydrologic Engineering, 4 (3), 232-239.

TRCA (2006) Etobicoke-Mimico watersheds coalition briefing book. http://www.trca.on.ca/dotAsset/159240.pdf (last accessed 29.10.2017).

United Nations, 2004. Living with Risk - A global review of disaster reduction initiatives. 1. https://www.unisdr.org/files/657_lwr1.pdf (last accessed 16.10.2017).

Whigham, P.A., Crapper, P.F., 1999. Time series modelling using genetic programming: An application to rainfall-runoff models. Advances in genetic programming, 3 (5), 89-104.

Young, P. ,2003. Top-down and data-based mechanistic modelling of rainfall–flow dynamics at the catchment scale. Hydrological processes, 17 (11), 2195-2217.