22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 3 to 8 December 2017
mssanz.org.au/modsim2017

Role of a 'combination rule' in hybrid short-term prediction of hydrological events

M.G. Erechtchoukova[a], P.A. Khaiter[a], M. Ditmans[b] and D. Khaiter[c]

[a] *School of Information Technology, Faculty of Liberal Arts and Professional Studies, York University, Canada*
[b] *Faculy of Science, York University, Canada*
[c] *Faculty of Health, York University, Canada*
Email: marina@yorku.ca

Abstract: Data-driven hydrological predictions based on supervised classification have recently gained momentum. This technique supports the classification of waterbodies and flood events that occur at different watersheds, predictions of a class of a hydrological event, e.g., 'high-' or 'low-flow', as opposed to forecasting magnitudes of streamflow characteristics generated by ANNs, regression models or other modelling tools. Flood management teams declare a state of emergency and/or take mitigation measures based on a set of business rules reflecting water level exceedance of an established threshold. Therefore, predicting a class of a hydrological event, e.g. 'flood' or 'no-flood', carries even more important information for operational flood managers than projected magnitudes of streamflow characteristics. When predictions of a class of an event are obtained based on data available in real-time, they can be easily deployed in flood management. Scientific literature has demonstrated the usefulness of various classification algorithms (inducers) in applied hydrology. The performance of these inducers, however, deviated notably on different data sets. To alleviate these deviations and generate forecasts with reduced generalization error, an ensemble of classifier can be constructed.

One of the important steps in developing an ensemble of classifiers is identifying the approach to aggregate individual predictions into a final judgement. The current study investigates the effect of various weighting schemes on the accuracy of the generated forecasts of hydrological events. The predictors were developed using C4.5, CART, REPTree, NBTree, Ridor, JRip, and Random Forest inducers trained on data collected by stream and rain gauges located on a small highly urbanized watershed during two hydrologically distinct years. The data sets were first transformed into time series of various granularity from 15 minutes to 60 minutes. Time series of the same granularity and corresponding to the same year were converted to an augmented phase space providing datasets for training and testing developed predictors. Ensembles were constructed using five combination rules: majority vote, maximum probability, minimum probability, average probability, and product of probabilities. The ensemble's generalization error was estimated using two measures: recall and *F*-score.

Combining the results of predictors constructed via training of individual inducers allows to develop a more robust model generating reliable predictions. However, the estimates of the ensemble's generalization error vary up to 28% depending on the combination rule used to aggregate individual predictions into the final judgement. The issue of selecting a combination rule which is the most suitable for an application domain has both theoretical importance and practical significance. Computational experiments revealed that the classifier constructed with the minimum probability combination rule outperformed the others. It consistently delivered the most accurate results for all investigated data sets and all lead time intervals. The performance of classifiers utilizing the maximum probability rule on all data sets was the weakest, contrasting to its interpretation as a rule which identifies a classifier with the highest estimated confidence. Although the results of data-driven analysis are site-specific, they suggest further investigation of this rule, including theoretical considerations and application of the rule to data sets from other watersheds. Another combination rule which should not be easily discarded for the given problem domain, is the majority vote.

Keywords: *Supervised classification, hydrological event, ensemble of classifiers, combination rule, short-term prediction*

1. INTRODUCTION

The supervised machine learning approach has been employed in hydrological modelling for several decades. Starting from regression problems where relationships between factors affecting natural conditions of a watershed and targeted hydrological characteristics of a corresponding waterbody were described as a black box, the approach delivered a large variety of artificial neural networks (ANNs) applied to different problems of water resources assessment and management. This artificial intelligence technique became very popular, so the American Society of Civil Engineering (ASCE) summarized the experience and provided recommendations on ANN applications in two papers ASCE (2000a) and ASCE (2000b). Since then, ANNs have evolved dramatically into Bayesian artificial neural networks (e.g., Humphrey et al., 2016), neuro-fuzzy systems (Nayak et al., 2005), or deep learning tools (e.g., Li et al., 2016) supporting stream-flow predictions of various temporal scales. Regression methods were incorporated into decision trees, e.g., M5 algorithms (Quinlan, 1992), and when applied to large volumes of data, became one of the branches of supervised machine learning. There is a growing number of publications on the application of other branches of artificial intelligence to environmental problems and, particularly, to the issues related to water resources management. The comparison of machine learning algorithms with respect to their hydrological applications was undertaken by Londhe and Charhae (2010) and Spate et al. (2003).

Data-driven hydrological analysis based on supervised classification have recently gained momentum. The technique supports the classification of waterbodies (Hewett, 2003) and flood events that occur at different watersheds (Sikorska et al., 2015), predictions of a class of a hydrological event, e.g., 'high-' or 'low-flow', as opposed to forecasting magnitudes of streamflow characteristics generated by ANNs regression models or other modelling tools (Damle and Yalcin, 2007). McColloch et al. (2008) and Erechtchoukova et al. (2016) proved usefulness of various classification algorithms (inducers) in hydrological predictions. At the same time, the studies demonstrated the compliance with the 'no free lunch' theorem (Wolpert, 1996), which states that there is no *a priori* superiority of the inducer, even for the same problem domain. The performance of the inducers deviated notably on different data sets. To alleviate these deviations and generate forecasts with reduced generalization error, an ensemble of classifier can be constructed.

Optiz and Maclin (1999) and, more recently, Rokach (2010) provided extensive reviews of ensemble methods, suggesting where application of ensembles can be beneficial and offering general guidance on ensemble development. The main selection criteria for ensemble members can be summarized in the following way: (1) to preserve the 'diversity of opinion'; (2) to maintain independence of individual members; (3) to allow classifiers to draw conclusions using specific knowledge (i.e., decentralization); and (4) to aggregate private judgments into the final decision (Rokach, 2010). There are two general ways to combine the results of individual predictors into a final judgement: weighting schemes and meta-learning techniques. Meta-learning methods train inducers on the results of classifiers constructed using data sets. Weighting methods assign fractions to each member's opinion which are used to calculate the final decision. The weights can be static and assigned *a priori* or can be calculated dynamically depending on the classifier's performance.

The study was focused on comparison of weighting schemes with the goal to determine the most suitable combination rule for short-term prediction of high-flow events in small watersheds. The individual classifiers were constructed by training supervised classification algorithms on the augmented phase space following formal problem articulation, and the framework for data pre-processing was described earlier (Erechtchoukova et al., 2016). The paper presents the summary of the problem articulation and framework for reconstruction of the phase space, sets of computational experiments conducted and the results of the comparison of investigated combination rules.

2. MODELLING TOOL

To declare a state of emergency, flood management teams base their operational decisions regarding a flood event on a set of business rules reflecting water level exceedance of an established threshold. Therefore, predicting a class of a hydrological event, e.g. 'flood' or 'no-flood', carries even more important information for operational flood managers than projected magnitudes of streamflow characteristics. When predictions of a class of an event are obtained based on data available in real-time, they can be easily deployed in flood management. The formal articulation of the problem of a hydrological event prediction using supervised classification has been presented in (Erechtchoukova et al., 2016). The predictors have been developed using various inducers and their ensemble trained on data collected by stream and rain gauges located on a watershed of interest. Such data are usually collected automatically with high frequencies of observations. To apply classification algorithms, the time series generated by stream and rain gauges must be transformed into a phase space augmented by class labels corresponding to the class of an event occurring at the cross-section of interest. Given that the near future hydrological conditions at the cross-section of interest depend on the current stream

state, upstream hydrological conditions and overall watershed meteorological conditions, the phase space has been re-constructed from time series obtained by all meteorological observation sites located on a watershed as well as hydrological sites located upstream of a cross-section of interest. In addition to that, a current state of a stream is predetermined by the conditions in the recent past. To account for this, a time delay embedding approach (Povinelli and Feng, 2003) was applied.

In this study, seven inducers were used to investigate the effect of a combination rule on the uncertainty of short-term predictions of flood-events in a small highly urbanized watershed. The set of inducers contained well-known machine learning algorithms: C4.5, CART, REPTree, NBTree, Ridor, JRip, and Random Forest implemented in WEKA 3.6.14 software package (Hall et al., 2009). The details of their implementation can be found in the corresponding manual. All individual classifiers produced class labels representing categorical values which can be interpreted as a type of a hydrological event, e.g., a 'high-flow' or 'low-flow' event.

3. COMBINATION RULES

It is obvious that an ensemble of classifiers performs better if its members disagree on some elements of a phase space. One of the most popular combination rules is the majority vote where the final judgement corresponds to the label most frequently assigned to a classified element of the phase space:

$$\widetilde{f}(z_t) = \underset{cl \in \{'high','low'\}}{\arg\max} \left(\sum_{j=1}^{J} g(f_j(z_t), cl) \right), \quad g(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \tag{1}$$

where *cl* is the class label, f_j is the event characterization function ascertaining the class label assigned to the tuple of input variables z_t by *j*-th classifier and *g* is an indicator function (Rokach, 2010).

Previous studies of the above inducers showed that they produce classifiers which perform differently on the data sets reconstructed from hydro-meteorological data. Some of them predicted 'high-flow' events more accurately than the others and *vice versa*. The performance of classifiers generated by different inducers varied notably on different data sets from the same watershed (Erechtchoukova et al., 2016). These results prompted investigation of the performance-based weighting schemes for aggregation of the predictions of individual classifiers. The 'no free lunch' theorem (Wolpert, 1996) implies the necessity to evaluate the performance of classifiers on a case-by-case basis. Therefore, four other combination rules derived from probabilistic schemes were studied: the average of probabilities, product of probabilities, minimum probability, and maximum probability. Formal definition of these rules was presented in (Duin and Tax, 2000):

$$\tilde{f}(z_t) = \arg\max\left(\underset{j=1,\ldots,J}{rule} \left(p_j(x_t)\right)\right), \tag{2}$$

where *rule* is one of the functions from the list {maximum, minimum, average, product}, p_j is the *posteriori* probability of correct classification of the *j-th* classifier, $j = 1,\ldots,J$.

These rules have been investigated by Duin and Tax (2000) and Kittler et al. (1998) with respect to their application to pattern analysis and hand-written text recognition. In this study, the rules are applied to classify elements of the phase space re-constructed from hydrological and meteorological time series using time delay embedding.

4. PERFORMANCE EVALUATION

There are several empirical estimates of a generalization error of a classifier. The following considerations were taken into account for the selection of ensembles' performance measures. Naturally, the low- and mid-flow events dominate high-flow events and particularly flood events on many urbanized watershed. Therefore, hydrological and meteorological data sets are imbalanced. From operational management perspectives, the accurate prediction of high-flow events is extremely important to be able to issue alerts and to undertake mitigation measures where possible. Although misclassification of low-flow events imposes economic and social burden, it is less dangerous and implies that measures reflecting the performance of a classifier on a minority class are more informative for the given problem. Two measures better reflect the performance of classifiers on a minority class: (1) the recall (or sensitivity), which is the ratio of correctly identified high-flow events to their total number expressing how well a classifier recognizes such events; and (2) the precision of a classifier determined as the ratio of correctly identified high-flow events to the total number of elements classified as high-flow events. The relationship between these two measures, however, is inversely proportional

making the selection of the best rule using both recall and precision problematic. There is an aggregate measure of a classifier performance – *F*-score:

$$F = \frac{2TP}{FP+FN+2TP}, \quad (3)$$

where *TP* is true positive or the number of high-flow elements classified correctly, *FP* and *FN* are the numbers of false positive and false negative elements, respectively. Recall and *F*-score were evaluated on testing sets containing elements of the phase space not included into the training sets for developing classifiers.

5. DATA SETS

The combination rules described in section 3 were tested on the data sets reconstructed from time series of water levels and precipitation collected at the Spring Creek watershed, Ontario, Canada, over a two year period from 2013 to 2014 by the Toronto and Region Conservation Authority (TRCA). The watershed was chosen due to its 'flashy' response to intensive precipitation during the warm season from April to November. The watershed is classified as small with an area of about 50 km^2. The Spring Creek flows over 25km through the highly urbanized and populated region which is composed of approximately 70% urbanized areas, 25% rural lands and 5% natural cover (TRCA, 2006). The stream daily average baseflow estimates are close to 0.20m^3/s. The study was conducted for the hydrologically wet year, 2013, when 750mm of total rainfall was recorded during the warm period, the average instantaneous water discharge was approximately 0.64m^3/s and the total annual water discharge was 0.02km^3. During the dry year, 2014, the corresponding characteristics were estimated as 539mm of the total precipitation, 0.47m^3/s for the instantaneous water discharge and 0.015km^3 as the total annual water discharge (Figure 1).

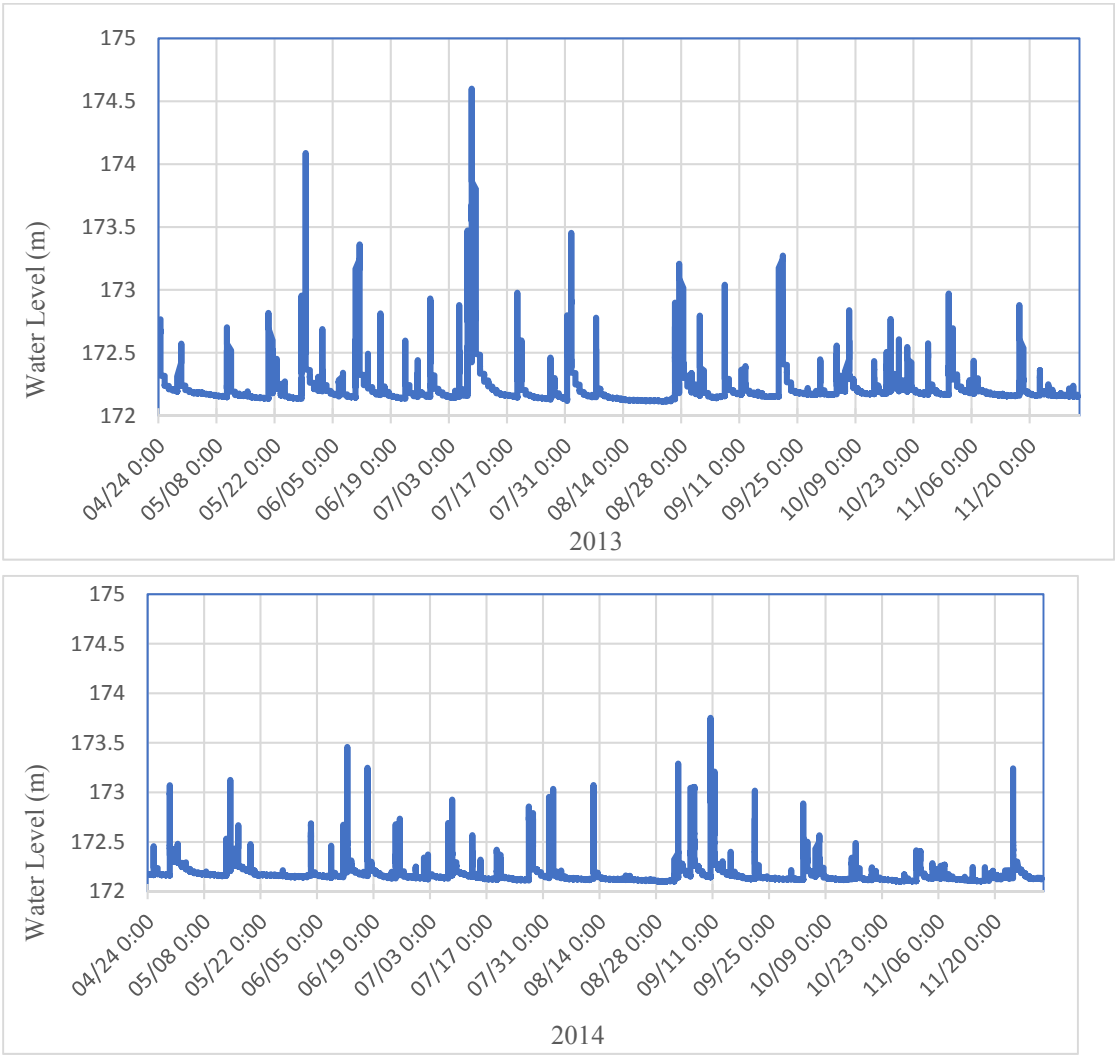

Figure 1. Hydrographs at the closing cross section of Spring Creek, Ontario, Canada.

Two stream gauges and two rain gauges are installed at the watershed. The downstream flow gauge was used at a cross section of interest where hydrological events were predicted. Time series from all four gauges were synchronized and transformed to the same level of granularity of 15 minutes. Time series of 30 minutes and one hour were obtained from the original data by way of aggregation.

These time series were used to reconstruct elements of the phase spaces applying the framework summarized in section 2. These elements can be defined by the following formula:

$$X(t+j\tau) = Y_1(t-1), Y_1(t-2), \dots, Y_1(t-K\tau), \dots, Y_M(t-1), Y_M(t-2), \dots, Y_M(t-K\tau), Class(t+j\tau), \quad (4)$$

where $X(t+j\tau)$ is the element of the phase space built to generate predictions with $j\tau$ lead time, $j = 1,\dots, K$, $Y_i(t)$ is the instantaneous measurement from *i-th* gauge at time t, $i = 1,\dots,M$, $Class(t)$ is the class label of an event at the investigated cross-section at the time t. The class label is determined by existing business rules set at the watershed by the management authority, and the underlying threshold value is obviously site-specific.

Each augmented phase space was split into two parts with two thirds of the elements belonging to the set used to train a classifier and the rest of elements reserved for testing.

6. RESULTS AND DISCUSSION

Six groups of experiments were conducted on data sets with three different levels of granularity and for two hydrologically distinct years to examine the effect of the combination rule on the final outcome of predictions using the WEKA software package (Hall et al., 2009). Each group of computations included performance evaluation of an ensemble constructed out of seven heterogeneous classifiers based on five investigated combination rules for several prediction lead time intervals. From a probabilistic point of view, the combination rules can be justified if they are applied to independent classifiers. Although in this study the relationships between classifiers outputs were not tested statistically, their independence is supported by applying different inducers. Each group of experiments used data of the same granularity from the same year. The recall and *F*-score reflecting the performance of each trained classifier were evaluated on records unseen during the training step. The results of computational experiments presented in Figure 2 and Table 1 showed notable difference in the accuracy of generated predictions.

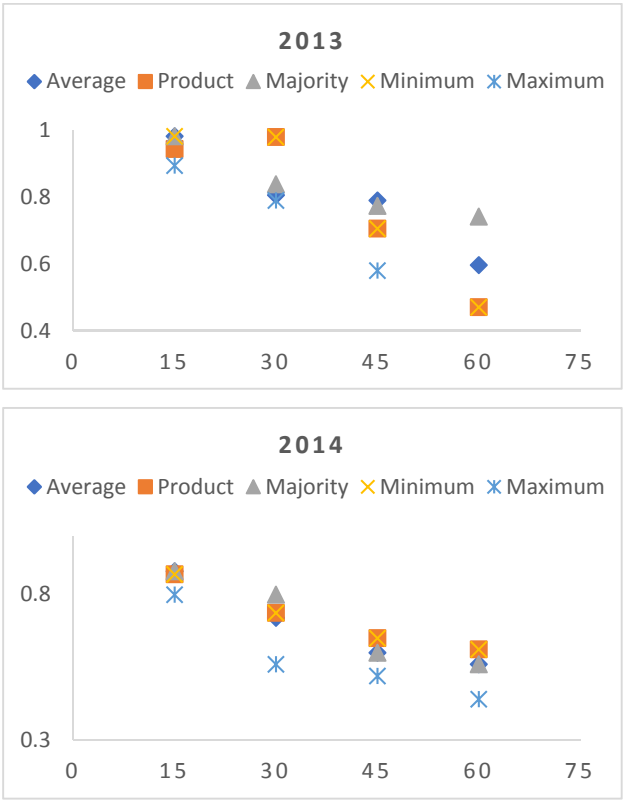

Figure 2. Recall of ensembles vs. prediction lead time developed on 15-minute time series.

Both estimates of a classifier's generalization error confirmed that accuracy of predictions declines with increasing lead time intervals. The deviations in performance of ensembles constructed with different combination rules rose with extended lead time intervals.

The classifier constructed with the minimum probability combination rule outperformed the others. It consistently delivered the most accurate results for all investigated data sets and all lead time intervals following its interpretation by Duin and Tax (2000) that the rule selects the least objected opinion.

Two combination rules, namely, minimum probability and product of probabilities delivered the same results on all investigated datasets. This result coincides with the evaluations of combination rules for non-precise observations where the product rule determines a subset of elements identified by the minimum rule.

The performance of classifiers utilizing maximum probability rule on all data sets was the weakest, contrasting to its interpretation as a rule which identifies a classifier with the highest estimated confidence. For the extended lead time intervals, the *F*-score of predictors with this rule was below the best score by 26-28%.

Table 1. F-score of developed classifiers

Granularity	Year	Combination rule	Lead time			
			30 min	60 min	90 min	120 min
30-minute	2013	Majority	0.9016	0.7857	0.6916	
		Average	0.9091	0.7676	0.6733	
		Product	0.9231	0.7381	0.6744	
		Minimum	0.9231	0.7381	0.6744	
		Maximum	0.8421	0.6078	0.5859	
	2014	Majority	0.8800	0.7727	0.6667	
		Average	0.8800	0.7727	0.6667	
		Product	0.9744	0.8125	0.7407	
		Minimum	0.9744	0.8125	0.7407	
		Maximum	0.8444	0.6667	0.5556	
1-hour	2013	Majority		0.8372		0.7592
		Average		0.8438		0.7358
		Product		0.8376		0.7586
		Minimum		0.8376		0.7586

The majority vote and the average probability rule produced very similar results on the hydrological data sets. The predictor with the average probability rule aiming at reducing the classification error outperformed other classifiers only once, suggesting that this rule is acceptable, but not the most suitable for the given problem domain. The same can be claimed for the majority vote combination rule with one reservation. This rule is easy to implement. It allows for a straightforward interpretation which is important for decision making in applied problems.

7. CONCLUSION

Classifiers can be useful for operational flood management in highly urbanized watershed equipped with stream and rain gauges. Combining the results of predictors constructed via training of individual inducers allows development of a more robust model that generates reliable predictions. However, the estimates of an ensemble's generalization error vary up to 28% depending on the combination rule used to aggregate the individual predictions into the final judgement. The issue of selecting a combination rule which is the most suitable for an application domain has both theoretical importance and practical significance. Investigation of the combination rules was previously done mainly for areas of pattern recognition. In this study, the classification algorithms were applied to find patterns in data reconstructed from time series and predict future hydrological events. This explains the disagreement of the results with other investigations of combination rules.

The study was conducted on data collected at a single watershed during two hydrologically distinct years using original and aggregated time series of different granularity. Therefore, the results allow for generalized conclusions. The minimum probability rule is the most suitable rule for both wet and dry hydrological years and data of granularity between 15- and 60-minute intervals. Although the results of data-driven analysis are site-specific, they suggest further investigation of this rule, including theoretical considerations and application of the rule to data sets from other watersheds. Another combination rule which should not be easily discarded for the given problem domain is the majority vote.

ACKNOWLEDGMENT

The analysis was implemented on the data collected by the Toronto and Region Conservation Authority. The authors thank the TRCA and, especially, J. Duncan and J. Cao for providing the data and necessary clarifications. The authors are grateful to editors and anonymous reviewers for their thoughtful suggestions and helpful comments on the manuscript.

REFERENCES

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000a). Artificial Neural Networks in hydrology. I: preliminary concepts. *Journal of Hydrologic Engineering*, 5(2), 115-123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000b). Artificial Neural Networks in Hydrology. II: hydrologic applications. *Journal of Hydrologic Engineering*, 5(2), 124-37.

Damle, C., and Yalcin, A. (2007). Flood predicting using time series data mining. *Journal of Hydrology*, 333, 305-316.

Duin, R.P.W., and Tax, D.M.J. (2000). Experiments with classifiers combining rules. Kittler, J., and Roli, F. (Eds.), MCS 2000, LNCS 1857, 16-29. Springer-Verlag, Berlin, Heidelberg.

Erechtchoukova, M.G., Khaiter, P.A., and Saffarpour, S. (2016). Short-term predictions of hydrological events on an urbanized watershed using supervised classification. *Water Resources Management*, 30(12), 4329-4343.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Hewett, R. (2003). Data mining for generating predictive models of local hydrology. *Applied Intelligence*, 19, 157-170.

Humphrey, G.B., Gibbs, M.S., Dandy, G.C., and Maier, H.R. (2016). A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. *Journal of Hydrology*, 540, 623-640.

Kittler, J., Hatef, M., Duin, R.P.W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), 226-239.

Li, C., Bai, Y., and Zeng, B. (2016). Deep feature learning architectures for daily reservoir inflow forecasting. *Water Resources Management*, 30 (14), 5145-5161.

Londhe, S., and Charhate, S. (2010) Comparison of data-driven modelling techniques for river flow forecasting. *Hydrological Science Journal*, 55(7), 1163–1174.

McCulloch, D.R., Lawry, J., and Cluckie, I.D. (2008). Real-time forecasting using updateable linguistic decision trees. *Fuzzy Systems*, 1935-1942.

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

Opitz, D., and Maclin, R. (1999). Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, 11, 169-198.

Povinelli R.J., and Feng, X. (2003) A new temporal pattern identification method for characterization and prediction of complex time series events. *IEEE Transactions on Knowledge and Data Engineering* 15(2), 339–352.

Quinlan, J.R. (1992). Learning with continuous classes. Proc. AI'92, World Scientific, Singapore, 343-348.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33, 1 – 39.

Sikorska, A., Viviroli, D. and Seibert, J. (2015). Flood-type classification in mountainous catchments using crisp and fuzzy decision trees. Water Resources Research, 51(10), 7959-7976.

Spate, J, Croke, B, and Jakeman, A. (2003). Data mining in hydrology. International Congress on Modelling and Simulation (MODSIM 2003), MSSAZ Inc., UWA UniPrint, 422-427.

Toronto and Region Conservation Authority (TRCA) (2006) Etobicoke-Mimico watersheds coalition briefing book. http://www.trca.on.ca/dotAsset/159240.pdf (accessed Feb. 27, 2015).

Wolpert, D.H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7), 1341-1390.