Short-term prediction of flood events in a small urbanized watershed using multi-year hydrological records

S. Saffarpour[a], M.G. Erechtchoukova[a], P.A. Khaiter[a], S.Y. Chen[a] and M. Heralall[b]

[a] School of Information Technology, Faculty of Liberal Arts and Professional Studies, York University, Toronto, Canada
[b] Toronto and Region Conservation Authority(TRCA), Downsview, Toronto, Canada
Email:marina@yorku.ca

Abstract: The application of various artificial intelligence techniques to short-term flood prediction shows that the majority of low-flow events are predicted accurately while the forecasting of high-flow events has a high level of uncertainty. The differences in forecast accuracy can be attributed to the imbalance of water flow data which mainly represent low-flow events with fewer measurements corresponding to flood events. Furthermore, the hydrological conditions of a watershed vary from year to year, so that the annual occurrences of high-flow events fluctuate notably from year to year. As a result, for some years with very few high-flow events the predictors cannot be trained. In data-driven modeling, the quality of a predictor depends on both the applied algorithm and the data set used for training it. To address the issues of imbalanced data in application of supervised machine learning algorithms to short-term flood prediction, we investigated approaches to constructing training data sets which could improve the prediction ability of classification algorithms. The approaches were examined on the hourly records of rainfall and water levels from April to December for the years of 2008, 2011, 2012 and 2013 in the highly urbanized Spring Creek watershed, Ontario, Canada. The Spring Creek is a small stream with a baseflow of about 0.20 m^3/s. Its watershed has an area of 50 km^2. A constructed ensemble of five classifiers, C4.5, CART, RepTree, NNge, and JRip, has provided very accurate predictions of low-flow events and demonstrated satisfactory performance for high-flow events for one-hour lead time in some years while for years with fewer high-flow events the prediction accuracy declined. In this study, we explored three approaches including: 1) maintaining low-flows of each year and oversampling high-flow data for the same year; 2) training the ensemble of classifiers using the wet year; and 3) maintaining low-flows of each year and combining high flows of multiple years. Two performance measures: F-score and misclassification rate were employed to compare the results of the ensemble on various data sets. The performance of the ensemble of classifiers on the raw data set was used as a benchmark. We ran the experiments for up to a two-hour lead time. Some of the classification algorithms in the given study worked better on non-flood events while others predicted high-flow events more accurately. From a management perspective, an improvement in the prediction of high-flow events is more important than that of low-flows, however, misclassification of low-flow events also has some negative effects. Therefore, the developed models have to improve both high- and low-flow predictions and minimize the number of misclassified tuples. The ensemble of classifiers produced the most accurate predictions for a one-hour lead time for all of the investigated years. Although oversampling is typically recommended for dealing with imbalanced data, the results of the study did not support this suggestion. Applying the ensemble of the classifiers on the oversampled data set increased the number of misclassified tuples. For all of the investigated wet years, the best predictions were generated by applying the ensemble on the combined data sets. For the dry year, the prediction results of ensemble on the combined data set were satisfactory but still lower compared to the wet years. The results suggest that the combined approach can be used for short-term hydrological forecasting. The developed classifier does not require a resource-consuming model parameterization. Its site-specific calibration is implemented at the training step using data readily available. It makes classification algorithms very attractive as a tool for flood management in small watersheds. However, training classifiers on real-time data at the beginning of the wet season may not be possible because tuples for flood events may not be available. The proposed combined approach may help to address this issue, and hydrological events can be predicted using historical records. The development of such a framework is the subject of further investigation.

Keywords: *Flood prediction, classification algorithms, imbalanced data, oversampling, combined approach*

1. INTRODUCTION

From a management perspective it is essential to predict flood events accurately and to issue informed and reliable flood warnings in a timely manner. Often in highly urbanized areas, floods can develop very quickly. Rivers swell rapidly with stormwater and submerge adjacent areas causing emergency conditions, damaging infrastructure and the environment, and increasing human casualties around the world (Sandford and Freek, 2014). Two main approaches are used for flow prediction: (1) process-based models and (2) data-driven techniques. While physical and process-based models have been commonly used in numerous flood studies (e.g., Feldman, 2000; Horritt and Bates, 2002; Vieux et al., 2004; Hunter et al., 2008), they have some drawbacks. First of all, these models employ a large number of parameters which require site-specific calibration. The process is computationally expensive and is based on detailed data collection on watershed conditions. The majority of available flood models predict seasonal variations of water levels and/or discharges, but they fail to predict the very rapid increases of water flow typical of flash floods. At the same time, it is important for practical reasons to know whether the high-flow in a stream will develop into a flood or not, i.e., to predict an event – flood or non-flood.

Data-driven techniques include classification algorithms (e.g. decision trees, Bayesian networks, rule-based approaches, neural network algorithms and support vector machines), cluster analysis, and association rule mining (Han et al., 2012). Classification algorithms work by categorizing elements into several classes. With respect to the flood prediction problem, hydrological and meteorological data on a watershed can be transformed into elements of a phase space such that each element corresponds to exactly one event which can be either flood or non-flood. Examples of this approach to flood prediction are found in papers of McCulloch et al. (2008), Segretier et al. (2012, 2013). In this approach, a model for flood prediction (predictor) is constructed by applying an appropriate classification algorithm (inducer) to the phase space. The quality of the developed predictor depends on the selected inducer and the data set used as a training set. This approach was applied to the Spring Creek urbanized watershed located in the Greater Toronto Area, Canada to generate predictions of flood events. An ensemble of five classifiers which combine the C4.5, CART, RepTree, NNge, and JRip algorithms was constructed in the WEKA software package (Hall et al., 2009). The ensemble predicted the majority of low-flow events accurately while the forecasting of high-flow events had a high level of uncertainty. The differences in forecast accuracy can be attributed to the imbalance of water flow data which mainly represent low-flow events with fewer measurements corresponding to flood events. In addition, the hydrological conditions of a watershed vary from year to year, so that the annual occurrences of high-flow events fluctuate notably from year to year. As a result, for some years with very few high-flow events the predictors cannot be trained.

The aim of this study was to develop an approach to constructing training data sets which could improve the prediction results of the ensemble. Two main approaches are used in data mining to deal with the problem of imbalanced data: (1) oversampling to increase the number of elements from the minority class (high-flows) by repeatedly adding tuples of this class to the training sets and (2) undersampling to reduce the dominance of elements representing the majority class (low-flows) by randomly deleting them from the training sets (Han et al., 2012). Undersampling of low-flow events increased the number of misclassified records for the low-flows. We explored different methods of constructing data sets and compared the performance of the developed classifiers to determine how the chosen methods affect the prediction of flood events in different years. Three approaches have been tested: (1) oversampling of high-flow events for the same year; (2) building the ensemble on the data for a wet year; and (3) combining all historical records corresponding to flood events in the training data set. The preferable approach is to combine all known records from different years for flood events into the training set along with the tuples for non-flood events for a given year. Although this approach has limitations, it improves the overall predictions on the majority of models and lead time intervals.

2. CASE STUDY: THE SPRING CREEK WATERSHED

The Spring Creek watershed is a sub-watershed of Etobicoke Creek in the West part of the Greater Toronto Area, Canada. The watershed has an area of 50 km^2 and a slope below 5%. It is highly urbanized and populated (Figure 1). Historically, the highest annual discharge in the Etobicoke watershed was observed at the beginning of spring, but the watershed flow regime has changed owing to over 40 years of urbanization. Recently, summer flash floods have became more common whereby water levels can rise up to 1 m in a short time (TRCA, 2011).

This study was carried out using the hourly data of rainfall and water levels in the Spring Creek watershed during the warm months of April to December. There were four monitoring sites in the watershed including two stream gauges and two precipitation gauges but some of them were periodically removed. Only two of the sensors (one rain gauge and one stream gauge) had hourly records for 2008, 2011, 2012, and 2013. Therefore, in order to be able to compare the performance of the developed classifiers on different data sets, only data from sensors A (rainfall) and D (water level) were considered.

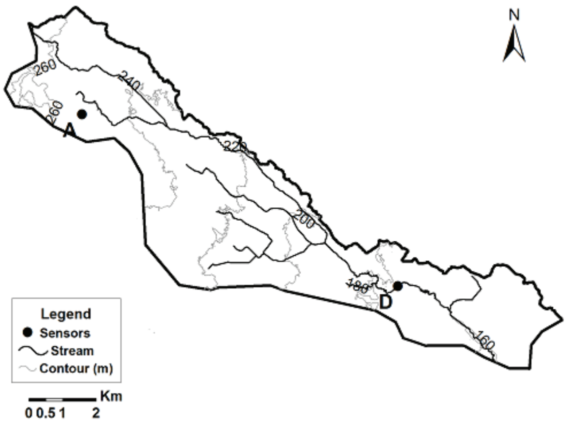

Figure 1. Spring Creek watershed and the sensors

3. PRELIMINARY ANALYSIS

The Spring Creek has low baseflows (daily average estimates are close to 0.20 m^3/s) but water levels could rise rapidly after rainfall events. Since the hydrological conditions of the watershed vary notably from year to year, the considered four years include hydrologically wet and dry years (Table 1). The total rainfall at sensor A from April to December varied between years. The lowest total rainfall of 597 mm was registered in 2012 and the highest total amount of rain of 749 mm was recorded in 2013. According to Toronto Region Conservation Authority's synthesis report (TRCA, 2008), it is estimated that in a normal year the Etobicoke Creek watershed receives an average annual precipitation of about 800 mm including approximately 115 mm mean annual snow and 685 mm mean annual rainfall. Comparison of the total amount of rainfall from senor A to the watershed normal condition (685 mm) suggests that 2012 can be considered as a dry year, while 2008, 2011, and 2013 were wet years. In addition, the mean annual instantaneous discharge and the total annual flow in the Spring Creek fluctuated significantly. The lowest mean annual instantaneous discharge of 0.44 m^3/s was in 2012 and the highest mean annual instantaneous discharge of 0.76 m^3/s was in 2011. The lowest total annual flow in the Spring Creek was observed for the year of 2012 when it dropped by more than 40% compared to 2011. These data are supported by the total annual flow data obtained from the Water Office Canada in the Etobicoke Creek (Environment Canada, 2015) which shows that the total annual flow dropped by about 60% in 2012 compared to other years. Based on the total rainfall amount of 749 mm at sensor A and the total annual flow of 20498400 m^3 in 2013, this year was identified as a wet year. As the total rainfall was 597 mm at sensor A and the total annual flow was 13819757 m^3 in 2012, this year was considered as a dry year. Therefore, the four investigated years are characterized by different hydrological conditions observed in the watershed.

Table 1. The total rainfall at sensor A and the total annual flow in Sensor D in the Spring Creek and total annual flow in the Etobicoke Creek at Brampton

Year	Total rainfall from April to December (mm)- Spring Creek*	Maximum hourly rainfall intensity from April to December (mm/hr) - Spring Creek	Mean annual instantaneous flow during the water year (m^3/s) - Spring Creek	Total annual flow during the water year (m^3) - Spring Creek	Etobicoke Creek at Brampton-total flow in dam^3
2008	715.4	13.6	0.58	18279481	31600
2011	729.6	24.2	0.76	23829795	26900
2012	597.4	22.4	0.44	13819757	13900
2013	749.4	47.8	0.65	20498400	30200

4. DATA PRE-PROCESSING

Data sets were constructed using up to 5 hours of hourly data of rainfall and water level at sensors A and D, respectively. The following expressions show the structure of a tuple used for constructing the data sets for hourly predictions:

For one-hour lead time: *A (t - 1) to A (t - 5), D (t - 1) to D (t - 5), Class label (t)*

For two-hours lead time: *A (t - 2) to A (t - 6), D (t - 2) to D (t - 6), Class label (t)* , (1)

where *A* (*t-j*) is the rainfall amount received from senor *A* at *j* hours preceding an event, *D* (*t-j*) is the water level measured at senor *D* at *j* hours preceding the same event and *Class label* (*t*) stands for the class label of each tuple at the current time *t*.

To apply supervised classification algorithms, class labels are assigned to each tuple. In this study, we assumed that a flood occurs when the discharge at the cross section of interest (i.e., sensor D) reaches 7 m^3/s. The corresponding water level threshold was estimated as 172.75 m. The class labels were determined using the observed water levels at sensor *D* at time *t*. Water levels above or equal to the threshold were labeled as high-flows and those below the threshold as low-flows.

With the assigned threshold for water levels, it was found that the highest number of tuples labeled as high-flow appeared in the data set for 2013, and that the 2012 data set had the lowest number of high-flow tuples. It should be noted that we could not meaningfully lower the water level threshold in order to increase the number of high-flow elements of the phase space so as to alleviate the issue of imbalanced data because the main purpose of the developed model was to predict flood conditions.

5. GROUPS OF EXPERIMENTS

The constructed tuples for each year were split into training and testing sets. Depending on the hydrological conditions of the various years, the size of the training and testing sets varied. The ensemble of classifiers was trained on one set and its performance was evaluated based on unseen tuples from the testing set.

We conducted four groups of experiments. In the first round, we investigated raw data and examined how well a classifier built on a training set of a specific year could predict high- and low-flows of the unseen data for the same year.

In the second group of experiments, we examined how a higher number of high-flows in the training sets affects the predictions of events from the testing sets. We used the common oversampling approach (Han et al., 2012). In order to reduce the effects of highly imbalanced data, for each individual year, we maintained low-flows from the training set and added duplicate tuples labeled as ‘high-flow’ from the same training set. The classifier was built on the training set with repetitions and we evaluated the classifier’s performance on the testing set for the same year. As a result, for each individual year, the testing set in the second group of experiments was exactly the same as the testing sets for the first and fourth groups of experiments.

In the third group of experiments, we checked whether a model built on a data set from one year can be used for other years. In this group of experiments, we assumed that a classifier built on historical data of a wet year could potentially improve the prediction results for other years. We selected data collected in 2013 as a training set because 2013 was a wet year with the highest number of elements labeled as high-flow. The most accurate one-hour predictions exceeding 80% were made based on the data for this year. We trained the classifier using a data set of the entire year of 2013 and tested the built classifier on data for other years. The testing sets on this stage of the study were different from the sets applied in the first, second, and fourth groups of experiments.

In the last group of experiments, for the combined high-flow method, we maintained the training set of each specific year and added all historical records corresponding to flood events from other years to the training set. For this group of experiments, the testing set was exactly the same as in the first and second groups of experiments.

The computational experiments were run for up to a two-hour lead time. Some of the inducers used in the study predicted better non-flood events, while others worked better on the high-flow events. From a practical perspective, the accurate prediction of high-flow events is more important but misclassification of the low-flow events also carries a negative impact so they should be minimized. To compare the performance of the ensembles trained on different data sets, we used two performance measures: F-score and Misclassification rate (Han et al., 2012). F-score was chosen as a main indicator of high-flow predictions. As a combined measure of precision and sensitivity, it decreases to 0 as fewer high-flow tuples from a testing set are identified correctly. An F-score of 100% indicates that all tuples are classified correctly. The F-score is calculated using the following formula:

$$F\text{-}score=\frac{(2*TP)}{(P+TP+FP)}*100\%, \qquad (2)$$

where *TP* or True positive represents the number of correctly classified tuples labeled as high-flows and *FP* or False positive indicates the number of low-flow tuples which are misclassified as high-flows. *P* is the total number of high flow tuples in the testing data set.

The misclassification rate or error rate is an overall combined measure of prediction errors for both high-flows and low-flows. A misclassification rate of 0 shows 100 % accuracy in the prediction results, and it means that all records were classified correctly. The misclassification rate is calculated using the following equation:

$$Misclassification\ rate = \frac{(FP + FN)}{Total\ number\ of\ records} * 100\%, \tag{3}$$

where *FN* or False negative shows the number of high-flow tuples which are misclassified as low-flows.

6. RESULTS

The first group of experiments demonstrated that the ensemble of five classifiers worked well on data sets corresponding to the wet years. The same ensemble failed to provide reliable predictions of flood events during the dry year of 2012 because the training set of data for that year did not have enough tuples corresponding to flood events to identify a pattern. This group of experiments was used as a benchmark for the three other groups.

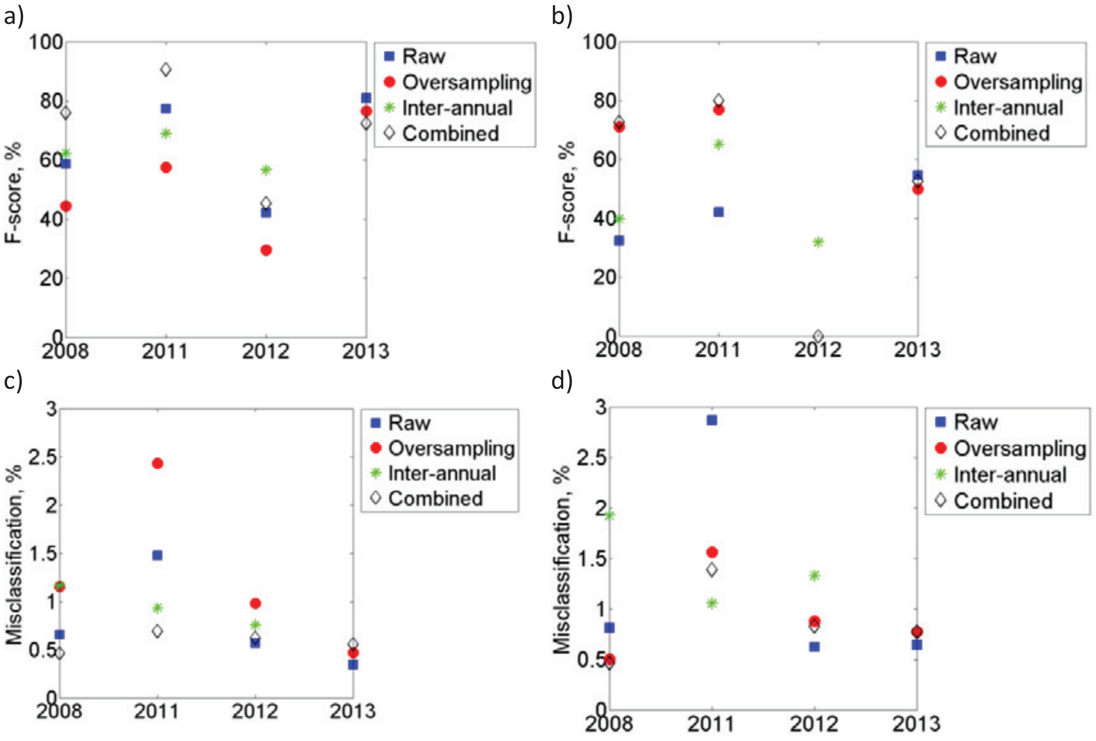

Figure 2. The comparison of up to two-hour performance for all data sets**:**
a) F-score, one-hour lead time; b) F-score, two-hour lead time;
c) Misclassification rate, one-hour lead time; and d) Misclassification rate, two-hour lead time

Figure 2 shows the comparison of performance measures for classifiers trained on all types of data sets for up to a two-hour lead time. The best predictions were made for a one-hour lead time for all investigated years. Increasing the lead time up to two hours reduced the predictive ability of the classifier trained on all data sets except oversampled data. Training the classifier based on the data sets with oversampling did not improve its performance on one-hour lead time, however, it helped to reduce misclassified tuples for two-hour predictions. F-score values indicated that the performance of the classifier varied notably when the classifier was trained on different data sets. For all of the studied years, the ensemble of classifiers trained on oversampling data sets reduced the F-score by up to 20 % compared to the raw data set for a one-hour lead time. However, the predictions were improved for the two-hour lead time for 2008 and 2011. The F-score results for the ensemble of classifiers demonstrated that, overall, the combined method produced the highest improvement in one and two-hour predictions.

Misclassification rates strongly support the interpretation of the F-score values. The ensemble of classifiers demonstrated the highest performance on the combined data set for one and two-hour lead times. In general, the misclassification rate shows that the most accurate predictions and the least number of misclassified tuples were achieved on ensembles of the combined data set for one and two-hour lead times for all years in the study. Similar to the F-score, the misclassification rate shows that an increase in lead time weakened the predictive ability of the classifiers. The ensemble performance on the oversampled data set demonstrated that the oversampling approach increased misclassified tuples for one-hour lead time, especially in 2011 and 2012. Classifiers trained on the oversampled data set correctly predicted classes of more tuples for a two-hour compared to a one-hour lead time in 2008 and 2011.

At the same time, in the dry year of 2012, training the ensemble on the combined data set did not improve the results. In that year, the combined approach generated results which are similar to the benchmark of the study. The performance of the classifier trained on inter-annual data was much better (approximately 40% improvement for two-hour lead time). Unfortunately, the traditional oversampling approach did not work well on these data. All other investigated approaches to constructing training data sets outperformed oversampling.

7. CONCLUSION

We found that the ensemble of five classifiers allowed for the best predictions on the combined data set for all wet years. Although this approach did not generate the best predictions for the dry year of 2012, its performance remained satisfactory compared to the other investigated methods. It should be noted, that out of four considered years, only one year (i.e., 2012) was dry with the least number of high-flow events observed at the cross section D. This means that very few tuples representing a dry season were available for training. To further discover flood patterns of dry years and also to leverage our understanding of watershed processes at various hydrological conditions, rainfall and water level data for multiple dry years have to be accumulated. This issue will be a subject of our further research.

The application of classification algorithms for short-term prediction of hydrological events has an obvious advantage that there is no need for labor- and time-consuming model parameterization. Site-specific calibration of the classifier is implemented at the training step using data on a watershed from rain and stream gauges which are available almost in real-time regime. It makes classification algorithms very attractive as a tool for flood management in a watershed. However, training classifiers on real time data at the beginning of the wet season may not be possible because tuples for flood events may not be available. The proposed combined approach may help to address this issue, and the hydrological events can be predicted using the historical records. The development of such a framework is the subject of further investigation.

ACKNOWLEDGEMENTS

This study was conducted as a part of MITACS Accelerate Project IT02708. The authors are thankful to TRCA, Canada, for providing data and clarifications necessary to complete this project. Partial financial and administrative support of the project was provided by the Faculty of Liberal Arts and Professional Studies, York University, Canada. The authors are grateful to Dean Singer and Associate Dean Adelson, Faculty of Liberal Arts and Professional Studies, York University for their help and support. The authors express their gratitude to anonymous reviewers for their thoughtful suggestions on the improvement of the manuscript.

REFERENCES

Environment Canada, Wateroffice, Daily Discharge for Etobicoke Creek at Brampton, http://wateroffice.ec.gc.ca/report/report_e.html?type=h2oArc&stn=02HC017 (Accessed June 27, 15).

Feldman, A. D. (2000). Hydrologic modeling system HEC-HMS: technical reference manual. *US Army Corps of Engineers, Hydrologic Engineering Center*. http://rivers.snre.umich.edu/639rivmod/hms_technical.pdf (Accessed July 17, 2015).

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.

Han, J., Kamber, M., and Pei, J. (2012). Data mining: concepts and techniques: concepts and techniques, third edition. Waltham, Mass.: Morgan Kaufmann Publishers

Horritt, M. S., and Bates, P. D. (2002). Evaluation of 1D and 2D numerical models for predicting river flood inundation. *Journal of Hydrology*, 268(1), 87-99.

Hunter, N. M., Bates, P. D., Neelz, S., Pender, G., Villanueva, I., Wright, N. G., Liang, D., Falconer, R. A., Lin, B., Waller, S., Crossley, A. J., and Mason. D. C. (2008). Benchmarking 2D hydraulic models for urban flood simulations. *In Proceedings of the Institution of Civil Engineers: Water Management*, 161 (1), 13-30. Thomas Telford (ICE publishing).

McCulloch, D. R., Lawry, J., and Cluckie, I. D. (2008). Real-time flood forecasting using updateable linguistic decision trees. *In Fuzzy Systems*, FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence). IEEE International Conference on (1935-1942). IEEE doi: 10.1109/FUZZY.2008.4630634

Sandford, R. W., and Freek. K. (2014). Flood forecast: climate risk and resiliency in Canada. *Rocky Mountain Books Ltd.*

Segretier, W., Clergue, M., Collard, M., Izquierdo, L. (2012). An evolutionary data mining approach on hydrological data with classifier juries. *In Proc. Evolutionary Computation* (CEC), 2012 IEEE Congress.

Segretier, W., Collard, M. and Clergue. M. (2013). Evolutionary predictive modelling for flash floods. *In Proc. Evolutionary Computation (CEC), 201, 844-851, IEEE Congress*.

Toronto and Region Conservation Authority (TRCA) (2008). Etobicoke Creek headwaters sub-watershed study synthesis report. http://trca.on.ca/dotAsset/96046.pdf (Accessed June 9, 2015).

Toronto and Region Conservation Authority (TRCA) (2011). West Etobicoke Creek South of Britannia road east erosion control project. http://www.trca.on.ca/dotAsset/120544.pdf (Accessed June 9, 2015).

Vieux, B. E., Cui, Z., and Gaur, A. (2004). Evaluation of a physics-based distributed hydrologic model for flood forecasting. *Journal of Hydrology*, 298(1), 155-177.