Addressing Imbalanced Data in Application of Predictive Analytics to Hydrological Problems

Maryum Saeed (Supervisor: Prof. Marina Erechtchoukova)

School of Information Technology, Faculty of Liberal Arts and Professional Studies, York University

Motivation

Due to changes in land use, there have been frequent accounts of intensive precipitation and flash floods in small watersheds. This encourages the need for an effective decision support tool for water resource management.

The research project entails the study on a data-driven framework for short-term prediction of hydrological events, e.g., floods or droughts, on a watershed based on machine learning techniques and data supplied by hydrological and meteorological monitoring networks.

The project targets the investigation of approaches for constructing training and testing sets which generate models of extreme hydrological events in small watersheds to ensure that forecasts are made timely and reliable. The predictions are considered reliable if their uncertainty is kept below 20%

Objective

The data-driven framework for short-term prediction of hydrological events [1] uses data routinely collected on watersheds. These data sets contain elements predominantly describing low-flow events leaving high-flow events underrepresented because these events happen in natural waterbodies occasionally.

The imbalanced nature of the data set used to develop models may lead to model bias and errors in predictions of flood-events.

The goal of the study was to investigate the effect of proportion of high-flow elements in data sets used for model construction on reliability of predictions generated by this model. The multi-year high-flow approach [3] was also compared with oversampling method utilized to mitigate imbalance in data.

Case Study

Data provided by Toronto and Region Conservation Authority (TRCA) was used to conduct experiments on a small urbanized watershed of the Spring Creek.

The highly urbanized and populated area of approximately 50 km² forms the Spring Creek watershed where the creek runs over 25 km.

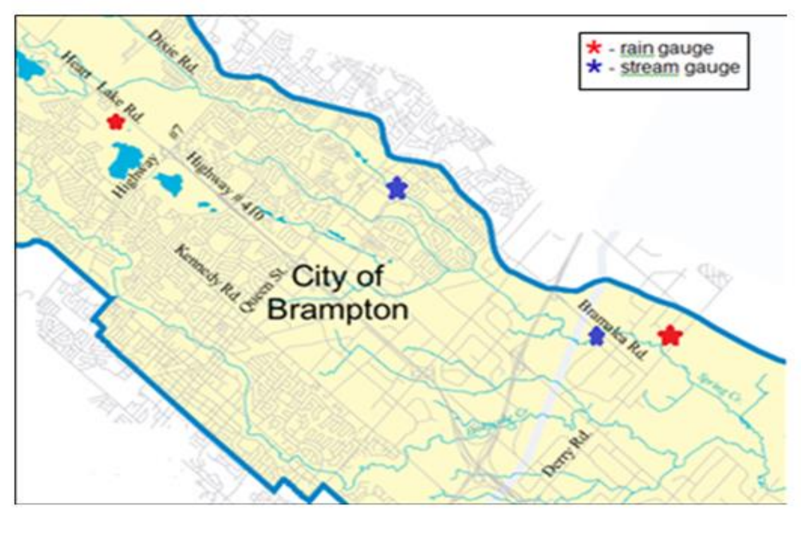


Figure 1 Spring Creek watershed scheme with monitoring sites (TRCA, 2006)

Our research used 15-min interval data of rainfall and water level observed at the watershed. The stream gauges are placed at Spring Creek North and Spring Creek South. Heart Lake and Mississauga Works Yard have rain gauges installed (Figure 1).

Time series of data on water levels and precipitation at all four monitoring sites for the warm period of 2013 – 2016 were used in computational experiments.

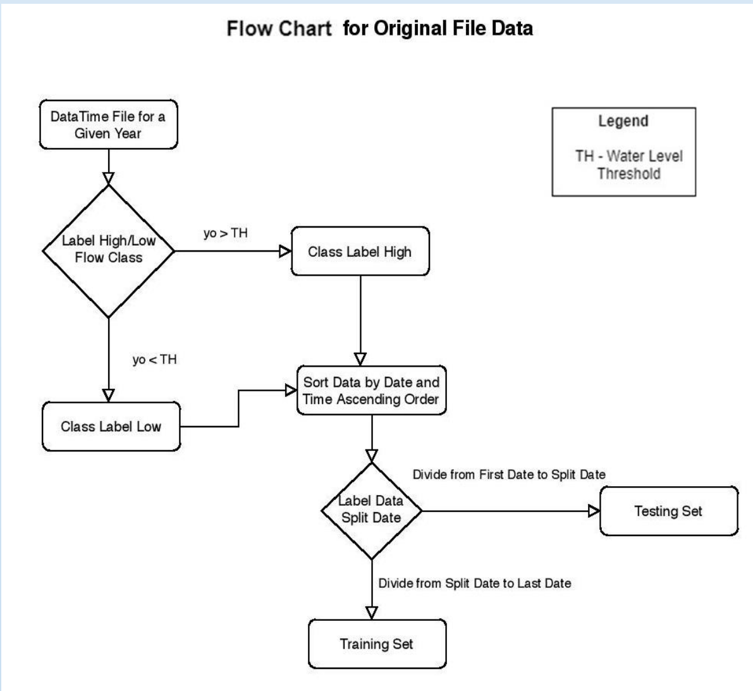
Methodology

The models were constructed following the framework for short-term prediction of hydrological events [1]. As a result, each element of the phase space can be reconstructed based on the following formula:

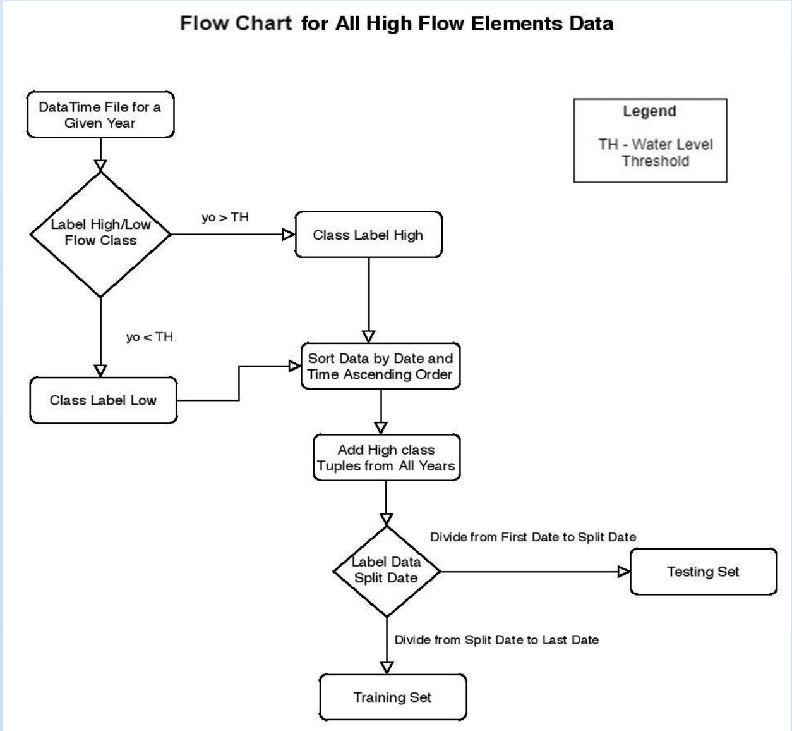
$$X(t+j\tau) = \; Y_1(t-1), Y_1(t-2), \ldots, Y_1(t-K\tau), \ldots Y_M(t-1), Y_M(t-2), \ldots, Y_M(t-K\tau), Class(t+j\tau)\,, \quad (1)$$

where *X(t+jτ)* is the element of the phase space constructed to generate predictions with *jτ* lead time, *j* = 1,…*K*, $Y_i(t)$ is the measurement from *i-th* gauge at time *t*, *i* = 1,…,*M*, *Class(t)* is the class label of an event at the investigated cross section at the time *t*.

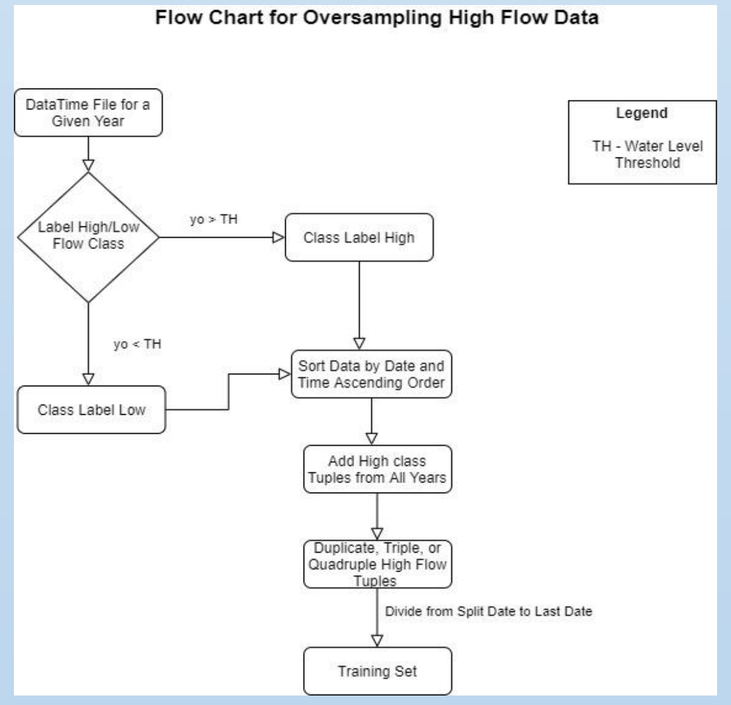
This original phase space was split into two disjoint subsets: one with 70% of records was used for training machine learning algorithms and the other one was left for testing developed models. The original phase space was further modified by adding historical high-flow records for several years.



Flow Chart for Original File Data
DataTime File for a Given Year
Label High/Low Flow Class
yo > TH
Class Label High
yo < TH
Class Label Low
Sort Data by Date and Time Ascending Order
Label Data Split Date
Divide from First Date to Split Date
Testing Set
Divide from Split Date to Last Date
Training Set
Legend
TH - Water Level Threshold



Flow Chart for All High Flow Elements Data
DataTime File for a Given Year
Label High/Low Flow Class
yo > TH
Class Label High
yo < TH
Class Label Low
Sort Data by Date and Time Ascending Order
Add High class Tuples from All Years
Label Data Split Date
Divide from First Date to Split Date
Testing Set
Divide from Split Date to Last Date
Training Set
Legend
TH - Water Level Threshold



Flow Chart for Oversampling High Flow Data
DataTime File for a Given Year
Label High/Low Flow Class
yo > TH
Class Label High
yo < TH
Class Label Low
Sort Data by Date and Time Ascending Order
Add High class Tuples from All Years
Duplicate, Triple, or Quadruple High Flow Tuples
Divide from Split Date to Last Date
Training Set
Legend
TH - Water Level Threshold

Both the original and multi-year high-flow records phase spaces were further transformed following oversampling method where high flow tuples were duplicated, tripled, quadrupled, and split into training and testing sets.

Five machine learning algorithms were trained and tested on these sets for three lead time intervals: 30-min, 45-min, and 60-min.

Proportion of high-flow records in training sets:

File Type	High Flow Records Percentage			
	2013	2014	2015	2016
Original Year High Flow	2.3%	1.0%	1.1%	0.4%
Four Year High Flow	4.7%	4.9%	4.7%	4.6%
Duplicated Original Year High Flow	4.4%	2.0%	2.1%	0.9%
Tripled Original Year High Flow	6.5%	3.0%	3.1%	1.3%
Quadrupled Original Year High Flow	8.5%	3.97%	4.1%	1.7%
Duplicated Four Year High Flow	8.9%	9.3%	8.9%	8.7%
Tripled Four Year High Flow	12.8%	13.3%	12.8%	12.5%
Quadrupled Four Year High Flow	16.3%	15.2%	16.4%	16.0%

Algorithms Used

- J48
- JRip
- PART
- RandomForest
- REPTree

The experiments were run in WEKA software [4]..

The performance of the developed models was evaluated based on the testing sets comprising elements of phase spaces unseen on the training step. The following performance measure were used:

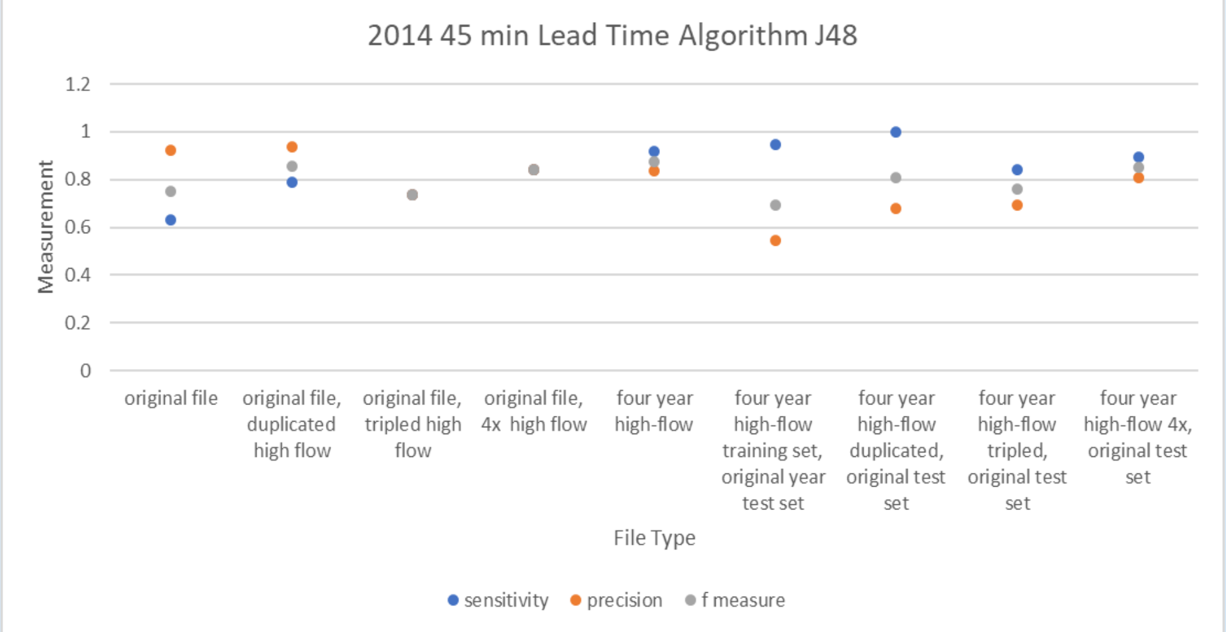
$$Sensitivity = \frac{true\_positive}{positive}, \qquad Precision = \frac{true\_positive}{true\_positive + false\_positive}. \qquad F = \frac{2 \cdot (True\ Positive)}{False\ Positive + False\ Negative + 2 \cdot (True\ Positve)}$$

where *true positive* represents the number of elements the model accurately classifies as positive, and *true positive* is the total number of positive elements in the testing set. *False positive* is the number of negative elements in a testing set the model incorrectly predicts as positive, and *false negative* is the number of positive elements in the testing set the model incorrectly classifies as negative.

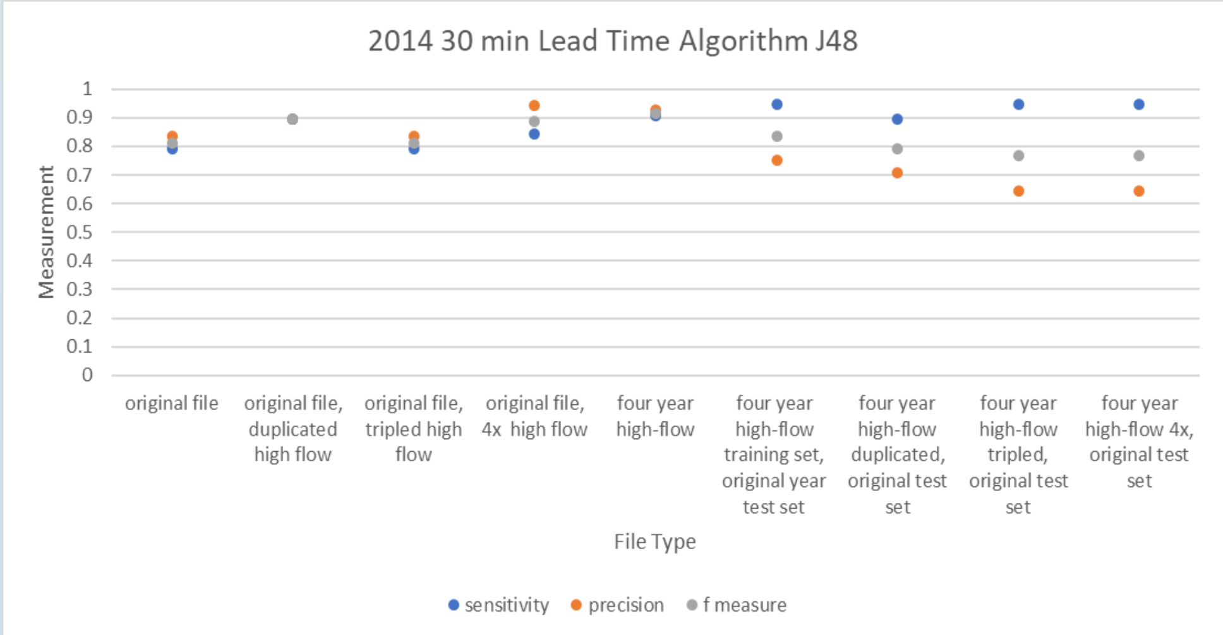
Results

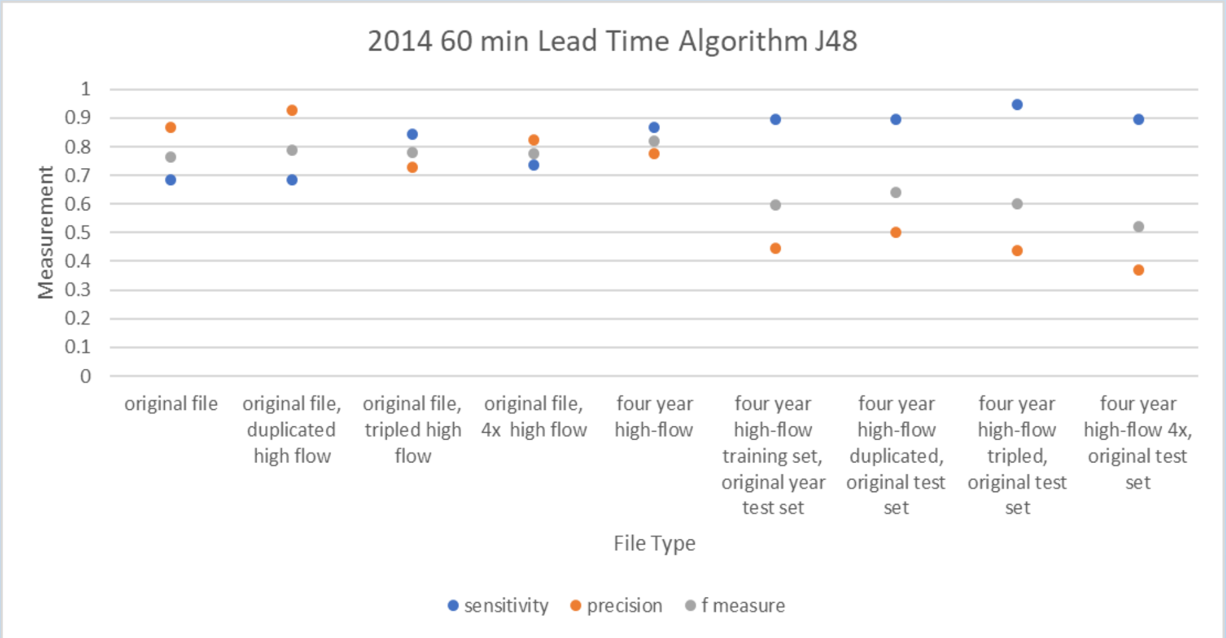
The performance estimates of the models trained on original phase spaces with different lead time intervals (no historical or duplicated records) were used as major benchmarks in the analysis of the results.

These estimates were compared with the performance of models trained on oversampled data sets and data sets with historical records representing high-flow events.

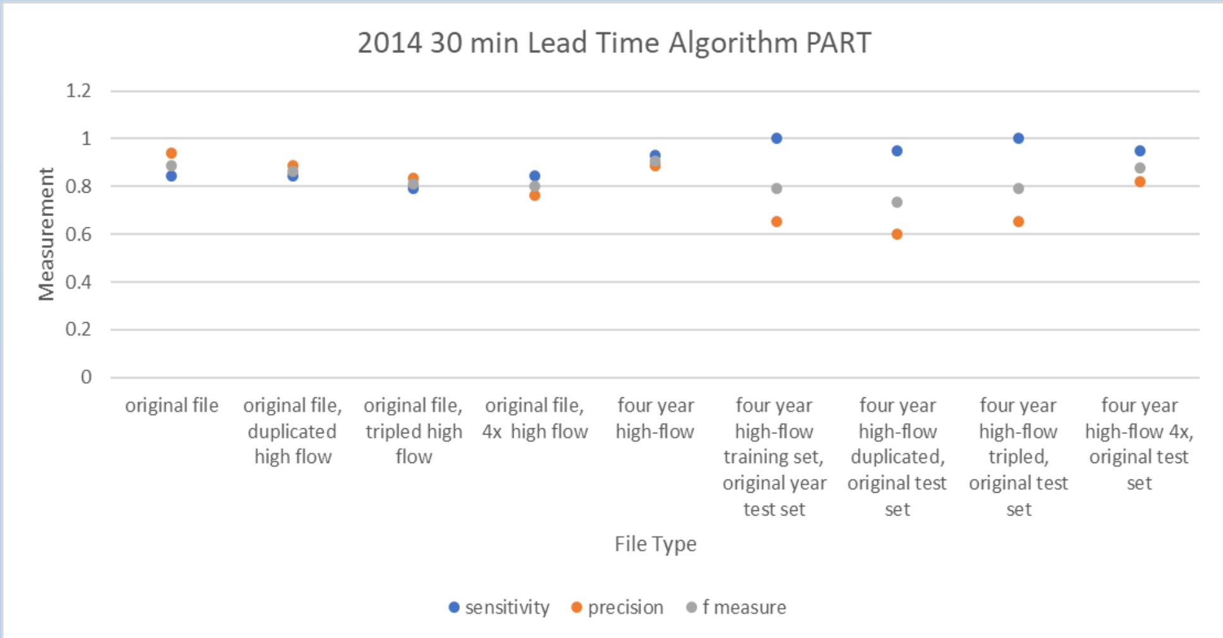


Model improvement from original file (benchmark)

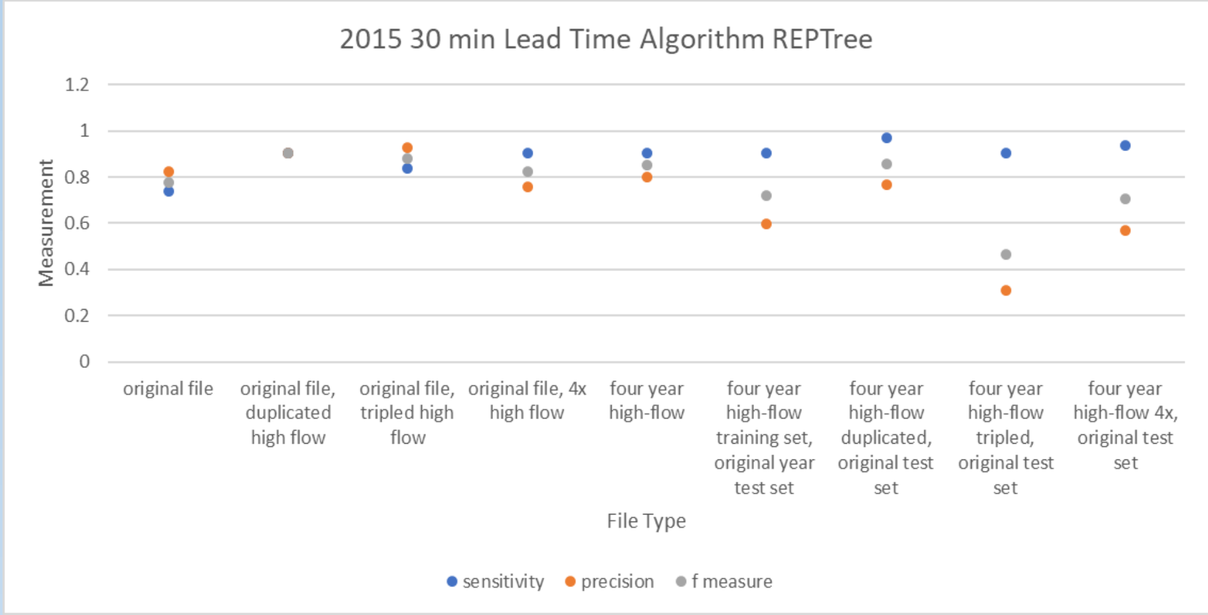


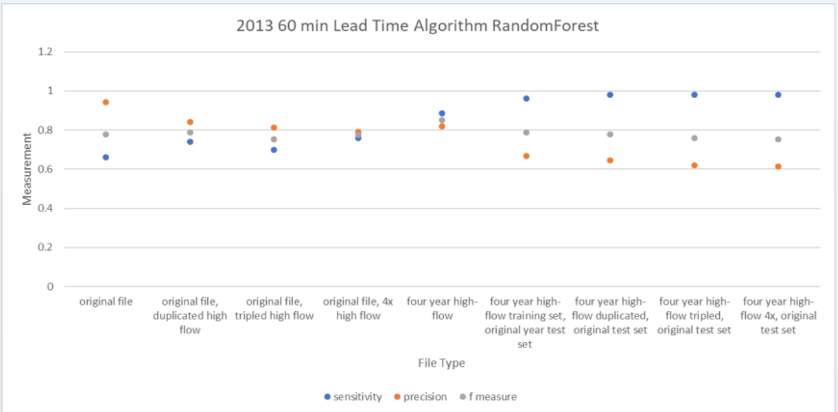


2014 Algorithm J48 showing the change in measures over 3 lead times

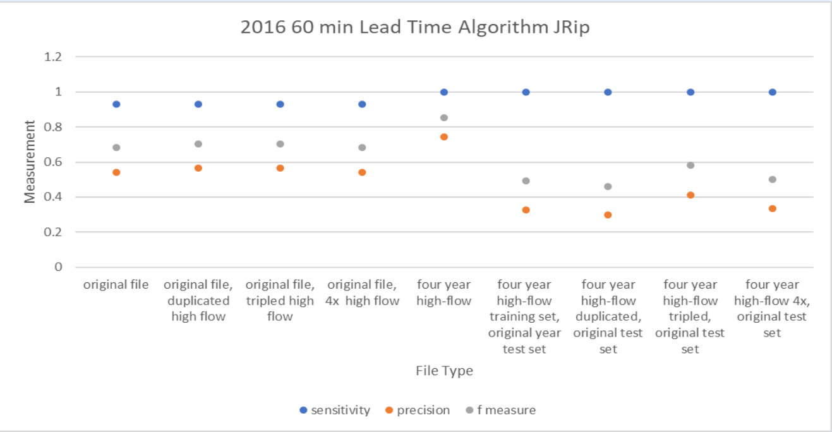


Good performance of original files





Model increase in sensitivity measure as files are further modified



Conclusions

Oversampling of the minority class in original phase spaces did not result in consistent improvements of the model performance. The obtained estimates varied notably for different algorithms and lead time intervals. Nevertheless, the original data sets allow for training reliable models. The results of this set of experiments validate application of all investigated algorithms for short-term prediction of hydrological events even on severe imbalanced data sets.

Oversampling of the minority class in phase spaces with historical records helped to improve model performance. The improvement is more significant for extended lead time intervals which is important for application of the developed models in early warning and flood management systems.

The most prominent and consistent improvements in model predictive ability were achieved by adding historical records on high-flow events observed on the investigated watershed to the training sets. This modification of original phase spaces brought the sensitivity of the developed models above 80% which is the threshold adopted by practitioners to determine reliability of forecasts. Subsequent oversampling of these data sets led to minor improvements of the results also confirming that data imbalance did not have notable effect.

The data used in this project were collected on the Spring Creek watershed, Brampton, Ontario. However, the predictive analytics developed in the project can be applied to any small urbanized watershed from any region in the world provided that sufficient data are available. The reliable predictions of extreme hydrological events for extended lead time intervals can significantly aid in damage control that municipalities can execute as mitigation measures to reduce the loss to structures and human lives due to flash floods.

References

1. Erechtchoukova, M.G., Khaiter, P.A., Saffarpour, S., 2016. Short-term predictions of hydrological events on an urbanized watershed using supervised classification. Water Resources Management, 30(12):4329-4343, DOI: 10.1007/s11269-016-1423-6.
2. Saffarpour, S., Erechtchoukova, M.G., Khaiter, P.A., Chen, S.Y., Heralall, M. (2015). Short-term prediction of flood events in a small urbanized watershed using multi-year hydrological records. In: Weber, T., McPhee, M.J. and Anderssen, R.S. (eds.) MODSIM2015, 21st International Congress on Modelling and Simulation. MSSANZ, December 2015, pp. 2234-2240. ISBN: 978-0-9872143-5-5.
3. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I, H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1).