The 2014 International Conference on Marine and Freshwater Environments

Joint Conferences:
The 2014 Annual Conference of the International Society for Environmental Information Sciences (ISEIS)
The 2014 Atlantic Symposium of the Canadian Association on Water Quality (CAWQ)
The 2014 Annual General Meeting and 30th Anniversary Celebration of the Canadian Society for Civil Engineering Newfoundland and Labrador Section (CSCE-NL)
The 2nd International Conference of Coastal Biotechnology (ICCB) of the Chinese Society of Marine Biotechnology and Chinese Academy of Sciences (CAS)

Peak Flow Prediction Using Fuzzy Linear Regression: Case Study of the Bow River

Usman T Khan[1*], and Caterina Valeo[1]

[1]*Mechanical Engineering, University of Victoria, BC, Canada*
**Corresponding author's e-mail:* utkhan@uvic.ca

Abstract

The 2013 floods in Alberta highlighted the need for better flood prediction. Though the mechanisms behind floods and extreme events in urban areas are understood and documented, the uncertainty in data during these events makes it difficult to accurately predict and assess the risk of floods. In this research, a fuzzy number based linear regression model is proposed that incorporates uncertainty and characterizes risk of extreme events in the Bow River at Calgary, Alberta, Canada. The proposed model uses a fuzzy linear regression model to predict peak flow rate using mean daily flow rate. Lagged data from one to seven days is also considered. Results of the research show that using a fuzzy number approach to predict uncertain extreme events outperforms traditional regression methods in the Bow River at Calgary. The developed model can accurately predict daily peak flow, including a flood event in 2005, up to 7 days in advance. In addition to this, fuzzy number model output can be used to further characterize the risk of peak flow magnitude. These results are extremely beneficial for water resource managers who implement flood mitigation and defence strategies.

Keywords: Flood, Fuzzy linear regression, Peak flow, Risk analysis, Uncertainty

Introduction

The flood that occurred in Alberta in June 2013 was one of the worst natural disasters to occur in Canada and the event highlighted the need for better flood prediction. The floods were the costliest natural disaster in Canada, causing approximately $6 billion in damage (Environment Canada, 2013). In addition to this, the floods caused four deaths and displaced more than 100,000 Albertans in over 30 communities (Alberta Government, 2014). The floods also impacted the environment; high flow rates caused permanent changes to watersheds in southern Alberta, including the transport of large amounts of sediment and the destruction of river banks, channels and aquatic ecosystems (Environment Canada, 2013).

Starting on 3 June 2013, the Government of Alberta posted High Streamflow Advisories for impacted watersheds. By 10 and 11 June 2013, a Flood Watch and Flood Warning were issued for selected watersheds. However a wider Provincial Advisory was not issued until 19 June 2013, following extreme precipitation events in southern Alberta (ESRD, 2013; Environment Canada, 2013). Thereafter, most of the Province was under a Flood Watch and Flood Warning from 20 June to 29 June 2013 (ESRD, 2013). The unprecedented heavy rainfall that occurred in June 2013 is largely believed to have caused the floods, however other contributing factors should not be ignored. For example, satellite imagery from May 2013 indicated that there was little capacity

for the watersheds in the region to store excess water, suggesting a risk of floods (Environment Canada, 2013).

Though the Government of Alberta uses numerical modelling as part of its flood mitigation strategy (Alberta Government, 2014), the flood modelling predictions did not provide a warning far enough in advance to trigger Advisories, Flood Watch or Flood Warnings from the Government. Thus, there is a need for improved flow or peak flow rate models, which incorporate the risk of flooding. This would be extremely beneficial in preparing for floods in the future. If the impact of a flood can be estimated in advance, there is the potential to reduce the enormous social, environmental and financial costs associated with it.

However, flood prediction is inherently uncertain. While the mechanisms behind floods in urban areas are understood and documented, the uncertainty in data during these events makes it difficult to accurately predict and assess the risk of floods in the future. In addition to this, physically-based models try to simulate complex physical systems by breaking them down into smaller, simpler units (Cox, 2003). These models also require specific and often hard to obtain data to parameterize the various sub-models. Both these factors, the simplification and the high data requirements, introduce additional uncertainty in physically-based models. This uncertainty is often difficult to propagate through the model and results in highly uncertain prediction of flow rate and other flood parameters.

An alternative approach to physically-based models that are typically used for flood prediction, mitigation and planning, is to use a data-driven approach. Data-driven models are based on generalized relationships, links or connections between input and output datasets (Solomantine & Ostfeld, 2008). The models can characterize a system with limited assumptions about it and often have similar, if not better performance than physically-based models. A simpler model structure means that the propagation of uncertainty from different sources is easier. While data-driven models may also be data intensive, the data can be collected from on-going monitoring systems, *e.g.* real-time flow rate data that is routinely collected by Environment Canada (Environment Canada, 2014), rather than data specific to the model in question. However, while a data-driven approach has its obvious advantages over physically-based models, they have different objectives. Data-driven models may improve our predictions of the future state of a system, but they might not provide a better physical understanding of the system. From this aspect, flood prediction for mitigation and planning purposes is an ideal candidate for a data-driven approach.

The nature of data-driven modelling means that these methods have intrinsic uncertainties associated with it. This uncertainty is not of purely random or probabilistic in nature, making it well suited for the use of fuzzy number (Dubois & Prade, 1997; Ozbek & Pinder, 2006). Fuzzy numbers use fuzzy set and possibility theory to describe uncertain or imprecise information. A fuzzy number is a specific type of quantity that expresses uncertain or imprecise quantities, measurements or observations (Zhang & Achari, 2010; Huang, et al, 2010). A major advantage of using fuzzy numbers is that they have the ability to provide more meaningful information compared to traditional techniques, especially in highlighting the possibility and probability of events like floods. Fuzzy numbers have been widely used in hydrology to represent uncertainty in the parameters of numerical models (Khan and Valeo, 2014; Khan et al 2013). The literature demonstrates the utility and advantage of using fuzzy numbers; a summary of these applications can be found in Khan and Valeo (2014).

Thus, there is potential to use a fuzzy number based data-driven model to predict the risk of floods like those that occurred in southern Alberta in 2013. A fuzzy linear regression (FLR) method, proposed by Khan and Valeo (2014) has demonstrated the utility of a fuzzy number based data-driven model. Previous studies using this method have shown improved prediction and risk analysis of environmental factors (Khan and Valeo, 2014; Khan et al 2013). In this paper, this FLR method is used predict peak flow in the Bow River, in Calgary, Alberta, Canada and the associated risk of floods.

The objective of this research is to improve flood prediction in the Bow River basin in Calgary, Alberta, Canada using a fuzzy number based data-driven approach. This model should provide accurate peak flow estimates to be able to quantify the severity of the flood, and also to provide the capacity for early flood mitigation and preparation, in a timely and effective manner. An FLR model is proposed, under the condition that it have minimal data requirements. Results from this method will be compared to simple linear regression methods. Also, a risk analysis using these results will be developed.

Methods

Data Collection

The Bow River originates from the Rocky Mountains and flows southeast through the City of Calgary, as shown in Figure 1. It has an average annual discharge of 90 m^3/s and it provides approximately 60% of the potable water for the city, is used for recreation purposes, and supports an aquatic ecosystem used for its fish resources. After flowing through Calgary, the river meets the Oldman River, and flows east as the South Saskatchewan River and ultimately draining into Hudson Bay (Robinson et al, 2009). Hourly flow rate for the Bow River in Calgary (Station Number 05BH004) was obtained from Environment Canada for the period from 2004 to 2008. During this period, one major flood occurred in 2005 – the maximum peak flow recorded that year was 728 m^3/s on 19 June 2005.

Figure 1. An aerial image of the City of Calgary showing the location of the flow rate monitoring station within the Calgary city limits. The insert shows the location of Calgary within Canada.

For this research, three years of data (2004, 2006 and 2008) was used to calibrate the FLR model, while the remaining two years data was used to validate the model. This selection meant that the flood of 2005 was not used to calibrate the model. This was done intentionally, so that the ability of the model to predict extreme events (like that in 2005) could be measured. Only the ice-free period, typically April through November was used in this study.

Figure 2 shows a trend plot of the daily peak flow (Q_p) and the mean daily flow (Q_d). Apart from the flood event in 2005, there is very little inter-annual variability. The similarity of the two trends suggests a high correlation between the variables; this implies that Q_d might be a good candidate as a variable to predict Q_p. Note that hourly data for December 2006 was missing; only daily mean values were available. Thus Q_p could not be calculated for this period.

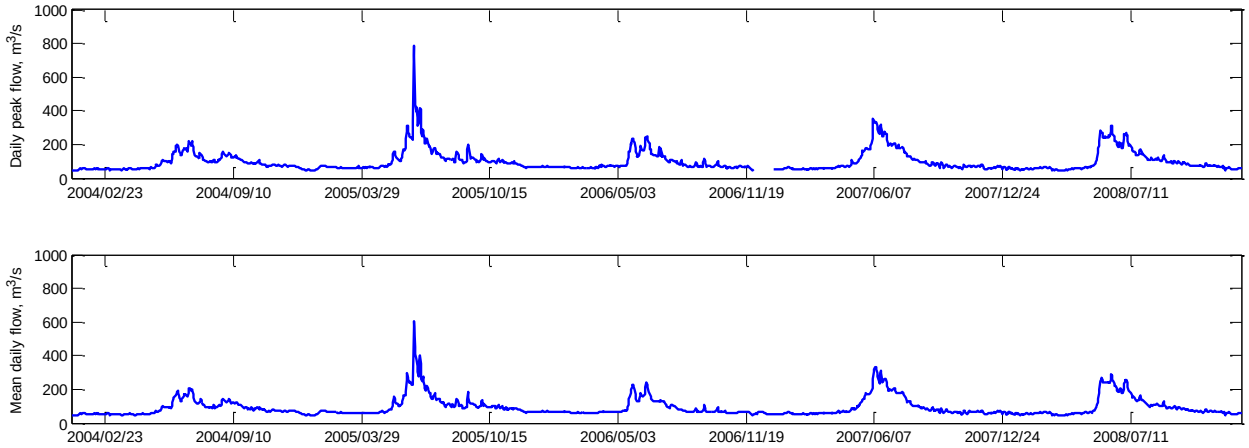

Figure 2. A trend of daily peak flow (top) and mean daily flow (bottom) for the Bow River at Calgary between 2004 and 2008.

A correlation analysis was conducted between the two variables to determine different time lags that can be potentially used to predict Q_p using Q_d. Figure 3 shows a plot of Q_p versus Q_d at a lag of 0 days (*i.e.* no lags, for reference), 1 day, 3 days and 7 days. As expected, the correlation decreases as the lag increases. However, even after 7 days, the correlation coefficient is high at 0.79. This means that if the uncertainty in predicting Q_p can be quantified, then accurate peak flow rates can be predicted a week in advance.

Based on this initial analysis, the following functional form of the FLR model was selected:

$$Q_p(t) = f(Q_d(t-d))$$

where $Q_p(t)$ is the daily peak flow on day t, Q_d is the mean daily flow on day t –d, and d is the time lag in days, selected as either 1, 3 or 7 days. However, to construct an FLR model, each of the model inputs, output and coefficient must be in fuzzy number format, before FLR can be conducted. This is detailed in the following section.

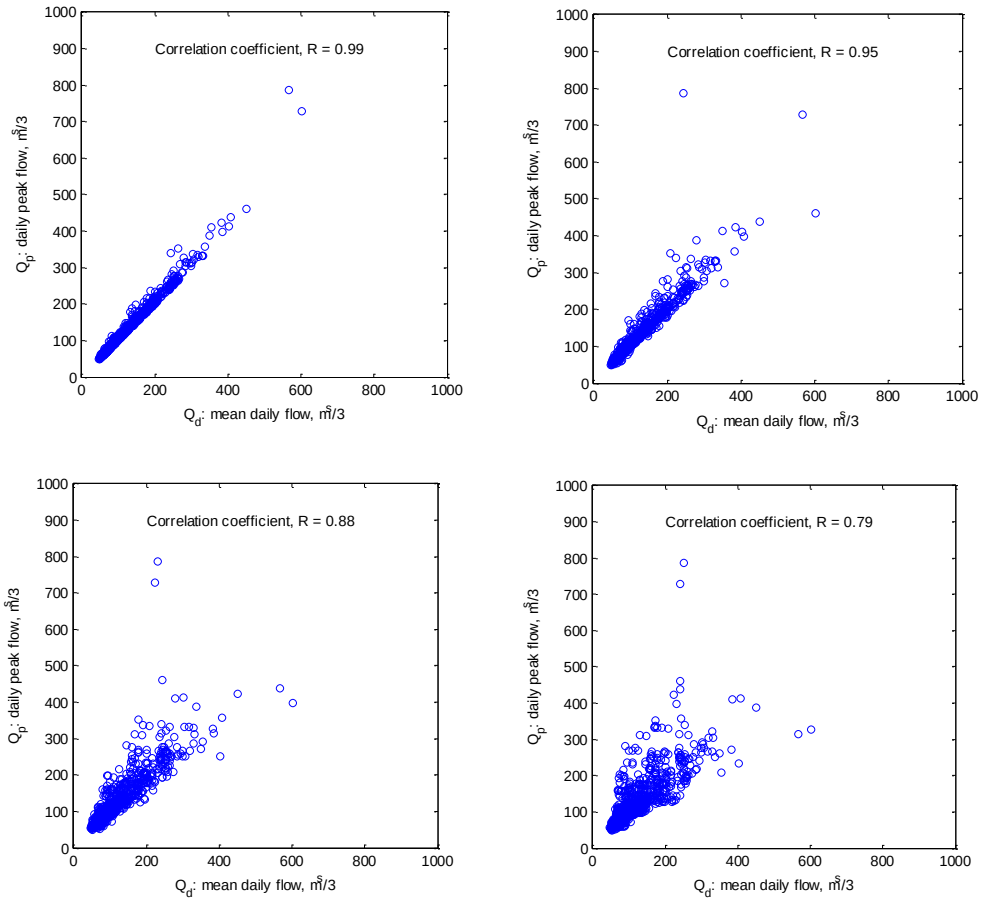

Figure 3. A comparison of correlation between Q_p and Q_d at lags of 0, 1, 3 and 7 days (clockwise from top left).

Fuzzy Linear Regression

Fuzzy linear regression is a method used to extend linear regression for applications involving fuzzy numbers [Khan and Valeo, 2014]. It provides an alternative method when simple linear regression may not be possible, *e.g.* when assumptions of linear regression are not met, or if there is obvious fuzziness in the underlying data or process. FLR tries to capture the vagueness, and the non-random or fuzzy error in the model structure: it is assumed that deviations are due to system fuzziness, *i.e.* the fuzziness of the regression coefficients (Chang and Ayyub, 2001).

A fuzzy linear regression is proposed in Khan and Valeo (2014) is implemented to predict Q_p in this research. The results from this an analysis are compared to observed data, and the results from simple linear regression for comparison purposes. This FLR method is unique in that fuzzy number inputs, outputs and regression coefficients are used, whereas other FLR techniques do not typically use fuzzy numbers for each of these variables (Khan and Valeo, 2014). In addition to this, the FLR method used here uses non-linear membership functions to define fuzzy numbers; this is much more suitable for analysis of flow rate, which is typically not symmetrically distributed. A probability-possibly transformation is used to construct convert the observed hourly flow data to daily mean flow rate. An example of this transformation for three different flow regimes (*i.e.* low flow, medium flow and high flow) is shown in Figure 4 for daily

mean flow. The background for this transformation can be found in Dubois et al (1993) and Dubois et al (2004) and is not discussed further here. Daily peak flow is handled differently: first the peak flow rate for each day is collected, and then a range of ±6% is used to determine the upper and lower fuzzy limits of the measurement (Khan et al, 2013).

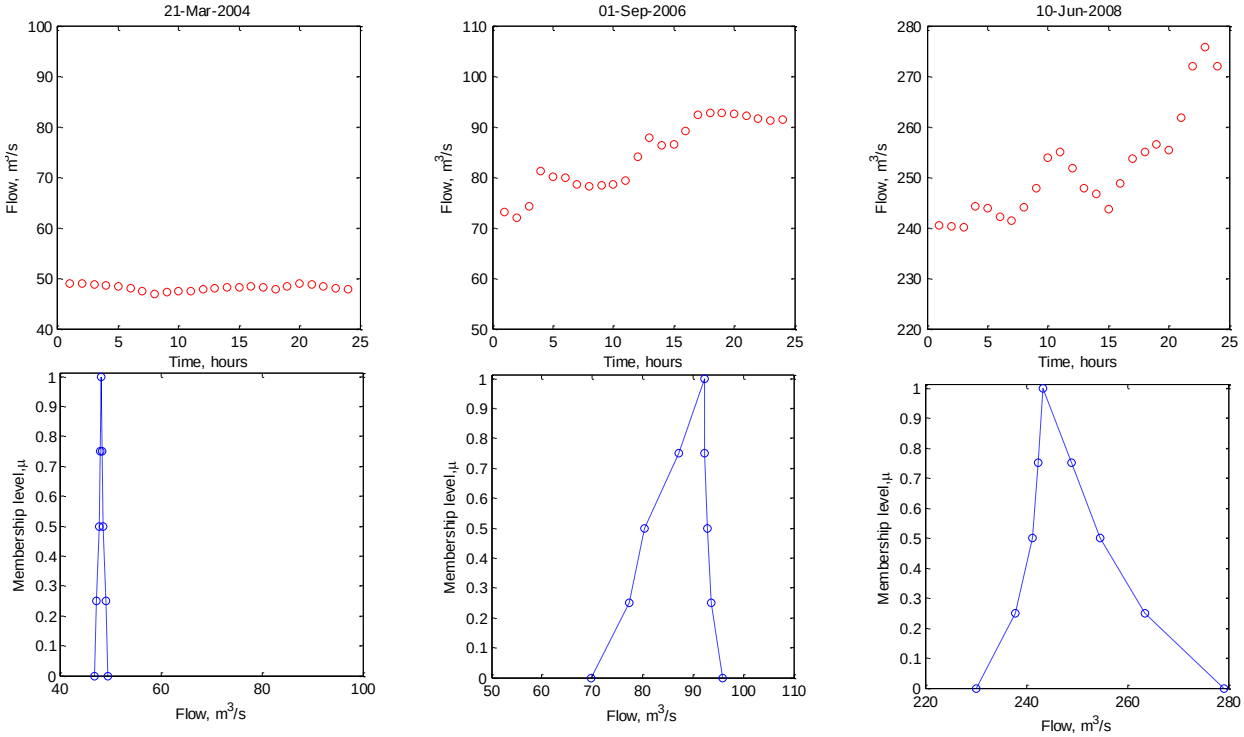

Figure 4. Three examples of probability-possibility transformations used to create fuzzy numbers from observed hourly flow rate data.

The objective of FLR method is similar to the simple linear regression, however, instead of minimizing the residual between an observed and regressed value, the distance between two fuzzy numbers is minimized instead. Given a set of fuzzy observations $\tilde{Q}_{d_i}$ and $\tilde{Q}_{p_i}$, and their corresponding membership functions, $\mu(\tilde{Q}_{d_i})$ and $\mu(\tilde{Q}_{p_i})$, for (i = 1, 2, ..., n) an FLR model is defined as:

$$\tilde{Q}_p = \tilde{A} + \tilde{B}\tilde{Q}_d$$

where the coefficients $\tilde{A}$ and $\tilde{B}$ are fuzzy numbers. The objective is to solve the following least-squares problem:

$$\min r(\tilde{A}, \tilde{B}) = \sum_{i=1}^{n} d^2\,(\tilde{Q}_{p_i}, \tilde{A} + \tilde{B}\tilde{Q}_{p_i})$$

where $d^2(\tilde{Q}_{p_i}, \tilde{A} + \tilde{B}\tilde{Q}_{p_i}) = \cup\,[\tilde{Q}_{p_i} - \tilde{A} - \tilde{B}\tilde{Q}_{p_i}]_\mu$ for i = 1, 2, ..., n and μ = 0 to 1. The metric d measures the sum of the squared-deviations of the observed ($\tilde{Q}_{p_i}$) and predicted ($, \tilde{A} + \tilde{B}\tilde{Q}_{p_i}$) intervals $[\ldots]_\mu$, for all alpha-cuts between μ = 0 and μ = 1. Using fuzzy arithmetic ensures that the coefficients $\tilde{A}$ and $\tilde{B}$ are normal and convex, a requirement of fuzzy numbers.

Error Analysis

The Nash-Sutcliff model efficiency (NSE), Root Mean Squared Error (RMSE), and the Mean Absolute Error (MAE) are used to measure the performance of both the FLR and simple linear regression models.

Results and Discussions

Figure 5 shows the results of applying the both simple linear regression and FLR method to predict Q_p using Q_d with a 7 day lag for the calibration dataset (*i.e.* data from 2004, 2006 and 2008). For the simple regression, the NSE was 0.66, RMSE was 30.19m^3/s, and MAE was 18.88 m^3/s. For FLR, the NSE was between 0.64 and 0.66, RMSE was between 29.18 and 33.22 m^3/s, and MAE was 17.08 to 22.62 m^3/s. The intervals are from the results of calculating the error at each membership level. Two important results can be seen from this figure, first in general, a 7-day lag can reproduce the general trend of peak flow. This is extremely important in highlighting advance warnings of a flood. Secondly, in comparing to the two methods, the FLR method is better able to reproduce higher values of Q_p within the $\mu = 0^L$ and 0^R interval (note that this interval is defined by the lower (L) and upper (R) values of Q_p when the membership level is 0).

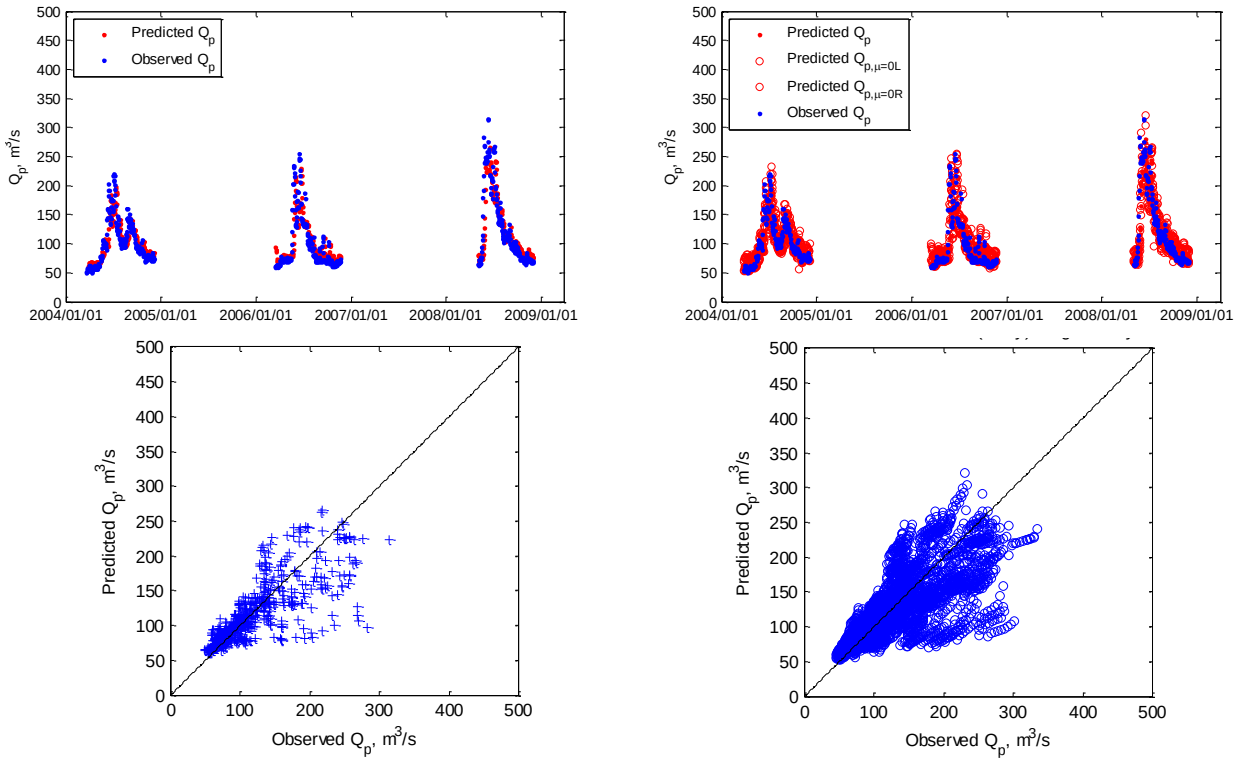

Figure 5. A comparison of result of simple linear regression (left) and fuzzy linear regression (right) for the model calibration dataset.

Figure 6 shows the same results for the validation dataset (2005 and 2007). This dataset included the floods that occurred in June 2005. While the simple linear regression method underestimates the flood peak flow on June 19 2005 by about 200 m^3/s, the FLR method produces results where the flood Q_p is nearly within the bounds of the $\mu = 0^L$ and 0^R interval. The significance of these results is that the FLR model predicted a risk of a flood 7-days in advance. The importance of

this fact can be highlighted by the fact that the training data used to construct the FLR model did not contain any major flood events.

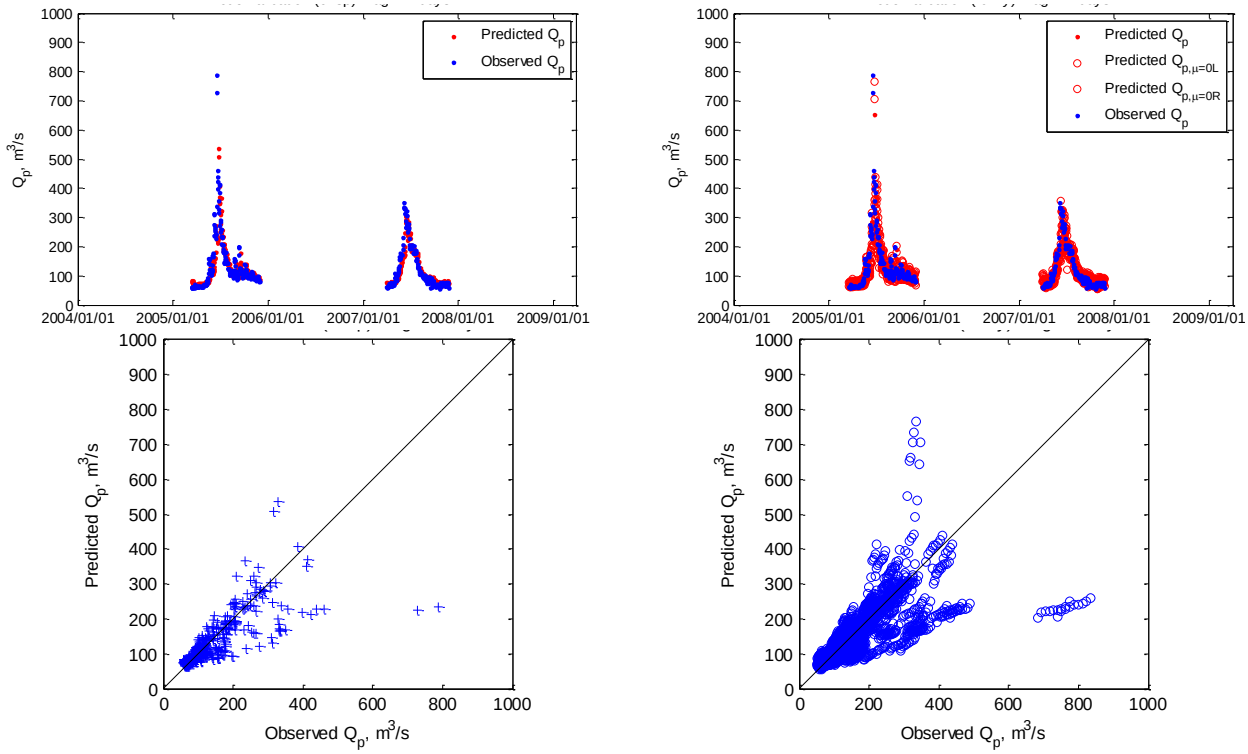

Figure 6. A comparison of result of simple linear regression (left) and fuzzy linear regression (right) for the model validation dataset.

The errors calculated using the validation dataset for the simple regression are an NSE of 0.60, RMSE was 53.36 m^3/s, and MAE of 24.33 m^3/s. For FLR, the NSE was between 0.54 and 0.61, RMSE was between 50.04 and 60.90 m^3/s, and MAE was between 22.58 and 29.29 m^3/s. These errors reduced as the lag was shortened from 7-days to 1-day. At 1-day, for the validation dataset (results not shown) were NSE of 0.87, RMSE of 30.16 m^3/s, and MAE of 9.67 m^3/s for simple regression. For FLR these values ranged between 0.80 and 0.88 for NSE, 30.33 and 35.78 m^3/s for RMSE and 9.71 and 12.50 m^3/s.

Risk analysis

One advantage of using FLR is the predictions from the model can be directly used to measure risk of flood events, using a possibility-probability transformation (Dubois et al, 1993; Dubois et al 2004). This transformation can be used to estimate the probability of Q_p occurring given a fuzzy number. In Figure 7 below, the predicted fuzzy number, the predicted Q_p using simple regression and the observed Q_p is shown for June 19 2005 using a 1 day lag. The results from the FLR method predict that the probability of Q_p to be greater than the observed Q_p of 728 m^3/s is approximately 24%. It also estimates that the probability of Q_p to be greater than the estimate for ordinary regression is approximately 58%.

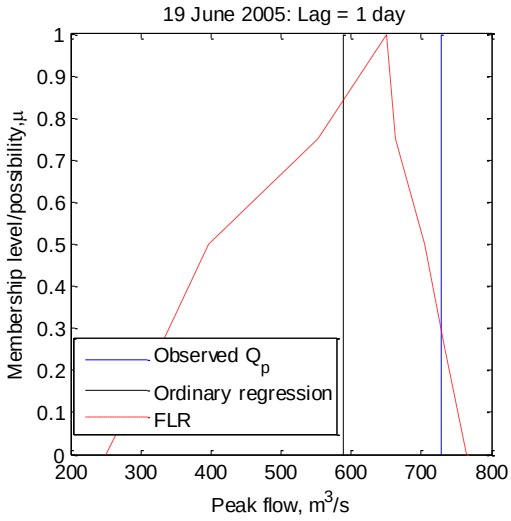

Figure 7. A comparison of results from simple linear regression, fuzzy linear regression and observed peak flow from the 2005 flood.

The importance of this aspect of fuzzy number analysis is that the risk of an event like a flood is directly included in the method. If a certain threshold of risk is set by water managers, municipalities, or provincial government, then the results of these simulations can be used to trigger an Advisory, Flood Watch or Flood Warning. The results from this analysis show that these triggers may be set as early as 7 days, if a risk of flood is predicted by the model. This means that flood mitigation and preparation.

Conclusions

The flood that occurred in Alberta in 2013 was one of the worst natural disasters in Canadian history. The event highlighted the need for improved flood prediction methods. A fuzzy number based data-driven method is proposed in this paper to predict peak flow in the Bow River in Calgary using mean daily flow as input. This analysis was conducted at various time lags (1, 3 and 7 days). The results show that peak flow, and the risk of flood, can be determined up to 7 days in advance. The 2005 flood in Calgary, Alberta, was used as an example to illustrate the utility of the method.

Acknowledgement

The authors would like to acknowledge NSERC, the BC Ministry of Education and the University of Victoria for funding this research, and Environment Canada for providing the datasets used in this study.

References

Alberta Government (2014). Respecting Our Rivers: Alberta's Approach to Flood Mitigation from https://pabappsuat.alberta.ca/albertacode/images/respecting-our-rivers.pdf visited on July 1, 2014.

Chang, Y.-H. O. and M Ayyub, B. (2001). Fuzzy regression methods–a comparative assessment. *Fuzzy sets and systems*, *119*(2), 187-203, doi: 10.1016/S0165-0114(99)00091-3.

Cox, B. A. (2003). A review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers. *Science of The Total Environment, 314-316*, 335-337. doi:10.1016/S0048-9697(03)00063-9

Dubois, D., & Prade, H. (1997). Bayesian conditioning in possibility theory. *Fuzzy Sets and Systems, 92*(2), 223-240. doi:10.1016/S0165-0114(97)00172-3

Dubois, D., Foulloy, L., Mauris, G. and Prade, H. (2004). Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing*, 10, 273-297, doi:10.1023/B:REOM.0000032115.22510.b5.

Dubois, D., Prade, H. and Sandri, S. (1993). On possibility/probability transformations. In R. Lowen and M. Roubens, eds. *Fuzzy Logic*, Dordrecht, Netherlands: Kluwer Academic Publishers, 103-112.

Environment Canada (2013). Alberta's Flood of Floods on Canada's Top Ten Weather Stories for 2013, http://ec.gc.ca/meteo-weather/default.asp?lang=En&n=5BA5EAFC-1&offset=2&toc=show visited 1 July 2014.

Environment Canada (2014). Real-time Hydrometric Data on the Wateroffice website, http://www.wateroffice.ec.gc.ca/index_e.html, visited 1 July 2014.

ESRD (2013). Archived Advisories & Warnings June 2013, http://environment.alberta.ca/forecasting/advisories/archives0613.html visited 1 July 2014.

Huang, Y., Chen, X., Li, Y. P., Huang, G. H., & Liu, T. (2010). A fuzzy-based simulation method for modelling hydrological processes under uncertainty. *Hydrological Process, 24*(25), 3718-3732.

Khan, U. T., & Valeo, C. (2014). A new fuzzy linear regression approach for dissolved oxygen prediction. *Hydrological Sciences Journal*, DOI:10.1080/02626667.2014.900558

Khan, U. T., Valeo, C., & He, J. (2013). Non-linear fuzzy-set based uncertainty propagation for improved DO prediction using multiple-linear regression. Stochastic Environmental Research and Risk Assessment, 27(3), 599-616.

Ozbek, M. M., & Pinder, G. F. (2006). Non-probabilistic uncertainty in subsurface hydrology and its applications: an overview. *Water, Air, & Soil Pollution: Focus*, *6*(1-2), 35-46. doi:10.1007/s11267-005-9011-4

Solomantine, D. P., & Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics*, *10*(1), 3-22. doi:10.2166/hydro.2008.015

Zhang, K., & Achari, G. (2010). Correlations between uncertainty theories and their applications in uncertainty propagation. In H. Furuta, D. M. Frangopol, & M. .. Shinozuka (Eds.), *Safety, reliability and risk of structures, infrastructures and engineering systems* (pp. 1337 - 1344). London, UK: Taylor & Francis Group