Environmental prediction and risk analysis using fuzzy numbers and data-driven models

by

Usman Taqdees Khan
BSc, University of Calgary, 2008
MSc, University of Calgary, 2011

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Mechanical Engineering

© Usman Taqdees Khan, 2015
University of Victoria

All rights reserved. This dissertation may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.
Supervisory Committee

Environmental prediction and risk analysis using fuzzy numbers and data-driven models

by

Usman Taqdees Khan
BSc, University of Calgary, 2008
MSc, University of Calgary, 2011

Supervisory Committee

Dr. Caterina Valeo (Department of Mechanical Engineering)
Supervisor

Dr. Bradley J. Buckham (Department of Mechanical Engineering)
Departmental Member

Dr. Yang Shi (Department of Mechanical Engineering)
Departmental Member

Dr. Alexandra Branzan Albu (Electrical and Computer Engineering)
Outside Member
Abstract

**Supervisory Committee**
Dr. Caterina Valeo (Department of Mechanical Engineering)
**Supervisor**
Dr. Bradley J. Buckham (Department of Mechanical Engineering)
**Departmental Member**
Dr. Yang Shi (Department of Mechanical Engineering)
**Departmental Member**
Dr. Alexandra Brzan Albu (Electrical and Computer Engineering)
**Outside Member**

Dissolved oxygen (DO) is an important water quality parameter that is used to assess the health of aquatic ecosystems. Typically physically-based numerical models are used to predict DO, however, these models do not capture the complexity and uncertainty seen in highly urbanised riverine environments. To overcome these limitations, an alternative approach is proposed in this dissertation, that uses a combination of data-driven methods and fuzzy numbers to improve DO prediction in urban riverine environments.

A major issue of implementing fuzzy numbers is that there is no consistent, transparent and objective method to construct fuzzy numbers from observations. A new method to construct fuzzy numbers is proposed which uses the relationship between probability and possibility theory. Numerical experiments are used to demonstrate that the typical linear membership functions used are inappropriate for environmental data. A new algorithm to estimate the membership function is developed, where a bin-size optimisation algorithm is paired with a numerical technique using the fuzzy extension principle. The developed method requires no assumptions of the underlying distribution, the selection of an arbitrary bin-size, and has the flexibility to create different shapes of fuzzy numbers. The impact of input data resolution and error value on membership function are analysed.

Two new fuzzy data-driven methods: fuzzy linear regression and fuzzy neural network, are proposed to predict DO using real-time data. These methods use fuzzy inputs, fuzzy outputs and fuzzy model coefficients to characterise the total uncertainty. Existing methods cannot accommodate fuzzy numbers for each of these variables. The new method for fuzzy regression was compared against two existing fuzzy regression methods, Bayesian linear regression, and error-in-variables regression. The new method was better able to predict DO due to its ability to incorporate different sources of uncertainty in each component. A number of model assessment metrics were proposed to
quantify fuzzy model performance. Fuzzy linear regression methods outperformed probability-based methods. Similar results were seen when the method was used for peak flow rate prediction.

An existing fuzzy neural network model was refined by the use of possibility theory based calibration of network parameters, and the use of fuzzy rather than crisp inputs. A method to find the optimum network architecture was proposed to select the number of hidden neurons and the amount of data used for training, validation and testing. The performance of the updated fuzzy neural network was compared to the crisp results. The method demonstrated an improved ability to predict low DO compared to non-fuzzy techniques.

The fuzzy data-driven methods using non-linear membership functions correctly identified the occurrence of extreme events. These predictions were used to quantify the risk using a new possibility-probability transformation. All combination of inputs that lead to a risk of low DO were identified to create a risk tool for water resource managers. Results from this research provide new tools to predict environmental factors in a highly complex and uncertain environment using fuzzy numbers.
### Table of Contents

Supervisory Committee ........................................................................................................................................................................... ii  
Abstract .................................................................................................................................................................................................................. iii  
Table of Contents ................................................................................................................................................................................... v  
List of Tables .......................................................................................................................................................................................... x  
List of Figures ........................................................................................................................................................................................ xi  
Acknowledgments .................................................................................................................................................................................... xx  
1. Introduction........................................................................................................................................................................................................ 1  
   1.1 Background and motivation ....................................................................................................................................................... 1  
   1.2 Research objectives ................................................................................................................................................................. 4  
   1.3 Dissertation outline ................................................................................................................................................................. 6  
2. Dissolved oxygen prediction in the Bow River using linear regression with non-linear fuzzy number membership functions ......................................................................................................................... 8  
   2.1 Chapter introduction ......................................................................................................................................................... 8  
      2.1.1 Fuzzy sets and uncertainty analysis ............................................................................................................................ 9  
      2.1.2 Chapter objectives ....................................................................................................................................................... 12  
   2.2 Methods .................................................................................................................................................................................... 13  
      2.2.1 Site description and data collection ............................................................................................................................ 13  
      2.2.2 Fuzzy sets and membership functions ....................................................................................................................... 17  
      2.2.3 Multiple linear regression ............................................................................................................................................... 23  
   2.3 Chapter results and discussion ................................................................................................................................................ 24  
      2.3.1 Risk assessment .............................................................................................................................................................. 36  
   2.4 Chapter summary ................................................................................................................................................................. 39  
3. Two-dimensional prediction and uncertainty analysis of dissolved oxygen in the Bow River using fuzzy linear regression ............................................................................................................................................ 41  
   3.1 Chapter introduction ....................................................................................................................................................... 41  
      3.1.1 Chapter objectives .............................................................................................................................................................. 43  
   3.2 Methods .................................................................................................................................................................................... 43  
      3.2.1 Site description and data collection ............................................................................................................................ 43  
      3.2.2 Non-linear membership functions for fuzzy numbers ................................................................................................. 49
### 3.2.3 Multiple linear regression

3.3 Chapter results and discussion

3.4 Chapter summary

### 4.

#### Comparing the proposed and existing fuzzy linear regression methods

4.1 Chapter introduction

4.1.1 Uncertainty analysis using fuzzy numbers

4.1.2 Fuzzy linear regression

4.1.3 Chapter objectives

4.2 Methods

4.2.1 Site description and data collection

4.2.2 Creating fuzzy numbers

4.2.3 Fuzzy linear regression methods

4.2.4 Model implementation and comparison

4.3 Chapter results and discussion

4.4 Chapter summary

### 5.

#### Comparing fuzzy and Bayesian linear regression methods

5.1 Chapter introduction

5.1.1 Chapter objectives

5.2 Methods

5.2.1 Site description and data collection

5.2.2 Model implementation

5.2.3 Bayesian linear regression

5.2.4 Fuzzy linear regression

5.2.5 Quantifying model performance

5.3 Chapter results and discussion

5.3.1 Model performance comparison

5.3.2 Low DO analysis

5.4 Chapter summary

### 6.

#### Fuzzy linear regression for flood prediction and risk assessment

6.1 Chapter introduction

6.1.1 Chapter objectives
6.2 Methods......................................................................................................................... 132
6.2.1 Site description and data collection .................................................................................. 132
6.2.2 Error-in-variable linear regression .................................................................................. 136
6.2.3 Fuzzy linear regression .................................................................................................. 138
6.2.4 Quantifying model performance ..................................................................................... 143
6.2.5 Model implementation .................................................................................................... 145
6.3 Chapter results and discussion .......................................................................................... 147
6.3.1 Model performance comparison ..................................................................................... 147
6.3.2 Flood risk assessment ...................................................................................................... 160
6.3.3 Model application: 2013 flood data ............................................................................... 163
6.4 Chapter summary ............................................................................................................... 167
7. Dissolved oxygen prediction and risk analysis using a fuzzy neural network .................. 169
7.1 Chapter introduction ............................................................................................................ 169
7.1.1 Possibility theory & fuzzy numbers for uncertainty analysis .......................................... 171
7.1.2 Data-driven models for DO prediction ............................................................................ 172
7.1.3 Chapter objectives ......................................................................................................... 173
7.2 Methods............................................................................................................................. 174
7.2.1 DO monitoring & modelling in the Bow River ................................................................. 174
7.2.2 Fuzzy neural networks .................................................................................................... 177
7.2.3 Risk analysis for low DO ............................................................................................... 187
7.3 Chapter results and discussion .......................................................................................... 189
7.3.1 FNN structure ................................................................................................................ 189
7.3.2 FNN training, validation and testing results ................................................................. 192
7.3.3 Risk analyses and low DO tool ...................................................................................... 205
7.4 Chapter summary ............................................................................................................... 212
8. Dissolved oxygen prediction using possibility theory based neural network ............... 213
8.1 Chapter introduction ............................................................................................................ 213
8.1.1 Fuzzy numbers and data-driven modelling ................................................................. 217
8.1.2 Chapter objectives ......................................................................................................... 218
8.2 Methods............................................................................................................................. 220
8.2.1 Data collection ............................................................................................................... 220
8.2.2 Probability-possibility transformations ................................................................. 222
8.2.3 Fuzzy neural networks ............................................................................................. 230
8.2.4 Risk analysis using defuzzification ......................................................................... 237
8.3 Chapter results and discussion .................................................................................. 239
  8.3.1 Examples of the probability-possibility transformation ............................................ 239
  8.3.2 Training the fuzzy neural network ......................................................................... 246
  8.3.3 Risk analysis for low DO events ............................................................................ 258
8.4 Chapter summary ........................................................................................................ 262
9. Conclusions .................................................................................................................. 264
  9.1 General conclusions ................................................................................................ 264
    9.1.1 Constructing fuzzy numbers ............................................................................... 264
    9.1.2 Data-driven modelling using fuzzy numbers ......................................................... 265
    9.1.3 Risk analysis ...................................................................................................... 266
  9.2 Novel contributions .................................................................................................. 267
  9.3 Future research directions ....................................................................................... 268
Bibliography .................................................................................................................... 272
Appendix A MATLAB code for Chapter 2 ...................................................................... 293
  1. Constructing triangular shaped fuzzy numbers ......................................................... 293
  2. Constructing Gaussian shaped fuzzy numbers ......................................................... 293
  3. Constructing Extreme Value III shaped fuzzy numbers ............................................ 294
  4. Multiple Linear Regression ....................................................................................... 296
Appendix B MATLAB code for Chapter 3 ...................................................................... 297
  1. Constructing Gaussian shaped fuzzy numbers ......................................................... 297
  2. Estimating 2D velocity profile .................................................................................. 297
Appendix C Some definitions for Chapter 4 ..................................................................... 301
  1. Fuzzy numbers ......................................................................................................... 301
  2. The α-cut and the extension principle ....................................................................... 302
  3. Probability-possibility transformation ..................................................................... 303
Appendix D MATLAB code for Chapter 4 ...................................................................... 304
  1. Constructing linear fuzzy numbers ........................................................................... 304
  2. Constructing non-linear fuzzy numbers ................................................................. 304
3. Fuzzy linear regression – Tanaka .......................................................... 314
4. Fuzzy linear regression – Diamond .................................................... 316
5. Fuzzy linear regression - Khan ............................................................ 317

Appendix E Bayesian linear regression explained ........................................ 319
1. Introduction to Bayesian linear regression ........................................ 319
2. Natural conjugate priors .................................................................. 320
3. Non–informative priors .................................................................. 321
4. Independent priors ....................................................................... 322

Appendix F Summary of results for Chapter 5 ........................................... 324

Appendix G MATLAB code for Chapter 5 ................................................ 326
1. Constructing fuzzy numbers .......................................................... 326
2. Fuzzy linear regression .................................................................. 330
3. Bayesian linear regression – non-informative prior ....................... 334
4. Bayesian linear regression – conjugate prior ................................. 338
5. Possibility-Probability transformation .......................................... 342

Appendix H Input variable selection for Chapter 6 ..................................... 345

Appendix I Model performance results for Chapter 6 .................................. 347

Appendix J MATLAB code for Chapter 6 .................................................. 349
1. Error-in-variables regression .......................................................... 349

Appendix K MATLAB code for Chapter 7 .................................................. 352
1. Fuzzy neural network with crisp inputs ........................................ 352
2. Cost function ............................................................................ 356
3. Inverse transformation ................................................................ 358

Appendix L MATLAB code for Chapter 8 .................................................. 360
1. Optimum bin-size and transformation algorithms ....................... 360
2. Fuzzy neural network with fuzzy inputs ....................................... 366
3. Cost function ............................................................................ 371
List of Tables

Table 4-1 Summary of model performance criteria for each method......................... 90
Table 5-1 Details of data used to calibrate and validate each model in the recursive modelling ........................................................................................................ 106
Table 6-1 Rating system used to rank and compare performance of the models (adapted from Moriasi et al. (2007))........................................................................................................ 144
Table 6-2 : Rating system used to compare the average rank of each simulation........ 144
Table 6-3 Distance results for all lags for an $e = 20\%$ for each site ......................... 160
Table 7-1 MSE and NSE for the neural network at a membership level equal to 1 ...... 193
Table 7-2 Percentage of data (%) (%) captured within each fuzzy interval for the FNN model .................................................................................................................................. 193
Table 8-1 A summary of low DO events in the Bow River between 2004 and 2012 and the corresponding minimum acceptable DO concentration guidelines ......... 221
Table 8-2 : Selected values for $P_{CI}$ for the FNN optimisation.................................. 236
Table 8-3 The $E_{MSE}$ and $E_{NSE}$ for each subset of the fuzzy neural network .......... 246
Table 8-4 Percentage of data captured within each $\alpha$-cut interval for the three subsets of data................................................................................................................................. 247
List of Figures

Figure 2-1 An aerial view of the City of Calgary, showing the locations of the various monitoring sites (a) Water Survey of Canada site “Bow River at Calgary”, (b) Bonnybrook, (c) Pine Creek, and (d) Stier’s Ranch ................................................................................................................. 14

Figure 2-2 Summary of flow, temperature and DO in the Bow River from May 22 - Nov 22 2006 .................................................................................................................................................. 16

Figure 2-3 Examples of a fuzzy membership function and an $\alpha$-cut for the fuzzy set $\mathbf{A}$ ............................................. 18

Figure 2-4 Examples of fuzzy membership function for $Q$ (top; linear, Normal and Gumbel-based); $T$ (middle; Normal and Gumbel-based); DO (bottom; Normal and Gumbel-based) for June 19, 2007 ........................................................................................................................................ 21

Figure 2-5 Predicted values from Model 1 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$ ........................................................................................................................................ 26

Figure 2-6 Predicted values from Model 2 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$ ........................................................................................................................................ 27

Figure 2-7 Predicted values from Model 3 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$ ........................................................................................................................................ 28

Figure 2-8 Predicted values from Model 4 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$ ........................................................................................................................................ 29

Figure 2-9 Predicted values from Model 5 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$ ........................................................................................................................................ 30

Figure 2-10 Observed versus Predicted DO concentration for all five models for the 2006-07 and 2008 data sets; clockwise from top left: Model 1, Model 2, Model 4, Model 5, and Model 3 ........................................................................................................................................ 31

Figure 2-11 $R^2$, RMSE and MAE for 2006-07 data set ............................................................................................................. 32

Figure 2-12 $R^2$, RMSE and MAE for validation (2008) data set ...................................................................................................... 33

Figure 2-13 Values for regression parameters, a, b and c for each of the five models ......................................................... 33

Figure 2-14 Variability of observed data captured by each model (Left: 2006-07 data; Right: 2008 data) ........................................................................................................................................ 35

Figure 2-15 Identifying the days in 2008 when the predicted DO concentration had a possibility to fall below the 5 mg/L limit (top: Model 2; bottom Model 4) ................................................................................ 37
Figure 2-16 Predicted probability of DO concentration to be less than the Alberta guideline of 5 mg/L using Models 2 and 4 .......................................................... 38

Figure 3-1 An aerial view of the City of Calgary showing (1) Bearspaw Dam, (2) Pumphouse, (3) Cushing Bridge, (4) upstream of Bonnybrook WWTP, (5) downstream of Bonnybrook WWTP, (6) upstream of Fish Creek WWTP, and (7) downstream of Fish Creek WWTP ........................................................................................................ 44

Figure 3-2 Examples of a cross-section profiles and water surface elevation at (a) the Bearspaw, (b) Cushing Bridge and (c) Upstream of Fish Creek WWTP sites for July 2007. .......... 45

Figure 3-3 (a) Water temperature and (b) dissolved oxygen data collected by the City of Calgary .................................................................................................................. 46

Figure 3-4 Velocity profiles at three locations (a) BS, (b) CS and (c) USFC for July 2007 .......... 48

Figure 3-5 Examples of qₖ, the fuzzy version of q, at three locations (a) BS, (b)CS and (c) UPFC for July 2007 .................................................................................................................. 50

Figure 3-6 Comparison of predicted DO at two α levels and observed DO ......................... 53

Figure 3-7 Illustration of the low inter-cross-section variation of DO at each site, with a degree of membership of 1, for the July 2007 data................................................................. 54

Figure 4-1 An aerial view of the City of Calgary, highlighting the Bow River, the Water Survey of Canada (WSC) monitoring station, and the Bonnybrook (BB), Fish Creek (FC) and the Pine Creek (PC) wastewater treatment plants ................................................................. 63

Figure 4-2 Daily observations of DO concentration, water temperature, T, and discharge, Q, in the Bow River for the ice-free period (May - November) in 2008 ......................... 65

Figure 4-3 A comparison of T, Q and DO observations from September 12, 2006 and their transformations into fuzzy numbers. Note that on the right column, the symbols correspond to µ = 0, 0.25, 0.50, 0.75, and 1 ............................................................. 70

Figure 4-4 A comparison of the values of the three regression coefficients calculated using the Tanaka (top row), the Diamond (middle row), and the proposed method (bottom row) . 81

Figure 4-5 Comparison of observed and predicted DO for model construction (2006 data), using the (a) the Tanaka method, (b) the Diamond method, and (c) the proposed method. Note that the observed data (black circles) correspond to µ = 0L and 0R, while the predicted data are shown for µ = 0L and 0R (black crosses), and at µ = 1 (white circles) ............ 83
Figure 4-6 Comparison of observed and predicted DO for model construction (2007 data), using the (a) the Tanaka method, (b) the Diamond method, and (c) the proposed method. Note that the observed data (black circles) correspond to μ = 0L and 0R, while the predicted data are shown for μ = 0L and 0R (black crosses), and at μ = 1 (white circles) ............. 84

Figure 4-7 Comparison of observed and predicted DO for model verification (2008 data), using the (a) the Tanaka method, (b) the Diamond method, and (c) the proposed method. Note that the observed data (black circles) correspond to μ = 0L and 0R, while the predicted data are shown for μ = 0L and 0R (black crosses), and at μ = 1 (white circles) ............. 85

Figure 4-8 A comparison of observed and predicted DO, for the construction and verification datasets: (a) and (d) the Tanaka method, (b) and (e) the Diamond method, and (c) and (f) the proposed method ................................................................................. 87

Figure 4-9 A comparison of (a) Nash-Sutcliffe coefficient and (b) RMSE values for the Diamond and proposed method calculated at μ = 0, 0.25, 0.50, 0.75 and 1.................................................. 89

Figure 4-10 A comparison of d2 for the Diamond and proposed method calculated at μ = 0, 0.25, 0.50, 0.75 and 1........................................................................................................ 91

Figure 4-11 A comparison of observed and predicted DO for the Diamond and proposed method, using the model construction dataset. The Alberta guideline for low DO (5 mg/L) has been highlighted........................................................................................................ 93

Figure 4-12 A comparison of membership function shapes for the predicted DO for (a) July 23, 2006 and (b) May 22, 2007.................................................................................................................. 94

Figure 5-1 Aerial image of the City of Calgary showing the locations of (a) Bearspaw water treatment plant, (b) Bonnybrook, (c) Fish Creek, and (d) Pine Creek wastewater treatment plants, and two sampling locations (e) Stier’s Ranch and (f) Highwood. Note that the Bow River flows from the northwest to the southeast, through the centre of the City... 104

Figure 5-2 Evolution of the BLR parameters for the 1–day lag and resolution = 96 case. The top figure shows the change in mean values of the two coefficients and the variance; the bottom figure shows the change in the approximate pdf of the variables for the M02, M04 and M09 cases........................................................................................................ 114

Figure 5-3 Evolution of the FLR coefficients for the 1–day lag and resolution = 96 case, shown here for the for the M02, M04 and M09 cases................................................................. 116
Figure 5-4 Trend plots for validation results for M01, M04 and M09 for the 1 day lag, resolution = 96 case .................................................................................................................................................. 117

Figure 5-5 (top) a comparison of observed vs. predicted DO for (a) M01, (b) M04 and (c) M09; the black circles and lines represent the Bayesian predictions, the grey boxes represent the fuzzy predictions at μ = 0; and (bottom) a comparison of observed data (dots), Bayesian (black lines) and fuzzy predictions (triangle at μ = 0) for (d) M01, (e) M04 and (f) M09 ................................................................................................................................................ 119

Figure 5-6 A summary of the performance metrics for each model; as more data is added (a, d and g); as the lag is increased (b, e and h); as the resolution is reduced (c, f and i) for the 1 day lag case; the black markers are the mean Bayesian results, and the lines are the limits of the fuzzy results ........................................................................................................................................... 121

Figure 5-7 Observed minimum DO compared to the lower limits of the predicted Bayesian and fuzzy (at μ = 0) intervals: (a) when both methods capture the minimum value, (b) when only the fuzzy interval captures the minimum; (c) when neither method captures the minimum; (d) when only the Bayesian interval captures the minimum ......................................................................................... 124

Figure 5-8 Sample results of inverse transformation showing the probability of the fuzzy method to predict the minimum observed DO. The grey area represents the interval of observed DO values, the dashed lines show the upper and lower limits of Bayesian predictions, and the solid black line with dots shows the fuzzy number ................................................................................................................................ 125

Figure 5-9 Sample results of inverse transformation, showing the probability of the fuzzy method to predict DO to be below 6.5 mg/L. The grey area represents the interval of observed DO values, the dashed lines show the upper and lower limits of Bayesian predictions, and the solid black line shows the 6.5 mg/L warning level, and the black line with dots represent the fuzzy number .................................................................................................................................. 126

Figure 6-1 The locations of the three sites used in this chapter: (from L to R) Bow River at Banff, Bragg Creek at Elbow River, and Bow River at Calgary ......................................................................................... 133

Figure 6-2 A sample of results of transforming observed hourly flow rate to fuzzy mean daily flow rate, and fuzzy mean peak flow rate, for July 2, 2005. An e value of 20% was used for these conversions ........................................................................................................................................ 141

Figure 6-3 Results from the proof-of-concept models for a lag of 1 day, and e of 20%: trend plots of predicted daily peak flow rate from the error-in-variables and fuzzy linear regression
methods for (top) the calibration (shown for 2010 only), and (bottom) the validation
dataset (shown for 2005 only) for the Bow River at Calgary ........................................ 149

Figure 6-4 Results from the proof-of-concept models for a lag of 1 day, and e of 20%: observed
versus predicted peak flow rate plots for the error-in-variables (black circle with line) and
fuzzy linear regression (black boxes) methods for (left) the calibration, and (right) the
validation dataset for the Bow River at Calgary ................................................................. 150

Figure 6-5 Results from the proof-of-concept models for a lag of 1 day, and e of 20%: observed
peak flow rate, the error-in-variables regression line, 95% confidence intervals, and fuzzy
interval at μ = 0 for the (left) calibration and (right) validation dataset for Bow River at
Calgary .................................................................................................................................... 151

Figure 6-6 RSR, NSE and PBIAS values for the validation dataset for e = 20% proof-of-concept
models: the markers (circle, square, rhombus, and triangle) represent results for different
lags (1, 2, 3 and 7 days, respectively): Banff (top), Calgary (middle) and Bragg Creek
(bottom) ................................................................................................................................... 153

Figure 6-7 Results from the recursive model for a lag of 1 day, and e of 20%: trend plots of
predicted daily peak flow rate from the error-in-variables and fuzzy linear regression
methods for the validation dataset (for 2005) for the Bow River at Calgary .................. 155

Figure 6-8 Results from the recursive models for all lags, and e of 20%: observed versus predicted
peak flow rate plots for the error-in-variables (grey lines) and fuzzy linear regression
(black boxes) methods for the validation dataset (for the year 2005 only) for the Bow
River at Calgary ...................................................................................................................... 156

Figure 6-9 Results from the recursive models for all lags, and e of 20%: observed peak flow rate,
the error-in-variables regression line, 95% confidence intervals, and fuzzy interval at μ =
0 for the validation dataset (for the year 2005 only) for Bow River at Calgary ............. 157

Figure 6-10 RSR, NSE and PBIAS values for the validation dataset for all ten recursive models
with e = 20%: the markers (circle, square, rhombus, and triangle) represent result for
different lags (1, 2, 3 and 7 days, respectively), for each of the three sites: Banff (top),
Calgary (middle) and Bragg Creek (bottom) ......................................................................... 159

Figure 6-11 Possibility to probability transformations for high peak flow rate days (>Q_{P2%}), for
Bow River at Calgary, with a lag of 1 day, and e of 20% ..................................................... 162
Figure 6-12 Results from the test case of 2013 for a lag of 1 day, and e of 20%: trend plot of predicted daily peak flow rate from the error-in-variables and fuzzy linear regression methods for the Bow River at Calgary. The insert shows the days of interest in June 2013 when the highest flows were measured ................................................................. 164

Figure 6-13 Results from the test case of 2013 for a lag of 1 day, and e of 20%: (left): observed versus predicted peak flow rate plots for the error-in-variables (grey lines) and fuzzy linear regression (black boxes) methods; and (right) observed peak flow rate, the error-in-variables regression line, 95% confidence intervals, and fuzzy interval at $\mu = 0$ for the Bow River at Calgary........................................................................................................... 165

Figure 6-14 Possibility to probability transformations for high peak flow rate days ($> Q_{P2\%}$), for Bow River at Calgary in 2013 during the flood, with a lag of 1 day, and e of 20%....... 166

Figure 7-1 An aerial view Calgary, Canada showing the locations of (a) Water Survey of Canada flow monitoring site “Bow River at Calgary”, (b) Bonnybrook, (c) Fish Creek, and (d) Pine Creek wastewater treatment plants, and two water quality sampling sites (e) Stier’s Ranch and (f) Highwood.................................................................................................................. 175

Figure 7-2 A generic three-layer multilayer perceptron feed-forward ANN, with two input neurons, the hidden layer neurons, and one output neuron. $W_{IH}$ are the weights between the input and hidden layer, $W_{HO}$ are the weights between the hidden and output layer, $B_H$ are the biases in the hidden layer, and $B_O$ is the bias in the output layer...................... 179

Figure 7-3 Sample results of the coupled method to determine the optimum number of neurons in the hidden-layer and percentage of data for training, validation and testing subsets; the mean (solid black line) and upper and lower limits (in grey) of (a) the MSE for the test dataset for each number of hidden neurons; (b) the number of epochs for training; (c) the MSE for a range of training data size; and (d) the number of epochs for 10 hidden neurons ................................................................................................................................. 191

Figure 7-4 Sample results of the FNN optimisation algorithm to estimate the fuzzy number values of each weight and bias in the model.................................................................................. 195

Figure 7-5 A comparison of the observed and predicted crisp (black dots) and fuzzy intervals at $\mu = 0$ (grey lines) for minimum DO in the Bow River for the training, validation and testing datasets.................................................................................................................. 198
Figure 7-6 Time-series comparison of the observations and FNN minimum DO for 2004 and 2006 .......................................................... 201
Figure 7-7 Time-series comparison of the observations and FNN minimum DO for 2007 and 2010 ............................................................. 203
Figure 7-8 Detailed view of time series for minimum observed DO and predicted fuzzy DO for 2004, 2006, 2007 and 2010, corresponding to days with low DO events ................. 204
Figure 7-9: Membership functions of the predicted minimum DO and observed minimum DO for corresponding to the lowest DO observation for each year between 2004 and 2012..... 207
Figure 7-10 Results of the low DO identification and risk analyses tool for DO less than 5 mg/L .......................................................................................................................... 209
Figure 7-11 Results of the low DO identification and risk analyses tool for DO less than 6.5 mg/L .......................................................................................................................... 210
Figure 7-12 Results of the low DO identification and risk analyses tool for DO less than 9.5 mg/L .......................................................................................................................... 211
Figure 8-1 An aerial view of the City of Calgary, Canada showing the locations of (a) the flow monitoring site Bow River at Calgary (Water Survey of Canada ID: 05BH004), three wastewater treatment plants at (b) Bonnybrook, (c) Fish Creek, and (d) Pine Creek, and two water quality sampling sites (e) Stier’s Ranch and (f) Highwood. ......................... 216
Figure 8-2 An example of a three-layer multilayer perceptron feed-forward ANN, with two input neurons, three hidden layer neurons, and one output neuron. ........................................ 233
Figure 8-3 Sample results of probability-possibility transformation for flow rate, Q ...................... 241
Figure 8-4 Sample results of probability-possibility transformation for water temperature, T .... 244
Figure 8-5 Sample plots of the produced membership functions for the weights and biases of the fuzzy neural network....................................................................................... 249
Figure 8-6 A comparison of the predicted and observed minimum DO at the $\mu = 0$ interval (black line) and at $\mu = 1$ (black dots) ......................................................................................... 251
Figure 8-7 A comparison of the observed and predicted minimum DO trends for: (top) 2004, and (bottom) 2006............................................................................................................. 254
Figure 8-8 A comparison of the observed and predicted minimum DO trends for three sample years: (top) 2007 and (bottom) 2010. ....................................................................................... 255
Figure 8-9 Zoomed in views of the trend plots for four sample year corresponding to important periods with low DO occurrences........................................................................................................ 257

Figure 8-10 Sample plots of low DO events and the corresponding risk of low DO calculated using a possibility-probability transformation for the (top) 5 mg/L, (middle) 6.5 mg/L, and (bottom) 9.5 mg/L guideline ........................................................................................................ 261

Figure F-1 Results for M01 to M09 using BLR and FLR models for the 1 day lag at all resolutions......................................................................................................................................... 324

Figure F-2 Results for M01 to M09 using BLR and FLR models for the 2 day lag at all resolutions......................................................................................................................................... 324

Figure F-3 Results for M01 to M09 using BLR and FLR models for the 3 day lag at all resolutions ......................................................................................................................................... 325

Figure F-4 Results for M01 to M09 using BLR and FLR models for the 7 day lag at all resolutions ......................................................................................................................................... 325

Figure H-1 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 0 day lag .................................................. 345

Figure H-2 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 1 day lag .................................................. 345

Figure H-3 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 2 day lag .................................................. 345

Figure H-4 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 3 day lag .................................................. 346

Figure H-5 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 7 day lag .................................................. 346

Figure I-1 RSR, NSE and PBIAS values for the validation dataset for $e = 10\%$ proof-of-concept models: the markers (circle, square, rhombus, and triangle) represent result for different lags (1, 2, 3 and 7 days, respectively): Banff (top), Calgary (middle) and Bragg Creek (bottom)................................................................................................................................. 347

Figure I-2 RSR, NSE and PBIAS values for the validation dataset for $e = 5\%$ proof-of-concept models: the markers (circle, square, rhombus, and triangle) represent result for different lags (1, 2, 3 and 7 days, respectively): Banff (top), Calgary (middle) and Bragg Creek (bottom)................................................................................................................................. 347
Figure I-3 RSR, NSE and PBIAS values for the validation dataset for all ten recursive models: the markers (circle, square, rhombus, and triangle) represent result for different lags (1, 2, 3 and 7 days, respectively). Error values of $e = 10\%$ (left), 10\% (middle), and 20\% (right) were used, for each of the three sites: Banff (top), Calgary (middle) and Bragg Creek (bottom).

Figure I-4 RSR, NSE and PBIAS values for the validation dataset for all ten recursive models: the markers (circle, square, rhombus, and triangle) represent result for different lags (1, 2, 3 and 7 days, respectively). Error values of $e = 10\%$ (left), 10\% (middle), and 20\% (right) were used, for each of the three sites: Banff (top), Calgary (middle) and Bragg Creek (bottom).
Acknowledgments

I would like to thank my supervisor, Dr. Caterina Valeo, for giving me the opportunity to pursue my PhD research under her guidance. Dr. Valeo allowed me to work with independence and encouraged me to find my own research direction, for which I am extremely grateful. Her patience and support during the initial, exploratory phase was invaluable, as was her recommendation to explore fuzzy set theory for modelling uncertainty and risk assessment. I particularly enjoyed the directed studies course she taught on fuzzy set theory – a great experience that helped cement my knowledge on the topic. Her sustained support over the last five years has been instrumental in completing this dissertation.

This research would not have been possible without funding from a number of agencies; I would like to thank: the Natural Sciences and Engineering Research Council of Canada for a two-year Postgraduate Scholarship; the British Columbia Ministry of Advanced Education, Innovation and Technology for a one-year Graduate Student Fellowship; the University of Victoria for a one-year Graduate Award and a President’s Research Scholarship; and the University of Calgary for a Dean’s Entrance Scholarship. I am also grateful for the Graduate Support and numerous Teaching Assistantships provided by the Department of Mechanical Engineering at the University of Victoria.

In addition, I would like to acknowledge the individuals and agencies that provided data – essential for any data-driven research – for this dissertation: Dr. C. Ryan and Dr. A. Chu at the University of Calgary; Mr. M. Wang at Alberta Environment; Mr. F. Frigo and numerous others at the City of Calgary – Water Resources; and lastly the staff at the Water and the Climate Offices at Environment Canada.

Finally, I would like to thank: Dr. J. He at the University of Calgary for a number of interesting discussions over the years, particularly on calibration-validation procedures; Dr. S. Alvisi at the Università degli Studi di Ferrara for providing the MATLAB code used as the basis for the work presented in Chapters 7 and 8; Mr. B. Kent at the University of Victoria for helping with MATLAB issues; and my supervisory committee: Dr. A. B. Albu, Dr. B. J. Buckham, and Dr. Y. Shi, for their feedback on my proposal which helped define the direction of this dissertation.
1. Introduction

1.1 Background and motivation

The supply of water for potable consumption, industrial and recreational use, and for the provision of adequate flows for sewage disposal and habitat protection is essential for all communities. The effects of urbanisation on hydrology are severe and extend beyond the limits of urban areas. An increase in population leads to an increase in water demand, resulting in problems in providing adequate water resources. Reduced availability, supply and access to water resources resulting from urbanisation compounds this problem (Hall, 1984). An increase in building density results in an increase in the amount of impervious areas. This alters the natural drainage patterns in urban areas. An increase in impervious area means a greater proportion of rainfall now contributes to runoff (the overland flow after rainfall). Storm sewers and the modification of natural streams result in the runoff being conveyed to waterways more rapidly, allowing for little storage in the natural landscape as was prevalent in undeveloped conditions. Along with this, impervious surfaces reduce the rate of groundwater recharge, resulting in lower baseflow in waterways. The increase in runoff volume, higher flow velocities and reduced baseflow affects the timing of hydrographs, results in higher peak flow rates, and increases the risk of floods (Hall, 1984).

The water quality aspect of hydrology is also strongly linked to urbanisation. Water pollution increases as a response to population growth due to the increase in waterborne waste and sewage. With an increase in impervious surfaces, many contaminants are washed from streets, roofs and paved areas to waterbodies during rain events. This phenomenon, coupled with the fact that there is lower baseflow in rivers and streams due to urbanisation, intensifies the deterioration of water quality in urban areas (Hall, 1984). Typically, parameters like total suspended solids, biochemical oxygen demand, nutrient concentration and dissolved oxygen (DO) are used as water quality parameters in freshwater sources. DO, is a common assessor of water quality, and is widely used to gauge the overall health of aquatic ecosystems because it is essential for the survival of many aquatic organisms (Dorfman & Jacoby, 1972). However, it is highly susceptible to, and influenced by urbanisation. Low DO can lead to hypoxic and anoxic conditions,
leading to the possibility of fish kills and the degradation of drinking water supplies (Kramer, 1987).

Many municipalities and governments have enacted water quality guidelines and criteria to protect aquatic ecosystems due to the effects of urbanisation. Typically a maximum or minimum value is assigned that parameters like DO must adhere to for the protection of ecosystems (Migliaccio & Angelo, 2010). In order to ensure compliance to these guidelines, numerical water quality models are designed to simulate the natural environment. Typically these models are physically-based models, i.e. they try to replicate natural phenomenon using simplified mathematical constructs. The broad objective of these models is to predict the effect of implementing various water management plans on water quality, and to be able to understand and forecast what conditions may lead to undesirable outcomes (e.g. a breach of guideline for a particular contaminant).

However, physically-based models are hard to develop as they are extremely data intensive, require a complete understanding of the physical system being modelled, including the complex relationships that exist between numerous factors. In fact in highly urbanised areas, it may be impossible to realistically account for all the factors that actually impact water quality or flow rate in rivers, either directly or indirectly. Even if these relationships are understood, it is still difficult and impractical to collect the necessary data to effectively calibrate these models.

In contrast to this, data-driven modelling relies on characterising a system with limited or no assumptions regarding the nature of the physical system being modelled. Typically, a model can be defined on the basis of generalised connections between input and output variables (Solomantine & Ostfeld, 2008). These models do not suffer from the same problems as physically-based models, i.e. the difficulty in calibrating variables and mathematically representing complex physical relationships.

The nature of data-driven models lends itself well to using secondary or indirect factors to predict water quality parameters, such as DO. The aim is not to explicitly characterise the system, but to use these factors to adequately forecast the variable in question, even if there is no direct relationship between them. For example, research (He et al., 2011; Rankovic et al. 2010) has shown that abiotic factors, which are non-living, physical and...
chemical attributes such as flow, temperature and radiation, can be used to effectively predict DO concentration using data-driven modelling techniques. These secondary factors are routinely collected by municipalities; thus, data-driven models can make use of these pre-existing datasets as inputs. One benefit of this approach is that the detailed complexities of the actual system and the parameterisation issues that plague conceptual or physically-based modelling are avoided. Also, the rationale for identifying which, and to what extent, these abiotic factors influence DO concentration, is that if the former can be controlled, the latter can be indirectly improved.

However, data-driven modelling has intrinsic uncertainties associated with it, which must be identified and propagated through the model before decisions based on model output can be made (Shrestha & Nestmann, 2009). Though probability theory is the most popular method used for uncertainty analysis in numerous engineering applications, other frameworks such as fuzzy set theory, possibility theory, Dempster-Schafer theory, imprecise probabilities, and upper and lower probabilities may be more appropriate. The use of these methods has not been as widespread as have methods based on probability theory, e.g. Bayesian inference methods. The resistance to these methods has been attributed to a “lack of formal education” (Bárdossy et al., 2006) and to a “lack of common language” between different fields (Dubois & Prade, 1993). Of these methods, fuzzy set theory has been the most popular method for many applications. The theory was introduced by Zadeh (1965) to express imprecision in complex systems and it is a consistent body of mathematical tools: it is a theory of fuzziness rather than a “fuzzy” theory. Several bridges between fuzzy set and probability theory have been proposed, and possibility theory is one considered to be at the cross-roads between the two. In the same way as there are multiple interpretations of probability theory (e.g. subjective vs. frequentist), there are multiple interpretations of fuzzy set theory. When interpreted from the basis of possibility theory, probability and fuzzy sets can be jointly considered as an enlarged framework for modelling uncertainty (Dubois & Prade, 1993). Fuzzy numbers are one particular case of using probability and possibility to define uncertainty (Dubois and Prade, 2015).

In general, a fuzzy number represents a set of all possible values that a variable may take on rather than a single deterministic value that is typically used. Fuzzy numbers are
particularly suited to applications where data is scarce, data from multiple sources is combined, and highly imprecise data is used. Fuzzy numbers are one way of dealing with the type of uncertainties associated with data-driven models. However, before fuzzy numbers can be used, it requires that ordinary (i.e. non-fuzzy) mathematical tools be adapted to handle fuzzy quantities. In this research, new methods to create fuzzy numbers from observed data are explored, and novel data-driven models techniques are created that use fuzzy numbers to predict environmental factors such as DO concentration or streamflow in rivers.

The use of fuzzy numbers ensures, for example, that instead of dealing with mean values of observations, such as mean daily DO or daily peak streamflow, the full spectrum of possible values of a quantity are incorporated into model construction and prediction. This lends itself well to risk analysis, and fuzzy numbers can be used to identify when and under what conditions there is heightened risk of either low DO or high streamflow. This is something that would not be possible using a typical data-driven approach. In this research, methods of risk analysis using fuzzy numbers are developed and used to assess the risk of low DO and high peak flow rates.

1.2 Research objectives

The following research objectives have been identified based on gaps in the current literature:

1. Develop a method to construct fuzzy numbers from high-frequency observations.

The first objective of this research is to develop a consistent and uniform technique to create fuzzy numbers from highly uncertain, observed environmental data. Fuzzy numbers are defined using a membership function \( \mu(x) \), which indicates to what degree an element \( x \) has a membership in the fuzzy number. Typically, these membership functions are linear or triangular shaped. However, the use of these types of functions is not appropriate to capture the nuances of the variability in many environmental variables, such as flow rate, water temperature and DO. Thus, various non-linear forms must be investigated. The developed method must follow fuzzy number and possibility theory principles. While generic methods to convert probability distribution functions to possibility distributions exist, methods when the distribution function is unknown need to
be analysed. The impact of measurement error in the data and sampling resolution on fuzzy number membership functions is needed.

2. **Develop data-driven models for environmental prediction that use fuzzy numbers to quantify and propagate the uncertainty in the input, output and parameters of the model.**

The second objective is to develop data-driven methods to predict environmental variables (specifically DO concentration) in a riverine environment. The model must use fuzzy numbers to characterise and propagate uncertainty. Two types of data-driven models are proposed, namely fuzzy linear regression and fuzzy neural networks. For the fuzzy linear regression, the inputs, output and model coefficients are all fuzzy numbers. Note that in traditional linear regression these variables are crisp numbers (i.e. non-fuzzy), and the coefficients are calculated by a least-squares approach. Thus, a conceptually equivalent approach to least-squares is needed for the fuzzy regression case. For the fuzzy neural networks case, the inputs, output and model weights and biases are all fuzzy numbers, representing the total uncertainty of the system. A new method to calibrate the model parameters is needed to deal with the fuzzy numbers. Note that for both models, fuzzy numbers represent the quantification of uncertain values rather than fuzzy logic based applications that are common in hydrology. Thus, this fuzzy neural network method differs from existing neuro-fuzzy methods in that the fuzzy number variables quantify the uncertainty rather than for fuzzy logic (if-then rules) purposes as is more common.

Note that for both models, fuzzy numbers represent the quantification of uncertain values rather than fuzzy logic based applications that are common in hydrology. The models will be evaluated based on their performance against observations as well as against other non-fuzzy modelling techniques. The fuzzy linear regression method will be compared to ordinary linear regression, Bayesian linear regression, error-in-variables regression and two existing fuzzy linear regression methods.

The fuzzy neural network method will be compared to an existing fuzzy neural network method, and a non-fuzzy network approach. An optimum method of selecting the network architecture will be explored.
New evaluation criteria that use fuzzy numbers need to be established. Different application of the models will be explored, including an extension from point-source to 2-dimensional predictions, recursive algorithms for real-time updating, and a time series based approach where lagged dependent variables are used as input. Lastly, the robustness of the regression method will be tested by a different application, namely peak flow rate prediction.

3. **Develop a technique to use fuzzy number outputs for risk assessments.**

Information gained from fuzzy number output from the data-driven models will be used to quantify the risk of low DO. The new representation of data will provide information that was not available in previous techniques. There is a need to develop a method to clearly identify and articulate the risk of low DO using possibility theory. In particular, different combinations of the abiotic input data will be used to identify the different conditions that result in low DO. This information will be used to create a risk analysis tool for water resource managers.

**1.3 Dissertation outline**

Each chapter of this dissertation is a modified version of a journal or conference article. Thus, each chapter includes its own introduction, literature review, specific objectives, methods, results and discussion. Chapter 2 begins with an exploration of the use of non-linear membership functions to represent the uncertainty in input data used for DO prediction. These fuzzy number inputs are implemented in a fuzzy linear regression method in the Bow River in Calgary, Alberta, Canada. Outputs from the models are used to analyse the risk of low DO. Chapter 3 uses the calibrated model presented in Chapter 2 to extend DO prediction from a point-source to a number of different 2-dimensional cross-sections of the Bow River in Calgary. The variability of predicted DO across the cross-section of the River is analysed.

In Chapter 4 a formal, mathematical definition of fuzzy linear regression is proposed and compared to two existing methods of fuzzy linear regression. In addition, a new method to create fuzzy numbers from observations is proposed. The impact of this method is analysed by comparison with other fuzzy linear regression techniques.
The fuzzy linear regression method develop in Chapter 4 is then compared to a probability based uncertainty analysis approach, namely Bayesian linear regression, in Chapter 5. Instead of abiotic factors as inputs, lagged DO at different time steps (i.e. an autoregressive method) is used as the input. The impact of sampling resolution on the data-driven methods is explored, as well as on the method to create fuzzy numbers (first established in Chapter 4). A new method to calculate the cumulative probability from the possibility distribution is developed and used to predict the risk of low DO.

In Chapter 6, the fuzzy regression model is compared against an error-in-variables regression model to predict peak flow rate in the Bow River. Lagged mean-daily flow rate is used as the input to the model. The impact of the error value used to construct the fuzzy numbers is explored. A number of different modelling performance metrics are also analysed. The risk analysis method from Chapter 5 is redeveloped to predict the risk of floods in the Bow River.

Chapters 7 and 8 focus on refining an existing fuzzy neural network model. In Chapter 7 crisp abiotic inputs are used to predict minimum DO in the Bow River using the neural network. An optimisation algorithm based on possibility theory is proposed to calibrate the fuzzy weights and biases of the network. A coupled method to optimise the network architecture is developed. The previously developed possibility-probability transformation (from Chapter 5) is used to predict the risk of low DO. The calibrated network is then used to create a risk analysis tool for water resource managers.

Chapter 8 further refines the fuzzy neural network developed in Chapter 7 by introducing fuzzy number inputs rather than crisp inputs. To create the fuzzy inputs, the method proposed in Chapter 4 to construct fuzzy number is further advanced by including a bin-size optimisation algorithm. The risk of low DO for various scenarios is analysed using the new model.

Lastly, Chapter 9 summarises the major conclusions from each chapter and novel contributions of this dissertation. Suggestions for future research are also listed.
2. Dissolved oxygen prediction in the Bow River using linear regression with non-linear fuzzy number membership functions

2.1 Chapter introduction

Dissolved oxygen (DO) is an important water quality parameter for assessing aquatic ecosystem health. It is vital for the survival of biotic (living) communities, which are highly influenced by DO concentration and fluctuation (He et al., 2011; Altunkaynak et al., 2005). Biotic and abiotic (non-living, chemical and physical attributes) factors, directly and indirectly influence DO concentration and fluctuations in riverine environments. Biotic factors such as algal and macrophyte growth cause variations in DO due to photosynthesis and respiration (Hauer & Hill, 2007; Pogue & Anderson, 1995). Abiotic factors, including pH, temperature, turbidity, water levels and climate also influence DO in aquatic systems. Urbanization can also effect DO concentration in aquatic ecosystems, possibly through an increase in nutrient concentration.

In the Bow River in Calgary, Alberta, low DO concentrations have been observed in a highly urbanised region downstream of a wastewater treatment plant’s (WWTP) effluent discharge. The City of Calgary has made efforts to reduce nutrient loading in the Bow River in an attempt to reduce the occurrence of low DO concentrations and maintain levels above the Alberta provincial guidelines for one-day minimum DO (5 mg/L) (Alberta Environment Protection, 1997). Consequently, both monitoring and modelling of DO concentrations are significant components of the City of Calgary’s compliance program. This has resulted in major efforts to model DO throughout a complex ecosystem with little understanding of the factors governing DO. In a recent study by He et al. (2011), daily DO concentration in the Bow River was modelled by analysing abiotic factors (temperature, radiation, flow, turbidity, and nutrients) that seemed to govern biotic influences on DO concentration. They suggested that if important abiotic factors could be identified and controlled, the DO concentration downstream of the WWTPs could be improved. Two data-driven modelling techniques, multiple linear regression (MLR) and a multiple layer perceptron (MLP) neural network were used to predict

---

minimum daily DO and daily DO variation. He et al. (2011) found that the minimum daily DO was best predicted by using these two parameters, while the daily DO variation (maximum daily DO minus minimum daily DO) was best predicted by including radiation as well. The data-driven techniques used were able to adequately predict DO for the given environment in the absence of sufficient biotic data and a complete understanding of the physical processes underlying DO trends in the river. However, their study did not consider the impacts of variability and uncertainty in the underlying data on the predictions.

2.1.1 Fuzzy sets and uncertainty analysis

Data-driven modelling has inherent uncertainties associated with it, and the proper identification and propagation of this uncertainty is critical for understanding and evaluating model prediction (Shrestha & Nestmann, 2009). Scarcity in data, measurement errors and variability over time scales (lower than the modelling time scale) can contribute to uncertainty in modelling projections (Zhang et al., 2009; El-Baroudy & Simonovic, 2006). Variability and uncertainty can have different meanings, e.g. the exact value of a variable may be known at a given location (no uncertainty) but the site-to-site value may be different (i.e. parameter variability). Uncertainty can be reduced by increasing the sampling frequency but variability cannot (Delhomme, 1979). Additionally, uncertainty may be increased when data from multiple sources are used, integrated and propagated (Porter et al., 2000; Shrestha & Nestmann, 2009; Shrestha & Simonovic, 2010; Hermann, 2011).

Fuzzy set theory (Zadeh, 1965) is one technique for representing and aggregating uncertainty. It provides a framework to analyse and propagate uncertainty and is growing in popularity (Bárdossy et al., 1990; Alvis et al., 2006; Shrestha & Nestmann, 2009). A fuzzy set can be described as a set without crisp boundaries that contain a band of continuous values that a variable can have with membership values between 0 (not belonging to the set) and 1 (completely belonging to the set), that are defined by the membership function (µ) (Zhang & Achari, 2010a). This is in contrast to the typical representation of variables as crisp-sets that are defined as either a member (µ = 1) or non-member (µ = 0). The fuzzy set method is useful for treating uncertainties when data is limited or imprecise (El-Baroudy & Simonovic, 2006; Guyonnet et al., 2003; Zhang &
Achari, 2010a; Huang et al., 2010). In this case, a probabilistic representation of parameter values or the parameter distribution cannot be developed as the exact values of the parameter are unknown and partial information (due to the small number of measurements) cannot be correctly represented in the probabilistic framework (Zhang, 2009; Shrestha & Simonovic, 2010). It is possible to transform parameter uncertainty from a traditional probabilistic representation to possibilistic functions when the exact probability distributions of a variable are unknown (Xia et al., 2000). Fuzzy techniques can then be used to propagate uncertainty in variables (Zhang & Achari, 2010b; Zhang & Achari, 2010a). However, there is no uniform technique to fuzzify or to develop the membership functions to represent the uncertainty and variability in a parameter (Zhang & Achari, 2010a).

The literature regarding applications of fuzzy sets abounds with linear membership functions because they are simple to implement and understand, and are believed to provide a reasonable representation of the possible values of uncertainty. Normal (i.e., Gaussian) representations are also common (Duch, 2005) and non-linear membership functions have been shown to be appropriate in data-driven modelling techniques (e.g. MLP applications) and when uncertainty distributions are fairly well known such as in medical data (Duch, 2005). Few however, have explored the use of Normal or other distributions for characterising the uncertainty in environmental data such as hydrometric and climatic data leading to risk assessment of a highly influenced biotic factor. Previous studies have shown that fuzzy sets are appropriate to represent uncertainty and variability for predicting discharge and other environmental factors, such as dissolved oxygen and contaminant transport (Bárdossy et al., 1990; Mujumdar & Sasikumar, 2002; Altunkaynak et al., 2005; Shrestha et al., 2007; Giusti & Marsili-Libelli, 2009; Shrestha & Nestmann, 2009; Zhang et al., 2009; Shrestha & Simonovic, 2010; Wang et al., 2012).

Shrestha & Simonovic (2010) used fuzzy sets to represent uncertainty in stage-discharge measurements and used the fuzzy-extension principal to aggregate uncertainty from different sources to represent an overall uncertainty in discharge measurements. For this case, linear membership functions were used, but the use of fuzzy representation of uncertainty was not extended beyond the parameter under investigation. Fuzzy analysis was used to provide more accurate results of measured discharge. Shrestha & Nestmann
(2009) used fuzzy sets to represent uncertainty in discharge measurements. Both physically-based and data-driven models were used to predict future discharge using fuzzy historical discharge data. The fuzzy representation of aggregated input discharge was a generalised bell function while the output of the models was constructed as a linear function. Similarly, Shrestha et al. (2007) used fuzzy sets to represent and propagate uncertainty in stage-discharge measurements and prediction. A linear and non-linear fuzzy regression method was used to define the membership function of discharge. A physically-based model was then used to predict peak discharge and the associated uncertainty in river channels.

Fuzzy sets were used by Zhang et al. (2009) to represent uncertainty in aquifer porosity and hydraulic conductivity. A physically-based model was then used to trace the fate of a contaminant in the aquifer. A random field generator was used to create variability in the input parameters, before transforming them to fuzzy numbers from probabilistic functions. Though the generation of the input fuzzy numbers was not from a linear process, the final representation was a three point function, with an interval based support (i.e., upper and lower bounds) and a most-likely value, essentially creating a linear membership function. The approach significantly altered the fate of the contaminant being modelled, compared to the no-uncertainty approach. The output was represented as a probability distribution based on a linear fuzzy number.

Mujumdar & Sasikumar (2002) defined the risk of the occurrence of low DO as a fuzzy event, and represented it with a linear membership function. Randomness associated with the measurement of low DO was then linked to the probability of the occurrence of the fuzzy event. A physically-based model to predict DO was used by Giusti & Marsili-Libelli (2009). The model used fuzzy pattern recognition to capture the variability of DO in historical data sets. Near-term DO dynamics could then be predicted using the historical fuzzy data series as well as a number of time-varying crisp physical parameters. The crisp parameters represented a number of biotic processes that occur in a lagoon environment, such as the photosynthetic productivity rate. Due to the nature of the input parameters used, the DO variability could not be adequately predicted beyond a couple of days. Altunkaynak et al. (2005) modelled monthly DO variation using a data-driven approach (artificial neural network) while incorporating fuzzy logic. In this case, the
uncertainty or variability in DO was not represented as fuzzy sets or fuzzy numbers, but rather a Mamdani-type approach was used to create a fuzzy inference system to predict monthly DO variation. Historical variation in DO was used to define rules to predict future DO concentration based on historical observations.

Overall, the use of fuzzy sets to represent uncertainty in predictive models has largely been limited to typical linear representation of the membership function in current literature. There is a need to study the effects of using different forms of the fuzzy membership function to represent and propagate uncertainty. Non-linear memberships have the potential to better capture the variation and error seen in data such as river discharge, water temperature and DO concentration. Data-driven and regression based prediction of DO concentration and its variation using fuzzy sets is limited. Typically, the uncertainties of the independent regressors have not been represented as fuzzy sets (linear or non-linear). Additionally, the model outputs need to be represented as a probabilistic set rather than a deterministic value. This is because a single output value is not sufficient enough to capture the variability in daily DO; a range of predicted values will be able to highlight potential risks and can mimic observed values better. The use of fuzzy sets has the potential to satisfy these needs.

2.1.2 Chapter objectives

In this chapter, the use of fuzzy sets to characterise and propagate the variability and uncertainty is explored. The objectives are to characterise the significant abiotic parameters (flow and temperature) that influence DO concentration, as fuzzy sets and to then determine if fuzzy techniques can adequately predict daily DO variation and uncertainty using data-driven methods. Five different combinations of linear and non-linear membership functions are used to transform the data into a fuzzy form, to determine if non-linear based membership functions are more appropriate considerations for DO prediction.

The approach used in this chapter is to develop a MLR model based on the finding by He et al. (2011) that predicts DO concentrations in an urbanised reach of the Bow River in Calgary, Alberta using abiotic factors. A new fuzzy set approach is used to characterise and propagate the uncertainty in the MLR model, using several different approaches to represent the fuzzy set functions in each dependent (DO) and independent variable (i.e.,
flow, Q and water temperature, T, the regressors). The output of the fuzzy model is compared to the output of the non-fuzzy case as a means of estimating model performance and evaluating the effectiveness of the fuzzy set based approach. Finally, the fuzzy output from the data-driven model is used to assess the risk or probability of predicted DO concentration to be less than a threshold value (i.e. minimum daily limit of 5 mg/L).

2.2 Methods

2.2.1 Site description and data collection

The headwaters of the Bow River are located on the eastern slopes of the Rocky Mountains in Alberta, western Canada. The river flows south-eastwardly through three distinct physiographic regions: the Rocky Mountains, the foothills and the prairies. Calgary, Alberta, the location for this study, is located approximately 200 km downstream of the headwaters of the Bow River. It is a rapidly growing city, with a population of approximately 1 million. The river flows through a largely undeveloped area before entering the City of Calgary’s limits and is regulated by Bearspaw dam upstream of the city limits. Nose Creek, Elbow River, Fish Creek and Pine Creek are tributaries that confluence with the Bow River within the city limits. Other significant pollutant point-sources within Calgary are the Fish Creek and Bonnybrook WWTPs. The average annual discharge in the Bow River is approximately 90 m$^3$/s; the average width and depth of the river within city limits are 100 m and 1.5 m, respectively.

For this study, flow data collected by the Water Survey of Canada (WSC) monitoring station 05BH004 (shown in Figure 2-1) at hourly intervals was used for the period of May to November, from 2006 to 2008 (Environment Canada Water Office, 2013). This period represents the ice-free flow period. Low flows are seen in the river until May when snowpack melt from the Rocky Mountains start to contribute to flows. Typically, discharge in the river peaks in mid-June due to the combination of snowmelt from the mountains and precipitation within the watershed. Generally, the flow then reduces until late October when baseflow conditions are reached. Low flows are seen from December

---

2 The MATLAB code associated with this section is included in Appendix A.
onwards until the onset of spring freshet, the following year (Alberta Environment River Basins, 2013).

Figure 2-1 An aerial view of the City of Calgary, showing the locations of the various monitoring sites (a) Water Survey of Canada site “Bow River at Calgary”, (b) Bonnybrook, (c) Pine Creek, and (d) Stier’s Ranch

Water quality data, specifically DO and T were collected using continuous YSI sondes from Pine Creek (for 2006 and 2007) and from Stier’s Ranch (2008). Pine Creek is approximately 20 km downstream of the Bonnybrook WWTP as seen in Figure 2-1. The
sondes were calibrated on a weekly basis and recorded water quality data at 15 or 30 minute intervals.

Data for each variable was collected for the period between May 22 through November 22 for the years 2006, 2007 and 2008. This period was selected for the analysis as it had the largest component of data points available for each variable over the monitoring period and also represented the ice-free period in the Bow River as mentioned above (no data was collected in the ice covered period for 2007 and 2008). Also, on an annual basis, the occurrence of low DO concentration tends to occur during the summer months, corresponding to low flows and high temperatures, i.e., the most critical period for low DO during the year. High daily minimum DO concentration and low daily DO variability are typically seen in the winter, during the ice covered period, when the risk to aquatic habitat is negligible. Missing data were ignored for the analysis. The average (minimum and maximum) Q for 2006, 2007 and 2008 was 99.10 (51.97 – 254.90), 127.44 (48.67 – 352.17) and 131.83 (60.83 – 314.49) m³/s, respectively. The average (minimum and maximum) T was 12.96 (0.29 – 22.44), 12.72 (-0.05 – 21.83) and 11.72 (0.86 – 19.99) °C, and the average (minimum and maximum) DO concentration was 9.52 (3.81 – 16.37), 9.79 (5.45 – 18.32) and 10.02 (6.01 – 16.57) mg/L, for 2006, 2007 and 2008, respectively. Note that DO concentration fell below the provincial guidelines of 5 mg/L occasionally in 2006, but not in 2007 or 2008. None of the parameters exhibited a significant trend (p = 0.05) during the monitoring period. Time series plots of the 2006 data set are shown in Figure 2-2 below.
Figure 2-2 Summary of flow, temperature and DO in the Bow River from May 22 - Nov 22 2006
2.2.2 Fuzzy sets and membership functions

The methodology used to create the different membership functions is described as follows. First, a brief description of fuzzy sets and the extension principle is given. A fuzzy set $\hat{A}$ has elements $[a_1, a_2, \ldots, a_n]$, each of which have memberships $\mu$ where $0 \leq \mu \leq 1$. When fuzzy sets are used to represent uncertainty, the membership of an element $\mu_{\hat{A}}(a_i) = \alpha$, means that an element in $\hat{A}$ has a value of $a_i$ with a possibility of $\alpha$. Also for any elements $a_1$ and $a_2$ in $\hat{A}$, $\mu_{\hat{A}}(a_2) > \mu_{\hat{A}}(a_1)$ represents the higher possibility of the variable to have a value of $a_2$ compared to $a_1$. A fuzzy set is normal if there exists at least one element $a_i$ in $\hat{A}$ with $\mu = 1$. A fuzzy set is said to be convex if for three real numbers, $a_1, a_2$ and $a_3$, where $a_1 < a_2 < a_3$, the membership at $a_2$, $\mu_{\hat{A}}(a_2) \geq \min(\mu_{\hat{A}}(a_1), \mu_{\hat{A}}(a_3))$ (Zhang et al., 2009; Wang et al., 2011). A fuzzy number is represented as a normal and convex fuzzy set and an example is shown in Figure 2-3.

Detailed description of fuzzy set theory is available in, Zadeh (1987), Novak (1989), Kosko (1997) and others.

The extension principle based fuzzy arithmetic can be used to generalise crisp mathematical operators to fuzzy sets (Zadeh, 1978). A widely used and simple method to deal with fuzzy arithmetic is known as the alpha-level cut ($\alpha$-cut) (Shrestha & Simonovic, 2010; Zhang et al., 2009; Zhang & Achari, 2010b; Wang et al., 2011; Wang et al., 2012). The $\alpha$-cut methodology reduces a fuzzy membership function to an interval containing crisp numbers. It provides a bridge to connect a fuzzy set to a classical set. Formally, for a fuzzy set $\hat{A}$, there exists an interval or subset $\hat{A}_{\alpha}$ corresponding to $\mu_{\hat{A}}(a) = \alpha$ (the $\alpha$-cut), that contain all the values of $\hat{A}$ in the interval $[a^-, a^+]$ with membership $\mu_{\hat{A}} \geq \alpha$. The elements $a^-$ and $a^+$ are the lower and upper bound of the $\alpha$-cut, respectively as shown in Figure 2-3.
To create fuzzy set membership functions for the MLR model, continuous data for each parameter, Q, T and DO, was discretised into daily-data arrays. Due to the frequency of sampling, this resulted in 24 flow samples per day for flow (i.e. hourly samples were grouped for each day) and on average 96 temperature and DO samples per day (corresponding to a sampling interval of 15 minutes). The sampling period, from May 22 – November 22 for 2006 to 2008 resulted in a total of 185 daily arrays for each variable per year.
Parameter uncertainty was then introduced into each data point. The flow was divided into three regimes: low (0 – 80 m$^3$/s), medium (80 – 220 m$^3$/s) and high (>220 m$^3$/s) flows as investigated by He et al. (2011). The error in each data point was selected as ±2% for the low flow regime, ±3% for the medium flow regime and ±8% associated with the high flow regime (Shiklomanov et al., 2006; Di Baldassarre & Montanari, 2009; McMillan et al., 2010). The error for each data point was calculated on a point-by-point basis, thus expanding the daily data array from 24 to 72 points. For T and DO, the accuracy rating for the YSI sonde was used (YSI Environmental, 2012). A ±0.3°C error was assigned to each temperature measurement, expanding the 96 daily data points to 288 points. For DO, the error assigned to each point was ±2% of the observed value or 0.2 mg/L, whichever was greater. This also resulted in a total of 288 data points per day.

Three types of membership functions were developed for each series of daily data. Linear (triangular) membership functions (L) and Normal membership based functions (N) were developed for all three parameters. Additionally, a Gumbel (G) distribution membership function was developed for Q. Linear and Normal membership based functions have been used in the literature but generally, in a multivariate analysis, all data have the same form of membership function. The distribution and underlying uncertainty in temperature and flow data are inherently different, and their respective influences on DO also have physical differences. Higher temperatures simply result in lower DO due to the chemical and physical nature of dissolved oxygen in water. However, the relationship between Q and DO is much more complex. Low Q not only presents less storage for DO but influences macrophyte and periphyton population densities and distributions within the varying bathymetry of the river. This in turn causes a complex response in both macrophyte and periphyton populations, and therefore, a highly unpredictable response in DO. The DO that is influenced by flow is an aggregation of a variety of unknown processes and thus, DO response tends to be delayed with respect to Q, which is in contrast to temperature influences because the response of DO to T is more immediate. This suggests that the uncertainty in T and Q and DO are likely all inherently different and require different forms of the membership functions. Linear and Normal membership functions were chosen because of their occurrence in fuzzy set analysis. A Gumbel distribution was chosen because it is commonly used for population distributions similar
in qualities to flow data and in particular, in extreme conditions such as low flow. In addition, a variety of permutations of these three distributions are tested.

For the linear membership function, the daily median value was assigned a fuzzy membership level of 1 (i.e. \( \mu = 1 \)). The minimum and maximum daily values were used as the upper and lower bounds of the fuzzy number, and assigned a fuzzy membership level of \( \mu = 0 \). The intervals \([x^-, x^+]\) corresponding to \( \alpha \)-cuts at \( \mu = 0.75 \), \( \mu = 0.50 \) and \( \mu = 0.25 \) were linearly interpolated between \( \mu = 1 \) and \( \mu = 0 \) (for \( x^- \)) or \( \mu = 0 \) (for \( x^+ \)). Thus, each linear membership function for each parameter consisted of a set of nine data points, corresponding to membership values or intervals at \( \mu = 1, 0.75, 0.5, 0.2, \) and 0. The \( \alpha \)-cuts at these specific locations were taken to give a broad representation of the possible values of a variable at different membership levels. This ensures that the fuzzy number generated fully represents the variable at different possibility levels. Examples of linear membership functions for Q, T and DO are shown in Figure 2-4 for June 19, 2007.

For the Normal based membership functions, the mean value of the daily data (\( \bar{x} \)) was assigned a membership value of \( \mu = 1 \). The upper and lower bounds of the daily data fuzzy numbers were calculated as \([x^-, x^+] = \bar{x} \pm 3s\), where \( s \) was the standard deviation of each daily data array. A membership value of \( \mu = 0 \) was assigned to the calculated upper and lower bounds of the daily data array. The values corresponding to the remaining three \( \alpha \)-cuts were calculated using the probability-density function (pdf) \( f(x) \) for the Normal distribution:

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu_N)^2}{2\sigma_N^2}} \tag{2-1}
\]

where \( \mu_N \) is the population mean and \( \sigma_N \) is the population standard deviation.
First, $f_{\mu=1}(x = \bar{x})$, i.e. the value of the pdf corresponding to the mean value of the data is calculated. From this, the corresponding values of the pdf at $\mu = 0.75$, 0.5 and 0.25 were calculated as:

$$f_{\mu=0.75}(x) = 0.75f_{\mu=1}(x = \bar{x})$$  \hspace{1cm} \text{Equation 2-2} $$
$$f_{\mu=0.5}(x) = 0.50f_{\mu=1}(x = \bar{x})$$  \hspace{1cm} \text{Equation 2-3} $$
$$f_{\mu=0.25}(x) = 0.25f_{\mu=1}(x = \bar{x})$$  \hspace{1cm} \text{Equation 2-4} $$

These values were then used to solve for the corresponding intervals of $x_\alpha$, using Equation 2-1. The interval, $[x^-, x^+]_\alpha$ corresponding to each $\alpha$-cut was then calculated as $[x^-, x^+]_\alpha = \bar{x} \pm x_\alpha$. Similar to the linear case, each daily data array consisting of either 72 (for flow) or 288 (for temperature and DO) had been reduced to a nine element
membership function, with values or intervals corresponding to the five membership levels (see Figure 2-4). The same methodology as the Normal-based membership functions was applied to obtain the Gumbel-based. The pdf of the Gumbel distribution is given by:

\[ f(x|\mu, \beta) = \frac{1}{\beta} e^{-\left(\frac{x-\mu_G}{\beta}\right)} e^{-e^{-\left(\frac{x-\mu_G}{\beta}\right)}} \]  \hspace{1cm} \text{Equation 2-5}

where, \( \mu_G \) and \( \beta \) are the location and scale factors determined from the population mean and standard deviation by:

\[ \beta = s \sqrt{\frac{6}{\pi}} \]  \hspace{1cm} \text{Equation 2-6}

\[ \mu_G = \bar{x} - \gamma \beta \]  \hspace{1cm} \text{Equation 2-7}

where \( s \) is the population standard deviation and \( \gamma \) is the Euler-Mascheroni constant.

The sample mean was used to calculate \( \mu_G \) and this value was assigned a membership level of \( \mu = 1 \). The upper and lower bounds of the daily data fuzzy numbers were calculated as \([x^-, x^+] = [\bar{x} - 2s, \bar{x} + 5s]\). A membership value of \( \mu = 0 \) was assigned to the calculated upper and lower bounds of the daily data array. The values corresponding to the remaining \( \alpha \)-cuts were calculated using Equation 2-5. From this, \( f_{\mu=0.75}(x) \), \( f_{\mu=0.5}(x) \), and \( f_{\mu=0.25}(x) \) were calculated as described for the Normal case and then used to calculate the intervals \([x^-, x^+]_\alpha\) at the respective membership level. An example for the Gumbel-based membership function for Q is shown in Figure 2-4.

By creating the membership functions, the three parameters, flow, water temperature and DO can be represented as normal, convex fuzzy numbers: \( \hat{\dot{Q}}, \hat{T} \) and \( \hat{DO} \). Each daily data array is thus reduced to a single fuzzy number, defined by nine elements, corresponding to the membership levels listed above. The fuzzy extension principle can be used to transform the possibility distribution (i.e. the membership function) to a probability distribution (Zadeh, 1987; Dubois & Prade, 1991). Dubois et al. (2004), Zhang (2009) and Zhang & Achari (2010a) describe a method to relate the pdf of a variable to its fuzzy membership function. The basic premise behind the transformation, from possibility to probability or from probability to possibility is relating the fact that
the area under a pdf is 1 and the maximum membership value of a fuzzy number is also 1 (Zhang, 2009). Using the $\alpha$-cut approach this relationship can be represented as:

$$p(x) = \int_0^1 \frac{1}{|A_\alpha|} d\alpha$$

Equation 2-8

where $p(x)$ is the probability density function of the fuzzy variable, $\mu(x)_\alpha$ is the membership function of the fuzzy variable, and $|A_\alpha|$ is the width of the $\alpha$-cut at $\mu = \alpha$. The membership function $\mu(x)_\alpha$ is upper semi-continuous and unimodal over the interval $[x^-, x^+]_\alpha$.

Using this approach, the uncertainty of a variable, represented as a fuzzy number can be transformed into a probabilistic representation of uncertainty. This is a useful method, since the spread of a variable $\hat{A}$ and its corresponding membership function, can be used to transform it to a crisp variable with an associated variance, which is represented in the pdf of the transformed variable. Some common transformations from probabilistic to possibilistic representation are presented in Duch (2005).

2.2.3 Multiple linear regression

Following the results of He et al. (2011), stepwise MLR was used to predict daily DO concentration, using only T and Q as candidates for the regressors. He et al. (2011) showed that these two abiotic factors were sufficient in predicting DO concentration in the Bow River; other abiotic factors (e.g. pH and radiation) were less significant. The algorithm was implemented and solved using the statistics toolbox in MATLAB (version R2011b). T, Q and DO data for 370 days from 2006 and 2007 was used to construct the MLR model, and 185 data points for each parameter from 2008 were used for validation of the model. By using the fuzzy data (described in the previous section), the structure of the MLR model was:

$$\hat{DO} = a + b\hat{T} + c\hat{Q}$$

Equation 2-9

where $\hat{DO}$, $\hat{T}$ and $\hat{Q}$ are the fuzzified form of the observed data and $a$, $b$ and $c$ are vectors of the regression parameters. It should be noted here that for the regression analysis, $\hat{Q}$ is the logarithm of the observed flow.
For incorporating the fuzzy nature of the variables, a regression analysis was conducted on each \( \alpha \)-cut. For example, values for \( a, b \) and \( c \) were calculated using inputs of \( \hat{DO} \), \( \hat{T} \) and \( \hat{Q} \) at \( \mu = 1 \), followed by another set of values for \( a, b \) and \( c \) using \( \hat{DO} \), \( \hat{T} \) and \( \hat{Q} \) at \( \mu = 0.75 \) and so on. These series of calculated regression parameters were then grouped as a vector. Fuzzy arithmetic (Klir, 1997; Dubois & Prade, 1991) was used to ensure that the calculated values of \( a, b \) and \( c \) at each \( \alpha \)-cut lead to predicted values of \( \hat{DO} \) that were normal and convex. For example, the value of \( a \) at \( \mu = 1 \) was always between the calculated interval of \([a', a^*]\) at \( \mu = 0.75 \), the interval of \([a', a^*]\) at \( \mu = 0.75 \) was always between the calculated interval of \([a', a^*]\) at \( \mu = 0.5 \) and so on.

### 2.3 Chapter results and discussion

Five MLR models were constructed using different combinations of fuzzy inputs. The first combination (Model 1) used a linear membership function for all three parameters. The second used a Normal-based membership function for all three parameters (Model 2). Model 3 used a Gumbel-based membership function for \( Q \) and linear functions for \( T \) and \( DO \). Model 4 again used a Gumbel-based function for \( Q \), but Normal-based functions for both \( T \) and \( DO \). Lastly, Model 5 used a Normal-based function for \( Q \) and linear functions for both \( T \) and \( DO \).

The models were evaluated using the coefficient of determination \( (R^2) \), root mean square error (RMSE) and mean absolute error (MAE). These error statistics were calculated using the following equations:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{Equation 2-10}
\]

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2} \quad \text{Equation 2-11}
\]

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \hat{x}_i| \quad \text{Equation 2-12}
\]
It should be noted that these statistics were calculated for each \( \alpha \) interval. For example, the interval of predicted \( \hat{D}0 \), \([DO^-, DO^+]\) at \( \mu = 0.75 \) was compared with the observed interval of \( \hat{D}0 \) at \( \mu = 0.75 \). Thus, an array of error statistics was developed for each of the five MLR models. However, the most important statistics were those corresponding to \( \mu = 1 \) as these values carry the highest possibility of occurrence; and are thus more sensitive to values that have a membership value closer to unity and a lower sensitivity to values that have a membership value closer to zero (Shrestha & Nestmann, 2009).

Figure 2-5 to Figure 2-9 show the results of the five model scenarios, including a comparison of the observed and modelled results for the 2006 and 2007 data sets and for the observed and modelled results for the validation data set (from 2008). In general, these figures show that the MLR approach which incorporates uncertainty and variability in the input data is better at predicting uncertainty and variability in the output data (DO) as compared to the non-fuzzy approach. This is demonstrated by the fact that each model not only captures the general DO trend that is visible in the observed data, but it also captures the upper and lower bounds of the observed variation, which are enclosed in the interval between the various \( \alpha \)-cut results that are displayed. Comparing the modelled variability of DO between each model, Models 2 and 4 have a higher variability in both the upper (i.e., \( \mu = [x^+]_\alpha \)) and lower bounds (i.e., \( \mu = [x^-]_\alpha \)) in the predicted data, whereas Models 1, 3 and 5 show a tendency of calculating larger variation in the upper bounds compared to the lower bounds. These three models all use linear membership functions for T and DO, indicating that this type of membership function may not be suitable for predicting low DO using MLR and fuzzy numbers. Details on the amount of variability captured and the relative performance of each model is discussed below.
Figure 2-5 Predicted values from Model 1 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$
Figure 2-6 Predicted values from Model 2 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$
Figure 2-7 Predicted values from Model 3 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$
Figure 2-8 Predicted values from Model 4 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$.
Figure 2-9 Predicted values from Model 5 for 2006 – 08; note the predicted fuzzy intervals from $\mu = [0, 1]$ are plotted against the observed values for $\mu = 1$
Figure 2-10 shows the observed and predicted values for DO concentration for both the 2006-07 data and the 2008 (validation) data against the 1:1 line. The figure illustrates that Models 1, 3 and 5 have a smaller variance in their predicted data compared to Models 2 and 4, particularly in the low DO range. Models 2 and 4 predict DO to be less than 5 mg/L in both the 2006-07 and 2008 data sets, although at low possibility values (i.e., $\mu \leq 0.25$ mg/L).

Figure 2-10 Observed versus Predicted DO concentration for all five models for the 2006-07 and 2008 data sets; clockwise from top left: Model 1, Model 2, Model 4, Model 5, and Model 3

Figure 2-11 and Figure 2-12 summarise the error statistics for each model; Figure 2-11 shows the calculated error values using the 2006 and 2007 data set, while Figure 2-12 uses the predicted values to calculate the error statistics for the 2008 data set. For the 2006 and 2007 data, the coefficient of determination, $R^2$ is higher for the upper bound ($\mu^+$) values of DO and Model 2 and 4 show a better fit at $\mu = 0.5^+$, 0.75$^+$ and 1, compared
to Models 1, 3 and 5. However, the same is not true for $R^2$ for the lower bound values, where Models 2 and 4 are lower than the other models. In general, the upper bound values have a better fit compared to the lower bound values. This highlights the fact that the shape of the membership function required to model DO concentration in this environment, may need to be skewed to the upper bound values, as in the Gumbel distribution used in Model 4. A similar trend in $R^2$ is noted for the predicted (i.e., 2008 DO concentration) values in Figure 2-12.

Figure 2-11 $R^2$, RMSE and MAE for 2006-07 data set

MAE and RMSE values were small for both data sets, averaging 10% of the observed values. For both statistics, the magnitude of error was higher for smaller values of $\mu^+$ and this error decreased as $\mu^+$ increased. In other words, the MLR model produced higher errors when predicting low-possibility upper bound values of daily DO concentration. In general RMSE was greater than MAE, while the difference between the two statistics was not large and indicated that overall, the MLR model predicted DO within a narrow window. RMSE and MAE for Models 2 and 4 were distinctly higher than the other three models for $\mu^+$ in the both data sets and this error increased with lower $\mu$ values. However, the magnitudes of errors for both the upper and lower bounds (i.e. $[x^-, x^+]\alpha$) were not large, and in general were lower for high membership levels (or at high possibility values).
Figure 2-12 $R^2$, RMSE and MAE for validation (2008) data set

Figure 2-13 shows the results of the regression parameters that were calculated using the 2006 and 2007 data set for each model at each membership level. The parameters at each membership level show that Models 2 and 4 both have parameter values that are markedly different than the other three models. Each parameter, $a$, $b$ and $c$, for these models has a larger interval compared to the regression parameters for the other three models. Additionally, the interval increases at lower possibility values. This highlights the fact that using non-linear membership functions in MLR analysis, produces coefficients that, along with the model, can better describe the variability in the observed parameter, DO.

Figure 2-13 Values for regression parameters, $a$, $b$ and $c$ for each of the five models
These error statistics cannot be used individually to determine whether or not the proposed methodology to propagate uncertainty through a data-driven model using fuzzy numbers is appropriate (He et al., 2011). The statistics need to be looked at collectively to produce a clearer picture. The statistics, in general, indicate that MLR is a suitable methodology to predict DO concentration in this environment, as the magnitude of errors produced is low, and the coefficient of determination is suitable ($R^2 \approx 0.65$ at $\mu = 1$ for Models 2 and 4). Together, the statistics show that representing uncertainty as a fuzzy number is appropriate. This is because the predicted fuzzy DO generated from the MLR models are comparable to the fuzzy DO constructed from the observed data. What this study highlights is that different combinations of fuzzy number representation (linear, Normal or Gumbel based) produce noticeably different results. The most pronounced case is when a non-linear, specifically a Normal-based membership function, is used to represent the independent regressors (i.e. T and Q). In this case the variability in predicted data is larger than the linear case. To assess whether or not this is a more accurate method of predicting uncertainty in DO, a new approach to measure variance is developed and tested. At each modelled $\alpha$-cut, the percentage of observed data falling within the upper and lower bounds of that particular $\alpha$-cut was determined for each model and compared. More weight was assigned to higher possibility values as a measure of accuracy. Figure 2-14 shows the results of this analysis, for both the 2006-07 data and the 2008 data.
Both sets of data indicate that Models 2 and 4 are able to capture the variability in the observed data at larger possibility levels. For example (using the 2006-07 data), Model 2 can capture 80% of the observed data within the first $\alpha$-cut (i.e. $0.75<\mu<1$) and 90% of the data at the next cut ($0.5<\mu<0.75$). Compared to this, Model 1 captured 55% of the data at the first $\alpha$-cut, 80% at the second and 90% at the third ($0.25<\mu<0.5$). This suggests that the Model 2 is a more robust predictor of the overall variance in DO compared to Model 1, as it can capture the 90% of the observed data with a possibility value between $\mu = 0.5$ and above, whereas Model 2 can capture 90% of the data at with a lower possibility, $\mu = 0.25$ and above. This trend is seen throughout the data set for both Models 2 and 4, which can predict a larger amount of data, with a higher possibility. This confirms that Models 2 and 4 use a better representation of fuzzy-based uncertainty compared to the other models. The differences between Model 2 (Q as a Gumbel-based fuzzy number) and Model 4 (Q as a Normal-based fuzzy number) are marginal and the better representation between these two models is not immediately evident. However, the different combinations of membership function types indicate that a non-linear representation of Q alone is not sufficient to reproduce DO variability, since the Gumbel-based Model 3 does not predict the variability as well as the other two non-linear models.
Similarly, non-linear membership functions are required for both T and DO to capture the variability in DO (Models 2 and 4) as none of the linear membership combinations of DO provided a good reproduction of DO variability. This, however, does not preclude that Normal-based functions are the best representation for T and DO, nor that the Normal or Gumbel-based functions are more appropriate for Q. It does show that when these particular combinations are used, the results are more accurate compared to the traditional linear representation of fuzziness.

2.3.1 Risk assessment

One of the overall objectives of creating fuzzy number representations of the input and output parameters was to be able to predict the probability of the DO concentration to be less than a threshold value, in this case 5 mg/L - the Alberta guideline for minimum daily DO concentration. To do this, the possibility of predicted DO to be less than 5 mg/L was converted to a probability (i.e., $P(\text{DO}<5)$) using the methodology outlined in Section 2.2.2. This concept is illustrated with an example, using the results of the two best performing models, Model 2 and Model 4.

First, for each model, specific days were identified when there was a possibility of predicted DO (for the 2008 data set) falling below the 5 mg/L limit. This categorization was conducted by looking at values of DO at each membership level ($\mu = 0, 0.25, 0.50, 0.75$ and 1) for the 185 days of available data from both model outputs. It was found that occurrences of predicted DO that were below 5 mg/L in 2008 were isolated at membership values of $\mu = 0.25^-$ and $\mu = 0^-$. This is illustrated in Figure 2-15 which shows the predicted DO concentration at $\mu = 0.25^-$ and at $\mu = 0^-$, along with the DO concentration at $\mu = 1$ for reference, for both Model 2 and 4. The days where there is possibility of DO to be less than 5 mg/L are highlighted by markers at the different membership levels. Model 2 predicted that there was a possibility of DO to be less than 5 mg/L for 97 of the 185 days at $\mu = 0^-$ and 11 out of 185 days at $\mu = 0.25^-$. Model 4 predictions were slightly different for the $\mu = 0^-$ case only; the model predicted 98 days where there was a possibility of DO to be less than the limit. The 11 days identified by both models occur in late July 2008, roughly corresponding to the period when the lowest annual observed DO in the river occurs.
Figure 2-15 Identifying the days in 2008 when the predicted DO concentration had a possibility to fall below the 5 mg/L limit (top: Model 2; bottom Model 4)
Since the fuzzy membership function for DO for the input in both these models was based on a Normal function, the predicted values of DO also followed a Normal-based membership function. Thus, the possibility of predicted DO to be less than the limit was transformed to a probability using a Normal-based cumulative distribution function. The probability, \( P(\text{DO} < 5) \) (i.e. theoretically in the interval \((-\infty, 5]\), but restricted to between \([0, 5]\) in this model since DO concentration cannot be less than 0 mg/L) was calculated at both membership levels. The calculated probabilities are displayed in Figure 2-16. The probability of DO to be less than the limit at \( \mu = 0 \) is very low (0.14%); this corresponds to the probability of a Normal distribution at the location \([x^-] = \bar{x} - 3\sigma\) since this location was assigned a membership value of 0 to develop the membership function.

**Figure 2-16** Predicted probability of DO concentration to be less than the Alberta guideline of 5 mg/L using Models 2 and 4
The probability $P(\text{DO} < 5)$, however, was significantly more noteworthy at the $\mu = 0.25$ level. For both models, the probability varied between 3.9% and 4.9%. Though, this probability may seem low, it roughly corresponds to a 1:20 or 1:25 year flood event. Return periods of this magnitude are frequently used in flood-prevention design guidelines for structures such as local roadways (i.e., smaller than main or major) in Alberta (Watt, 1989). Thus, though the probability of occurrence, and consequently the risk, appears to be low, it is in fact on par with commonly used risk-prevention guidelines in hydrology. It is also interesting to note that the actual (non-fuzzy) observations of DO from the corresponding dates average between 6.5 and 11 mg/L. Thus, if only mean values are used for model construction and prediction, the ability to pinpoint the days where there is chance of DO to be less than the limit would not be possible. This fuzzy approach can identify the risk at any day for the DO concentration to be lower than any predetermined value. Another way of looking at the predicted probability is that, in models that do not incorporate uncertainty and variability, the predicted values over this 11 day period would not be out of the ordinary, with respect to the rest of the time series. However, both Models 2 and 4, show distinctly lower DO concentration over the same period.

2.4 Chapter summary

DO is an important water quality parameter for assessing aquatic ecosystem health. Studies have shown that DO concentration in a riverine environment can be adequately predicted using a MLR model with abiotic factors (flow and water temperature) as the regressors. However, the uncertainty and variability in the input data are important and need to be identified and propagated through the model to successfully quantify DO and the risk of the occurrence of low DO. Limited research has been conducted that incorporates fuzzy set uncertainty analysis for data-driven DO prediction. It is important to represent the predicted DO values as a probabilistic set rather than a deterministic value, to mirror the variability in observed daily DO in riverine environments.

In this chapter, a fuzzy set based methodology is developed to represent uncertainty and variability in the regressors, Q and T. Typically linear (or triangular) membership functions are used to define fuzzy sets in uncertainty analysis. However, non-linear functions, such a Normal or Gumbel-based functions may be more appropriate and are
proposed for this application. Thus, three types of membership functions were created: linear, Normal and Gumbel-based functions. From this, five different MLR models were developed and tested using different combinations of these membership functions. This methodology was tested on data from the Bow River in Calgary, Alberta where low DO has been observed occasionally. Because of this the City of Calgary is interested in developing tools that can help predict DO concentration, in the absence of a physically-based model.

The results show that using non-linear membership functions are superior to the traditional linear membership function approach. The two models that used Normal or a mixture of Normal and Gumbel based membership functions performed better in predicting DO and its variability than their linear counterparts. This was based on a series of error statistics as well as a measure of the ability of the model to capture the observed variability.

In addition to this, a comparison of the results from a fuzzy set based to a traditional application of the MLR model show that the former application was more suitable as it provided a range of values of predicted DO, whereas the traditional approach only provided a single deterministic value. This issue was further highlighted in the risk analysis component of the study. A possibility-probability transformation was used to calculate the risk of DO concentration in the Bow River to be less than the provincial guideline of 5 mg/L. The two models used (i.e. the non-linear membership function models) predicted that for the period corresponding to the lowest measured DO in the Bow River, the chance of DO falling below the guidelines was between 3.9 and 4.9%. In doing so, the models highlight the advantages of using a fuzzy set methodology to represent uncertainty, as it was able to predict the possibility and probability of DO to be less than the limit, whereas a conventional data-driven modelling approach could not have captured this information.
3. Two-dimensional prediction and uncertainty analysis of dissolved oxygen in the Bow River using fuzzy linear regression

3.1 Chapter introduction

Dissolved oxygen (DO) is an important water quality parameter; it is often used to assess the overall health of an aquatic ecosystem. Generally, higher concentration of DO is favourable, as it is vital for the survival of organisms. In addition to this, the amount of variation of DO per day in a system is also an important indicator of ecosystem health, where lower variability is preferred (He et al., 2011). Low DO concentrations can result in hypoxic and anoxic conditions, affecting fish and other aerobic organisms’ populations (Lehman et al., 2004). Thus, accurately predicting the DO concentration in aquatic environments is important, so that steps to mitigate these adverse effects may be taken.

Characterising DO can be very complex: a number of primary and secondary drivers are known to affect its concentration in riverine environments. These drivers are a function of the river system and thus vary from site to site. Primary drivers, that directly influence DO, include the re-aeration rate, amount of imported DO (i.e. upstream contributions), photosynthesis, biochemical oxygen demand, sediment oxygen demand and the residence time. Examples of secondary drivers, that indirectly affect DO concentration, include water temperature, flow velocity, channel geometry, algal biomass, turbidity and nutrients (Lehman et al., 2004; Jassby & Nieuwenhuyse, 2005). These drivers can be grouped in to biotic (e.g. photosynthesis and respiration) and abiotic (non-living, chemical and physical) factors, which can influence DO concentration in aquatic ecosystems (Pogue & Anderson, 1995; Hauer & Hill, 2007). Often, sufficient biotic data is not available to be able to characterise the complex physical system that controls DO concentration. Recent research has shown that abiotic factors can accurately predict DO concentration in riverine environments. He et al. (2011) showed that daily DO could be predicted using abiotic factors such as flow, water temperature, solar

---

radiation, turbidity and nutrient concentration. The advantage of using abiotic factors to characterise DO is that if the important factors can be identified, there is potential to control and improve DO in aquatic systems (Khan et al., 2013).

Another difficulty in DO prediction is that the physical system that affect DO concentration, either primary or secondary drivers, or abiotic or biotic, are quite complex and hard to represent mathematically. A major drawback of using physically-based models is the data needed for simulations, which may not be easily available, as well difficulties in model calibration. This is especially true when a number of different processes (e.g., re-aeration rate, photosynthesis, biochemical oxygen demand) are grouped together. Similarly, data may be available for one region but may not be applicable for another. On the other hand, data-driven models, which are based on the analysis of the data characterising a system, require little or no assumptions regarding the nature of the physical system being modelled (Solomatine & Ostfeld, 2008; Khan et al., 2013). Data-driven models are defined on the basis of the connections between concurrent input and output variables (Solomatine & Ostfeld, 2008). The advantage of a data-driven approach is that if strong relationships between input and output parameters are found, a simple yet robust predictive model can be constructed.

He et al. (2011) and Khan et al. (2013) showed that data-driven techniques were able to adequately predict DO in riverine environments. He et al. (2011) did not consider the impacts of variability and uncertainty in the underlying data on the predictions. However, data-driven modelling has inherent uncertainties associated with it, and the proper identification and propagation of it is essential for understanding model prediction (Shrestha & Nestmann, 2009; Li et al., 2009). Uncertainty can exist due to many reasons, e.g. scarcity in data, measurement errors, variability over time, and combing and propagating data from different sources (Zhang et al., 2009; El-Baroudy & Simonovic, 2006; Shrestha & Nestmann, 2009). Fuzzy numbers (derived from fuzzy set theory; Zadeh, 1965) is an effective technique of representing and aggregating uncertainty in data-driven models (Khan et al., 2013).

A number of studies have shown that fuzzy numbers are a viable method of representing uncertainty for predicting environmental factors (such as dissolved oxygen) and contaminant transport, and have been used in conjunction with data-driven models
( Bárdossy et al., 1990; Mujumdar & Sasikumar 2002; Altunkaynak et al., 2005; Shrestha et al., 2007; Shrestha & Nestmann 2009; Zhang et al., 2009; Wang et al., 2012). For brevity, the details of fuzzy sets and numbers are not included here and the reader is directed to Khan et al. (2013) for more information.

### 3.1.1 Chapter objectives

Low DO concentration has been observed in a highly urbanised reach of the Bow River, in Calgary, Alberta, Canada. The City of Calgary has made efforts to reduce sediment and nutrient loadings into the river in an effort to maintain DO concentration above 5 mg/L (the provincial 1-day minimum DO guideline; Alberta Environment Protection, 1997). As part of its compliance programme, the City of Calgary has made efforts to predict the DO concentration in the River.

The objective of this research is to use a data-driven technique (specifically multiple linear regression or MLR) to predict DO concentration at different sites along the Bow River within the Calgary city limits. This research expands previous work by He et al. (2011) and Khan et al. (2013), by extending DO prediction from a point location to a 2-dimensional (2D) cross-section representation at seven sites. Abiotic factors (water temperature and flow) are used as the independent variables to predict daily DO concentration. Further, fuzzy numbers are implemented to represent parameter uncertainty in the independent and dependent variables. The resultant fuzzy predictions are used to evaluate the risk of low DO occurring under different scenarios.

### 3.2 Methods

#### 3.2.1 Site description and data collection

The Bow River originates in the Rocky Mountains in Alberta and flows south-eastwardly through the Foothills before entering Calgary (see Figure 3-1). The average annual discharge is approximately 90 m$^3$/s and measures 100 m wide and 1.5 m deep. A number of tributaries flow into the Bow River within the City limits; these include Nose Creek, the Elbow River, Fish Creek and Pine Creek. In addition to this, three wastewater

---

4 The MATLAB code associated with this section is included in Appendix B.
treatment plants (WWTPs) discharge into the River: Bonnybrook WWTP, Fish Creek WWTP and Pine Creek WWTP.

For this research, seven sites along the Bow River within the city limits were selected as locations for cross-sectional DO analysis and prediction. The selection of these sites was based on two factors, firstly, the availability of suitable data, and secondly a uniform representation of the Bow River in Calgary. These sites, illustrated on Figure 3-1, are (1) Bearspaw Dam (BS), (2) Pumphouse (PH), (3) Cushing Bridge (CS), (4) upstream of Bonnybrook WWTP (USBB), (5) downstream of Bonnybrook WWTP (DSBB), (6) upstream of Fish Creek WWTP (USFC), and (7) downstream of Fish Creek WWTP.

Figure 3-1 An aerial view of the City of Calgary showing (1) Bearspaw Dam, (2) Pumphouse, (3) Cushing Bridge, (4) upstream of Bonnybrook WWTP, (5) downstream of Bonnybrook WWTP, (6) upstream of Fish Creek WWTP, and (7) downstream of Fish Creek WWTP
3.2.1.1 Bathymetry data
To extend DO concentration prediction from a point to 2D cross-section, bathymetry data for the seven sites listed above was collected. This data was collected by Golder Associated Ltd in partnership with Alberta Environment and the City of Calgary as part of the “Bow River BioSonics Pilot Survey with Water Quality Ground-truth Monitoring” study (Wang & Berlando 2011). The horizontal and vertical coordinates of the profile of the cross-section were collected for each location. Water surface elevations and the corresponding average flow velocity, flow and flow area were also collected. Data from 19 measurements was selected. These measurements were taken between April and October 2010, with flow measurements ranging between 50 and 230 m³/s; a sample of a cross-section is shown in Figure 3-2. Details of equipment, data measurement and collection is included in Wang & Berlando (2011).

Figure 3-2 Examples of a cross-section profiles and water surface elevation at (a) the Bearspaw, (b) Cushing Bridge and (c) Upstream of Fish Creek WWTP sites for July 2007.

3.2.1.2 Water temperature and DO data
Water temperature (T) and DO data was provided by the City of Calgary. The City routinely collects several water quality samples annually from a number of locations along the Bow River and its tributaries. For this research approximately monthly samples for the period May to November for the years 2007 to 2011 were selected for each site.
This data is illustrated in Figure 3-3 below; note that the horizontal axis represents each monthly sample in chronological order (e.g. ‘1’ is May 2007, ‘35’ is November 2011).

Figure 3-3 (a) Water temperature and (b) dissolved oxygen data collected by the City of Calgary
3.2.1.3 Flow data

Flow data, Q, for the 35 dates corresponding to the date of the water quality samples, were collected from the Water Survey of Canada Archived Hydrometric Data for the location Bow River at Calgary 05BH004 (WSC, 2013). The data described in the previous section were then used to create stage-discharge relationships using the 19 water surface elevation and flow data points. This was used to estimate the corresponding water surface elevation and average velocity, at each location, for each of the 35 Q collected from the Water Survey of Canada.

It was assumed that the velocity profile decreased linearly from a maximum value at the mid-point on the surface to zero at the bottom and edges of the channel. The cross-section was divided into roughly 100 by 100 nodes. The linear velocity profile was then used to calculate velocity at each node. The corresponding area of each node was then used to estimate the flow at each node, q. The sum of all q over a cross-section equalled to the measured Q. Examples of velocity profiles for three sites is plotted below in Figure 3-4.
Figure 3-4 Velocity profiles at three locations (a) BS, (b) CS and (c) USFC for July 2007
3.2.1.4 Real-time water quality and flow data
Real-time T and DO (every 15 to 30 minutes), and Q (every hour) data were collected between 2006 and 2008 by the University of Calgary. Details on this data set can be found in He et al. (2011) and Khan et al. (2013). This data was used to characterise daily variation in each parameter: the mean daily value (μ) and standard deviation (σ) were calculated and used in the analysis described below.

3.2.2 Non-linear membership functions for fuzzy numbers
The collected T, Q and DO data described above were converted to fuzzy numbers to represent the uncertainty in the data. Fuzzy numbers are a way of representing uncertain information. By definition, they are convex and normal fuzzy sets. For this research, these fuzzy numbers were described using Gaussian-shaped membership functions, where μ was given a degree of membership (f(x=μ)) of 1, while f(x=μ ± 3σ) had a value of 0. The values of these fuzzy numbers were evaluated at 5 α-level intervals, where α = f(x) and equalled 0, 0.25, 0.50, 0.75 and 1. Further details on the choice of membership functions, the α-level cuts and construction of membership functions for this data set can be found in Chapter 2.

3.2.2.1 Flow data
For each of the 7 cross-sections and 35 Q, each q (the flow at a node) was fuzzified. The observed q was assigned a f(x_q) = 1 and f(x_q) = 0 was assigned μ_q ± 3σ_q. Here σ_q was calculated from the real-time daily flow data, weighted to the representative area of each node. This resulted in a 9 element interval (ranging from f(x_q) = 0 to 1) characterising the fuzzy q at each node. Examples of qf are illustrated in Figure 3-5 below.
Figure 3-5 Examples of $q_f$, the fuzzy version of $q$, at three locations (a) BS, (b) CS and (c) UPFC for July 2007
3.2.2.2 Water temperature and dissolved oxygen data

For T, it was assumed that the temperature did not vary across the cross-section of the river. Thus, T was identical at each of the nodes. T was fuzzified using $\mu_T$ and $\sigma_T$ calculated using the real-time data, similar to that of q; the observed T corresponded to $f(x_T) = 1$ and $\mu_T \pm 3\sigma_T$ corresponded to $f(x_T) = 0$. Similarly, DO was also fuzzified to represent the “observed” fuzzy form for comparison with the predicted DO from the MLR model.

3.2.3 Multiple linear regression

Following the results of He et al. (2011) and Khan et al. (2013), stepwise MLR was used to predict the daily DO at each site along the Bow River, using Q and T as the regressors. The algorithm was implemented in MATLAB (version 2012b). The structure of the MLR model, from Khan et al. (2013) was:

$$DO_f = a + bT_f + cQ_f$$  \hspace{1cm} \text{Equation 3-1}$$

where $DO_f$, $T_f$ and $Q_f$ are the fuzzy forms of DO, T and Q. a, b and c are fuzzy regression parameters. Here, both Q and $Q_f$ are the logarithms of the observed flow. For this research, Model 2 from Khan et al. (2013) was selected, for which, as mentioned earlier, each variable was converted to Gaussian-shaped fuzzy number. The real-time data was used for model construction and validation. Details on model development, performance and results are described in Khan et al. (2013). To adapt this MLR equation for 2D prediction, Equation 3-1 was adapted to include $q_f$ (the flow at each node) in place of $Q_f$:

$$DO_f = a + bT_f + dq_f$$  \hspace{1cm} \text{Equation 3-2}$$

where $q_f$ is the flow at each node, and d is the area weighted regression coefficient calculated as follows:

$$d = c^*a/A$$  \hspace{1cm} \text{Equation 3-3}$$

Equation 3-2 was used to calculate $DO_f$ at each node, at 9 membership levels, using fuzzy arithmetic. This was done for each of the 7 sites, for each of the 35 samples.
3.3 Chapter results and discussion

Figure 3-6 below shows an example of the comparison between predicted DO$_f$ and observed DO at each site, for samples taken on July 2007. The predicted DO$_f$ is shown for 2 membership levels, $f(x) = 1$ and 0.75 (both lower and upper bound of the interval). For this Figure 3-6, the DO$_f$ values over the cross-section were averaged for illustration purposes. The figure shows that the overall trend from the upstream site, BS, to the downstream site, DSFC, is not accurately predicted, as is visible when comparing DO$_f$ (at $f(x) = 1$) and the observed DO. However, it should be noted that the predicted DO$_f$ here is the mean of the entire cross-section, whereas the observed data, is from a point location.

In addition to this, the observed variability, falls within the variability exhibited by DO$_f$ within the first $\alpha$ cut ($f(x) = 0.75$).

A possibility-probability transformation (Khan et al., 2013) was conducted to calculate the probability of DO$_f$ to be a given value. The probability that the DO$_f$ ($P(\text{DO}_f)$) was between the lower bound $f(x) = 0.75$ and $f(x) = 1$ was approximately 24%. This means that, on average for each site, the probability of DO$_f$ has values between 0.5 and 9 mg/L (see Figure 3-6) is 24%. In addition, to this the $P(\text{DO}_f)$ when DO$_f$ is between 5 and 9 mg/L, when calculated to be 18%, and more importantly $P(\text{DO}_f)$ when DO$_f$ is between 0.5 and 5 mg/L to be 6%. This is a significant result as it indicates that the methodology can highlight when there is probability of low DO (i.e. below the Alberta guideline) even when the “most-possible” predicted value is above the guideline. An ordinary MLR model would have only predicted DO to be 9 mg/L, whereas this methodology can calculate the probability of DO for any interval. Thus the fuzzy number approach, with its lower possibility levels, is able to quantify the probability and the corresponding magnitude of the variability of the predicted DO.

Another important result of this research is that the variability within the reach (upstream to downstream) is not as pronounced as the variability within a single cross-section. This shows that bathymetry of river plays an important role in DO concentration in riverine environments. This result also highlights the importance of including uncertainty, specifically through fuzzy numbers.

However, when looking at the inter-cross-section variability of DO at one $\alpha$ level at a time, the variability is quite low. This is illustrated in the surface plots below in Figure
3-7, where the DO distribution for July 2007 is plotted for four sites for f(x) = 1 (similar trends are seen at other membership values). The reason for this is that water temperature T has a higher contribution to DO compared to the node flow q, this is apparent by the relative magnitudes of each (e.g. 10 °C and 0.03 m$^3$/s) and their respective regression coefficients in the MLR equation ($a = 9.98$, $b = -0.1867$ and $d = 2 \times 10^{-4}$; at f(x) = 1). Since Q was converted to q and varied across the cross-section (decreasing with depth as function of velocity), it meant that T, which was constant, kept the value of DO relatively constant as well.

![Graph showing comparison of predicted and observed DO for July 2007](image)

**Figure 3-6** Comparison of predicted DO at two $\alpha$ levels and observed DO
Figure 3-7 Illustration of the low inter-cross-section variation of DO at each site, with a degree of membership of 1, for the July 2007 data.
3.4 Chapter summary

In this chapter, a new method to predict DO in a river using a multiple linear regression model, with flow and water temperature as regressors was outlined. The model used fuzzy numbers instead of crisp numbers to represent the input and output data. The use of fuzzy numbers allows the uncertainty, from a number of different sources, to be combined and propagated (via fuzzy arithmetic) through the model. The model was applied to a river in Calgary, Alberta, Canada and predicted DO concentration in 2-dimensions (i.e. over the cross-section) at seven sites, for a collection of daily data between May and November, 2007 to 2011.

The results of the model show that fuzzy numbers are a suitable methodology of representing uncertainty for use in data-driven techniques. On average, the model predicted large variability in predicted DO at each site, and provided the probability (through a possibility-probability transformation) of the occurrence low DO. For each site, the probability of DO to be less than the Alberta Guideline (5 mg/L) was approximately 6%.
4. Comparing the proposed and existing fuzzy linear regression methods

4.1 Chapter introduction

Dissolved oxygen (DO) is an important water quality parameter and is often used as an indicator of overall aquatic ecosystem health. DO concentration in waterbodies influences biotic communities and is vital for their survival (He et al., 2011). Due to these reasons, DO is a parameter of concern and is routinely monitored by water resource managers. A number of factors can cause low DO and thereby increasing the risk of adverse effects to the aquatic environment. Both biotic and abiotic (non-living, physical and chemical attributes) factors influence DO concentration in riverine environments. These affects can be direct or indirect. For example, macrophyte growth can directly cause variation in DO concentration through photosynthesis and respiration (Pogue & Anderson, 1995; Hauer & Hill, 2007). In contrast, flow rate influences macrophyte population, which in turn affects DO concentration.

The physical processes that govern the behaviour of DO in the aquatic environment are quite complex and poorly understood. In addition, modelling DO behaviour requires the parameterization of a large number of variables (Khan et al., 2013). Thus, predicting DO concentration in riverine environments is not only challenging but filled with uncertainty that is generally unknown. This increases the difficulty in decision-making for water resource managers and also increases the risks to aquatic ecosystems. Thus, a method is needed that can accurately predict DO concentration that is not limited by the lack of understanding of governing processes and at the same time, circumvents the parameterization issues and data hurdles that plague modelling. In addition, the method must capture the uncertainty and variability in DO prediction so that complete and useful information can be provided to managers. In this chapter, a new method to predict DO concentration that satisfies these two requirements is proposed. It uses a combination of abiotic factors, data-driven techniques and fuzzy set theory.

In a highly urbanised reach of the Bow River, in Calgary, Alberta, Canada, occasional low DO concentrations have been observed. A number of abiotic factors (influenced by

---

urbanization) seem to affect DO concentration and variability at sites downstream of a wastewater treatment plant’s effluent discharge (He et al., 2011; Khan et al., 2013). The City of Calgary is mandated to maintain DO concentration above the Alberta provincial guideline of 5 mg/L (one day minimum limit) (Alberta Environment Protection, 1997). Thus, monitoring and modelling DO concentration is part of the City of Calgary’s compliance program. Recent studies by He et al., (2011) and Khan et al., (2013) have used data-driven modelling techniques (e.g. multiple linear regression and multiple layer perceptron neural networks) with abiotic factors to predict DO concentration in the Bow River in the absence of both, sufficient biotic data, and knowledge of the physical processes underlying DO trends in the river. In these studies, abiotic factors such as water temperature, river discharge, solar radiation, and pH, were analysed to see if they can be used to predict daily DO concentration and variability. The rationale of identifying which, and to what extent, these abiotic factors influence DO concentration, is that if the former can be controlled, the latter can be indirectly improved. The results from these studies show that a data-driven approach is an appropriate method to predict DO concentration in the River.

Another implication of the success of these studies is that the developed models can be used to predict when (and possibly where) low DO concentration may occur in the River. This extends the application of these models from simple predictions to risk assessment. For example, if forecast data suggest that the abiotic factors are expected to be such that the data-driven models predict potentially low DO concentration, a probability and risk of “DO being below 5 mg/L” can be assigned. However, data-driven modelling has intrinsic uncertainties associated with it. This uncertainty must be identified and propagated through the model before important decisions based on the output can be made (Shrestha & Nestmann, 2009). Scarcity, measurement errors, randomness and scaling issues are a few examples of the different sources of uncertainty, each of which can affect model projections (Zhang et al., 2009). In addition, uncertainty may also increase when data from numerous sources are used, integrated and propagated (Porter et al., 2000; Shrestha & Nestmann, 2009; Hermann 2011). Fuzzy numbers and its extension, possibility theory, are one approach for dealing with these types of uncertainty.
4.1.1 Uncertainty analysis using fuzzy numbers

Fuzzy set theory was first proposed by Zadeh (1965) and can be described as a generalization of classical set theory. In classical set theory, an element $x$ either belongs or does not belong to a set $A$. In contrast, using fuzzy set theory, the elements $x$ of a fuzzy set $\tilde{A}$ have a degree of membership within that set. This degree of membership is defined as a value between 0 and 1. Here a membership level, $\mu$, equal to 0 means that $x$ does not belong in $\tilde{A}$, a value $\mu = 1$ means that it completely belongs in set $\tilde{A}$, while a value $\mu = 0.5$ means that it is only a partial member of $\tilde{A}$. By these definitions, a traditional set (known as a crisp set to distinguish it from a fuzzy set) is a special case of a fuzzy set: when $x$ is a member of $A$, its membership level $\mu$ is equal to 1, and when $x$ does not belong to $A$, $\mu$ is equal to 0. The collection of membership levels that define a fuzzy number is known as its membership function. Fuzzy sets have recently been used in a variety of hydrological and environmental applications (Wang & Huang, 2012; Han et al., 2013; Wang & Huang, 2013a; 2013b).

Fuzzy set theory was developed to describe uncertain or imprecise information. A fuzzy number is an extension of fuzzy set theory, and expresses an uncertain or imprecise quantity. In much the same way that a fuzzy set generalises a crisp set, a fuzzy number generalises a real number. A fuzzy number does not refer to a single value, but represents the set of all possible values that define it. The nature of fuzzy numbers lends itself well to uncertainty analysis and is one technique for characterising and aggregating uncertain data (Khan et al., 2013). It is particularly useful for dealing with uncertainties when data is limited or imprecise (Bárdossy et al., 1990; Guyonnet et al., 2003; Zhang & Achari, 2010a; Huang et al., 2010). In these cases, a probabilistic representation of parameters may not be possible since the exact values of parameters may be unknown or only partial information is available (Zhang, 2009). In such a situation, fuzzy techniques can be used to transform uncertainty from traditional probabilistic to possibilistic representation (Xia et al., 2000).

Fuzzy numbers, along with possibility theory, has been widely used in hydrology to represent uncertainty in the parameters of numerical models. A popular application has been to represent water level and flow rate as fuzzy numbers to improve stage-discharge relationships (Alvisi et al., 2006; Alvisi & Franchini, 2011; Shrestha et al., 2007).
Another approach has been to model contaminant transport phenomenon using fuzzy numbers to represent uncertainty in groundwater systems, including contaminant concentration and subsurface characteristics (Zhang et al., 2009; Kumar et al., 2006; Abebe et al., 2000; Freissinet et al., 1999; Dou et al., 1997). Fuzzy number based methods have also been developed to link the contaminant concentration to health risk assessment in groundwater systems (Kumar et al., 2006; Guan & Aral, 2004; Kentel & Aral, 2004; Chen et al., 2003). In a different type of application, Maskey et al. (2004) used fuzzy numbers to describe a rainfall-runoff process, where the temporal uncertainty in precipitation was represented by fuzzy numbers. Another example, an early application in hydrology, was illustrated by Bárdossy et al. (1990), where fuzzy numbers were used to study an imprecise relationship between soil electrical resistivity and hydraulic permeability. In more recent studies, fuzzy set theory has been used in water resource management problems, where uncertainties from multiple sources and types have been combined.

Applications involving DO have also been developed, though they do not specifically use fuzzy numbers, but use other fuzzy set theory based methods, such as fuzzy logic and fuzzy pattern recognition. Mujumdar & Sasikumar (2002) defined the risk of the occurrence of low DO concentration as a fuzzy event, i.e. instead of assigning one value to the probability of a low DO event, a range of values in the form of a fuzzy set are defined. Giusti & Marsili-Libelli (2009) used a fuzzy pattern recognition technique to define the natural variation, and to predict the dynamics of DO in a lagoon. Altunkaynak et al. (2005) also modelled DO dynamics, and used a fuzzy logic inference system for forecasting using artificial neural networks.

The literature clearly demonstrates the utility and advantage of using fuzzy numbers in environmental problems. In the examples discussed above, the representation of uncertain data (e.g. hydraulic conductivity or contaminant concentration) as fuzzy numbers has been widely used, though with inconsistent techniques. However, the representation of fuzzy relationships or systems has been limited. In the case of the Bow River in Calgary, uncertainty exists not only in the abiotic parameters used to predict DO concentration, but also in the relationships between the abiotic factors and DO. Thus, a method that uses fuzzy numbers to represent parameter and system uncertainty to predict DO
concentration using a data-driven approach would be highly advantageous to water resource managers. Mathematical techniques have been developed, each suited to the particular application, in order to use fuzzy numbers in crisp relationships (e.g. in physically-based models). To correctly incorporate fuzzy numbers into numerical models, be it physically-based or data-driven, classical mathematics needs to be adapted to account for fuzzy numbers. Fuzzy arithmetic (covered in detail in Kauffman & Gupta 1984, and also briefly described in the Appendix C), including the extension principle and the \( \alpha \)-cut approach, have been developed to provide these tools.

4.1.2 Fuzzy linear regression

Linear regression is a widely used data-driven technique for prediction. The general purpose of regression is to explain the uncertainty and variability in a dependent variable through a series of explanatory or independent variables, leading to a prediction equation (Bárdossy 1990; Bárdossy et al., 1990; Peters 1994; Chang & Ayyub, 2001; Kahraman et al., 2006). When large data sets are available regression may provide realistic results, but in its absence, as is often the case in hydrological applications, statistical regression should not be used (Bárdossy et al., 1990). In addition to this, statistical regression is inappropriate when it is difficult to verify strict distribution assumptions (including whether or not the data are mutually independent and identically distributed), and vagueness and ambiguity exists in the relationships between variables (Peters, 1994; Kim et al., 1996; Kahraman et al., 2006). Fuzzy linear regression is an attempt to extend linear regression for applications involving fuzzy numbers. It provides an alternative method in situations where crisp linear regression may not be possible, whether it be due to the inability to meet strict assumptions or if there is obvious fuzziness in the underlying data or process (Bárdossy et al., 1990; Kim et al., 1996; Kahraman et al., 2006). A number of different fuzzy linear regression methods have been created, each aiming to mimic the concept of crisp regression but with fuzzy numbers.

In crisp regression, the deviations from the true data can be caused by measurement and modelling error. Though measurement error may be well defined, modelling error (which would occur even if using precise data) is often not (Bárdossy, 1990; Chang & Ayyub, 2001). However fuzzy regression 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. These two distinct types of uncertainty are only part of the overall uncertainty. Ordinary linear regression provides tools that are suitable for dealing with crisp data, and only when the true relationship between the data are known (Peters, 1994; Chang & Ayyub, 2001; Lee & Chen, 2001; Kahraman et al., 2006).

There are two major types of fuzzy regression. The first is known as the possibilistic fuzzy regression developed by Tanaka et al. (1982), which is a linear programming approach that aims to minimise the fuzziness of the system. The second is a fuzzy least-squares approach, where in broad terms the distance between two fuzzy numbers is minimised (Diamond 1988; Kahraman et al., 2006). The goal of each method is to fit fuzzy data within a specific fitting criterion. The choice of criterion whether it is to minimise system fuzziness or to minimise the distance between fuzzy numbers governs the final regression equation (Chang & Ayyub, 2001). In addition to this, there is also a choice in the type of parameters to be used: fuzzy or crisp input data, fuzzy or crisp output data, and fuzzy or crisp regression coefficients (i.e. fuzzy relationships). This type of flexibility allows fuzzy regression to provide application specific formulations.

The advantage of using fuzzy regression is that some of the strict assumptions of the statistical model can be relaxed (Bárdossy et al., 1990; Peters, 1994). In classical regression it is assumed that the each error in the predicted variable is the same, whereas in fuzzy regression, the individual membership function describes the variable’s unique uncertainty. In addition to this, fuzzy regression can provide insight to physical systems even with small sets of data; in fact the performance of fuzzy regression models improve (as compared to crisp regression) when the size of the datasets are reduced (Bárdossy et al., 1990; Kim et al., 1996). Another benefit of fuzzy regression is that it can provide better generalization of data trends or patterns compared to crisp data (Bargiela et al., 2007). This is because the membership functions associated with fuzzy numbers have significant and useful information that is not incorporated in crisp regression.

### 4.1.3 Chapter objectives

In this chapter the use of fuzzy numbers to characterise and propagate the uncertainty in a data-driven system for predicting daily DO concentration in a riverine environment is explored. Specifically, the objectives are to: i) characterise the abiotic factors (limited to
the flow rate $Q$ and water temperature $T$) that influence DO concentration as fuzzy numbers; and ii) determine the efficacy of using fuzzy numbers to predict daily DO concentration and its uncertainty with a fuzzy linear regression model. To meet the first objective, an algorithm is described to convert a series of imprecise, highly variable and uncertain crisp measurements into fuzzy numbers. The second objective is accomplished through an improved fuzzy linear regression technique proposed to predict fuzzy daily DO concentration using fuzzy abiotic input parameters. The proposed new method is compared to two popular fuzzy linear regression techniques to highlight its novelty and significance. The results from this method are then used to illustrate the benefit of using fuzzy numbers as opposed to crisp numbers. Data from the Bow River in Calgary are used to illustrate this methodology.

4.2 Methods

4.2.1 Site description and data collection

The Bow River flows south-eastward from its headwaters in the Rocky Mountains through the foothills and the prairies before reaching the City of Calgary. The average annual discharge in the Bow River within the city limits is approximately $90 \, \text{m}^3/\text{s}$ and the average width and depth of the river are 100 m and 1.5 m, respectively. Three wastewater treatment plants (WWTP) located at Fish Creek, Bonnybrook and Pine Creek discharge their effluent into the Bow River. These WWTP are illustrated in Figure 4-1.

The objective of this study is to predict daily DO concentration using two abiotic factors, namely the flow rate, $Q$, and river water temperature, $T$. The reasons for selecting these two parameters have been discussed previously in He et al. (2011) and Khan et al. (2013). The required data were collected from various sources. Hourly $Q$ measurements from the Water Survey of Canada (WSC) were collected for the monitoring station “Bow River at Calgary 05BH004” which is highlighted in Figure 4-1 (Environment Canada Water Office 2013). Water quality data, specifically DO concentration and $T$, were measured using continuous YSI sondes from Pine Creek (PC in Figure 4-1) which is approximately 20 km downstream of the Bonnybrook WWTP. The sondes were

---

6 The MATLAB code associated with this section is included in Appendix D.
calibrated on a weekly basis and recorded water quality data every 15 minutes; though the interval increased to 30 minutes on some days.

Figure 4-1 An aerial view of the City of Calgary, highlighting the Bow River, the Water Survey of Canada (WSC) monitoring station, and the Bonnybrook (BB), Fish Creek (FC) and the Pine Creek (PC) wastewater treatment plants

Data for each variable was collected for the period between May 22 through November 22 for the years 2006, 2007 and 2008. This period coincides with the ice-free period in
the Bow River, beyond which these data are typically not collected. During this period, the average annual Q ranged from 99.10 to 131.83 m$^3$/s while the minimum and maximum observed Q was 48.67 and 352.17 m$^3$/s, respectively. Low flows are typically observed from January until the onset of spring freshet in May when snowpack melt from the Rocky Mountains starts to contribute to flows. Discharge peaks in mid-June due to the combination of snowmelt from the mountains and precipitation within the watershed. The discharge then reduces until late October when baseflow conditions are reached (Alberta Environment River Basins, 2012).

Average annual DO concentration ranged from 9.52 to 10.02 mg/L throughout the study period, while the minimum and maximum observed concentration was 3.81 and 18.32 mg/L, respectively. On an annual basis, the occurrence of low DO concentration tends to occur during the summer months, corresponding to low flows and high temperatures. This critical period corresponds to the selected experiment period for this research. Typically daily minimum DO concentration is high and daily DO variability is low during the winter, ice covered period, when the risk to aquatic habitat is negligible. Note that DO concentrations fell below the provincial guidelines of 5 mg/L in 2006, but not in 2007 or 2008. The average annual T for the study period ranged from 11.72 to 12.96°C, while the minimum and maximum were –0.05°C and 22.44°C, respectively. None of the parameters exhibited a significant trend ($p = 0.05$) during the monitoring period. A sample time series plot of sub-daily observations for the three variables in 2008 is shown in Figure 4-2.
Figure 4-2 Daily observations of DO concentration, water temperature, T, and discharge, Q, in the Bow River for the ice-free period (May - November) in 2008
4.2.2 Creating fuzzy numbers

The definition of a fuzzy set can be found in many textbooks on the subject (e.g. Kauffman & Gupta, 1984; Dubois & Prade, 1980; Novak, 1989). A fuzzy quantity is a numerical representation of imprecise data using fuzzy sets. A fuzzy number is a specific type of fuzzy quantity, one that aims to represent one particular quantity, rather than an interval of values. The mathematical definition of a fuzzy number is detailed in the Appendix C, including that of a triangular (“L-R”) fuzzy number, which is the most common and simplest realization of fuzzy numbers. These types of fuzzy numbers have membership functions defined by only three points (the lower and upper bound where \( \mu = 0 \), and the modal value where \( \mu = 1 \)) and assumes that the membership level monotonically increases, linearly from the lower bound to the modal value, and then monotonically decreases, linearly to the upper bound. These types of fuzzy numbers are believed to provide a reasonable representation of the possible values of a variable. Being linear functions, triangular fuzzy numbers are computationally inexpensive but may not sufficiently describe the membership level of observed data. Other shapes for membership functions are also prevalent and may provide more realistic representation of the observed data (Khan et al., 2013). By necessity, these non-linear functions are defined at more than three membership levels (e.g. \( \mu = 0.25, 0.50, 0.75 \) and so on), and are thus referred to as higher resolution membership functions.

The membership function of a fuzzy number is mathematically equivalent to its possibility distribution, and the terms are often used interchangeably (Zadeh, 1978). However, it is important to note that possibility or degree of membership of a variable should not be confused with the probability or likelihood of occurrence. Possibility and probability are conceptually different: possibility theory is built on the weakness of probability theory, and does not subscribe to additivity (Zhang, 2009). In probability theory, the probability of an event completely determines the probability of the contrary or complementary event (i.e. it is self-dual) (Dubois et al., 2004). However, in possibility theory the event and contrary event are only weakly linked (Dubois & Prade, 1988). Since these are distinct concepts, it is possible to define the probability associated with some fuzzy number being measured (e.g. the probability that observed DO will be below the Alberta guideline), and the corresponding possibility distribution of the observed DO,
given by its membership function (e.g. $D\bar{O} = \{5.0, 4.8, 5.2\}$ mg/L, using notation described in Equation C-2). Detailed discussion and a derivation of the possibility distribution can be found in Zadeh (1978) and Dubois & Prade (1988).

Appendix C details three concepts, $\alpha$-cuts, the extension principle, and probability–possibility transformations, which are all used to convert crisp data to fuzzy numbers. Briefly, the $\alpha$-cut technique provides a bridge to connect fuzzy sets to crisp sets (Zhang et al., 2009; Zhang & Achari, 2010b; Wang et al., 2011; Wang et al.; 2012). The extension principle (Zadeh, 1975) provides a method for extending crisp mathematical functions to deal with fuzzy numbers. A probability-possibility transformation uses the $\alpha$-cut and extension principle to transform crisp data to fuzzy numbers (Dubois et al., 1993; Dubois et al., 2004; Zhang et al., 2009).

The data described in Section 4.2.1 were converted to a fuzzy number using a probability-possibility transformation detailed in Appendix C. The intention was to combine sub-daily measurements of a variable (e.g., Q, T or DO) into one fuzzy number representation of a daily measurement. Using this method, as opposed to the common approach of using mean daily data, all available information is contained in the fuzzy number. This fuzzy number not only characterises the magnitude of the modal value of the daily observations, but also the variability, i.e. the support (the left and right bounds) and the probability of the data distribution, indirectly through the membership function. The specific algorithm developed to conduct this transformation for the T, Q and DO data is described below.

A daily data array, $X$ of length $N$, representing the $N$ observations in a 24 hour period, were collected for each variable. The first step is to calculate the probability density function (PDF) of the given array. The data were sorted into bins, and the number of observations in each bin was calculated. This series represented the PDF of the daily data. This PDF has to be created to ensure it is uni-modal, which means that it is a function of the bin size. Another constraint on the bin size is that it cannot be smaller than the sensitivity of the instrument used to measure each of the variables. This error term, $\varepsilon$, was subtracted from the minimum ($X_{\text{min}}$) and maximum ($X_{\text{max}}$) values of $X$, essentially to extend the support of the fuzzy number. The magnitude of $\varepsilon$ is $\pm 0.3^\circ C$ for T, the greater of $0.2$ mg/L or $2\%$ of $X$, for DO, and $2, 3$ or $8\%$ of $X$, when Q is less than $80$ m$^3$/s,
between 80 and 200 m$^3$/s, and greater than 200 m$^3$/s, respectively. The $\varepsilon$ terms for T and DO are based on the accuracy of the YSI sonde used for measuring the data (YSI Environmental, 2012). The three intervals for Q correspond to low, medium and high flow regimes (as defined by He et al., 2011), while the corresponding values of error are based on the literature (Shiklomanov et al., 2006; Di Baldassarre & Montanari 2009; McMillan et al., 2010). The bin size $\Delta$ was calculated as simply:

$$\Delta = \frac{(X_{\text{max}} - X_{\text{min}})}{N}$$

Equation 4-1

Once the bin size was defined, the start and end point of the bins had to be defined. The first bin, $X_{B1}$ was from $(X_{\text{min}} - \varepsilon - \Delta)$ to $(X_{\text{max}} - \varepsilon)$, and the last bin was from $X_{BN} = (X_{\text{min}} + \varepsilon + \Delta)$. Here the extra term $\Delta$ is included to ensure that the developed membership function has $\mu = 0$ at both the left and right supports. The bin corresponding to the maximum probability is identified; this value, $X_M$, corresponds to the kernel of the membership function $\mu(X)$. From here the objective is to find bins that have equal probabilities, i.e. for each $p(X) = H$, determine $X^L = p^{-1}(H)$ and $X^R = p^{-1}(H)$. In most cases, the data are not symmetric around $X_M$, and the pairs of $X$ and $p(X)$ from the histogram cannot be directly used to calculate $X^L$ and $X^R$. So, all $X^L$ and $P^L(X) = \sum p^L(X)$, and $X^R$ and $P^R(X) = \sum p^R(X)$ are collected in a table. A linear interpolation algorithm is then used to calculate the $X^R$ corresponding to $P^L(X)$, and the $X^L$ corresponding to the $P^R(X)$. This results in a membership function described at N elements, $X_{B1}$ to $X_{BN}$, each with $\mu(X) = \{p^L(X) + p^R(X)\}$. If for example, $N = 100$ and $X_M$ is located at the midpoint of $X$, the length of the vector $X_B = [X_{B1}, X_{B2}, ..., X_{BN}]$ is 102, which means that the length of the membership function (as a paired dataset of $X_B$ and $\mu(X)$) is also 102.

However, the membership function does not need such high a resolution to be described. For this research, five points, at $\mu = 0$, 0.25, 0.50, 0.75 and 1 are selected, to represent the fuzzy number. Here each $\mu$ has two corresponding $X_B$ values: $[X_B^L, X_B^R]_{\mu=\mu}$. Note that in the case of triangular fuzzy numbers, only two values of the membership function, at $\mu = 0$ and 1 are considered. In that case, it is assumed that the possibility distribution linearly increases between $\mu = 0^L$ and $\mu = 1$, and then linearly decreases from $\mu = 1$ to $\mu = 0^R$. However this assumption is not always suitable (Khan et al., 2013) and even a slight change in the possibility of a fuzzy number can have a larger impact in the probability
distribution of its crisp representation. Thus, in this algorithm, the membership function is defined using five values of $\mu$ rather than two.

The algorithm described above was used to develop the fuzzy numbers $\tilde{T}$, $\tilde{Q}$, and $\tilde{D}$ corresponding to the crisp data $T$, $Q$, and $D$, respectively, for the 2006, 2007, and 2008 datasets. An example of the observed and fuzzy number version of each variable is illustrated in Figure 4-3. Note that in this figure, convention is followed by connecting the markers in the right column with a straight dashed-line. Technically speaking, the membership functions developed using the algorithm described above is a discrete process, and does not calculate, nor assumes that the values of $\mu(X)$ between the selected points (at $\mu = 0, 0.25, 0.50, 0.75,$ and $1$) is a straight line. The algorithm only guarantees that the function is monotonically increasing, as required by the definition of a fuzzy number. In many applications, especially those using triangular membership functions, using two points (at $\mu = 0$ and $\mu = 1$), it is assumed that the function is continuous and linear at all $X$ and $\mu$. The results from this analysis challenge this concept and are clearly illustrated in Figure 4-3. Had the typical triangular approach been used, a straight line between the kernel value and the two support values would result in a membership function drastically different than the ones generated here. This difference is pronounced when it is assumed that the triangular membership functions are symmetrical, i.e. the left and right bounds are equidistant from $X_M$. Thus, if fuzzy numbers are to be incorporated into physically-based or data-driven models, it is important to accurately portray the prevailing “fuzziness” in the data (Peters, 1994). The algorithm proposed here arguably develops a more appropriate fuzzy number representation of uncertain environmental data compared to typical triangular fuzzy numbers.
Figure 4-3 A comparison of T, Q and DO observations from September 12, 2006 and their transformations into fuzzy numbers. Note that on the right column, the symbols correspond to $\mu = 0, 0.25, 0.50, 0.75,$ and 1.
4.2.3 Fuzzy linear regression methods

In the following sections three fuzzy linear regression methods are defined: the Tanaka method (Tanaka et al., 1982), the Diamond method (Diamond, 1988) and a new proposed method. A short derivation of each method is provided, followed by a description of their application to predict DÔ.

4.2.3.1 The Tanaka method

The Tanaka method for fuzzy regression is a linear programming problem, formulated as follows (from Tanaka et al., 1982). Given a set of crisp observations \( x_i \) and \( y_i \) \((i = 1, 2, ..., N)\), a fuzzy equation is required to predict a symmetrical triangular fuzzy number \( \tilde{Z} \):

\[
\{ z, z - e, z + e \},
\]

where:

\[
\tilde{Z} = A_0 + A_1 x + A_2 y
\]

Equation 4-2

where \( A_0, A_1, \) and \( A_2 \) are fuzzy number coefficients, defined as \( \{ A_j, A_j - c_j, A_j + c_j \} \), where \( A_j \) is the kernel and \( c_j \) is the half width of the support of \( \tilde{A}_j \), for \( j = 0, 1, 2 \). In other words the support of \( \tilde{A}_j \) is defined by \( [A_j - c_j, A_j + c_j] \). The objective of this fuzzy linear regression method is to minimise the fuzziness \( J \) (or as originally referred to, the “vagueness”) of the system, which is defined as:

\[
J = c_0 + c_1 + c_2
\]

Equation 4-3

Subject to \( c_j \geq 0 \) and:

\[
-A_0 - (1 - H)c_0 - A_1 x_i - (1 - H)c_1 - A_2 y_i - (1 - H)c_2 \leq -z_i - (1 - H)e_i
\]

\[
A_0 - (1 - H)c_0 + A_1 x_i - (1 - H)c_1 + A_2 y_i - (1 - H)c_2 \leq z_i - (1 - H)e_i
\]

Equation 4-4

for \( i = 1, 2, ..., N \)

In Equation 4-4, \( H \) is a user-defined variable referred to as the h-certain factor. This factor, which is a membership level, restricts the amount of data being used in the regression problem. For example, if \( H = 1 \), the problem is to minimise \( J \), only considering the crisp values of \( \tilde{Z}_i \), i.e. where \( \mu = 1 \). As \( H \) is reduced, more variability is introduced into the problem, until \( H = 0 \), where the full triangular fuzzy data for \( \tilde{Z}_i \) is being used.

High values of \( H \) mean that less uncertainty is incorporated into the regression model. This means that the resulting regression equation will not be able to propagate the
uncertainty that exists in the observed data. Essentially, \( H \) can be seen as another constraint in the problem: \( J \) is to be minimised subject to the constraint of the degree of membership fit, and the constraints in Equation 4-4 (Kahraman et al., 2006). By specifying \( H \), the linear problem is trying to minimise the residual between the observed and predicted equations, for only data at membership levels equal to or above the cut-off \( H \).

This method has drawn criticism since it was introduced; the key short coming being that least-squares is not utilised, and thus, the problem does not reduce to crisp regression if crisp parameters are used (Diamond, 1988; Chang & Ayyub, 2001). In essence, this method has replaced the random error component with the fuzzy error component, rather than combining the two. Another issue is that as the number of observations increases, the constraints, and hence, the amount of computation, increases proportionally (Chang & Ayyub, 2001). Also, the results of this technique are dependent on a number of factors: the symmetry assumptions of \( \mu \), the value of \( H \) used, and the choice of the fuzziness criterion \( J \) (which may be defined in a number of different ways, e.g. the sum, mean or maximum of the \( c_j \)) (Bárdossy, 1990; Bárdossy et al., 1990; Chang & Ayyub, 2001). Thus, the success of the model is directly influenced by the selection of these parameters, which are selected specifically to conform to the application, rather than a universal approach, as is expected of regression techniques (Bárdossy 1990). Another limitation of the Tanaka model is that the regressors are required to be crisp. However, even if the model is adapted to use fuzzy regressors (Hojati et al., 2005), the fundamental problems with the model described above are not resolved.

4.2.3.2 The Diamond method

A method developed by Diamond (1988) is described as follows. A key feature of this method is that it incorporates least-squares, which was a key shortcoming in the Tanaka method. A number of other fuzzy regression techniques that use least-squares with linear or quadratic programming have also been developed. These include the maximum compatibility method (Celminš, 1987) and the minimum fuzziness method (Savic & Pedrycz, 1991). Though each of these methods has its advantages, only the Diamond method is discussed here in detail.
In the Diamond method the objective is to minimise $D^2$ (a metric defined as the distance between fuzzy numbers) in the least-squares sense, and to determine the regression coefficients as a linear problem. Of the three models described by Diamond (1988), the method that incorporates fuzziness in the input, output and in one coefficient is extended to be applicable to cases with multiple regressors and non-symmetrical fuzzy numbers. This is described below.

Given a set of fuzzy observations $\hat{X}_i: \{x, x^L, x^R\}_i$ and $\hat{Y}_i: \{y, y^L, y^R\}_i$, ($i = 1, 2, ..., N$) a fuzzy equation is required to predict a triangular fuzzy number $\hat{Z}_i$: \{z, z^L, z^R\}, where:

$$\hat{Z} = \hat{A}_0 + a_1 \hat{X} + a_2 \hat{Y}$$  

Equation 4-5

where the coefficient $\hat{A}_0$: \{A_0, A_0^L, A_0^R\} is a triangular fuzzy number, and the coefficients $a_1$ and $a_2$ are crisp numbers. The objective of the fuzzy least-squares problem proposed by Diamond is:

$$\text{minimise } r(\hat{A}_0, a_1, a_2) = \sum_{i=1}^{N} D^2(\hat{Z}_i, \hat{A}_0 + a_1 \hat{X}_i + a_2 \hat{Y}_i)$$  

Equation 4-6

where

$$D^2(\hat{Z}_i, \hat{A}_0 + a_1 \hat{X}_i + a_2 \hat{Y}_i) = (z_i - A_0 - a_1 x_i - a_2 y_i)^2$$

$$... + (z_i^L - A_0^L - a_1 x_i^L - a_2 y_i^L)^2 + (z_i^R - A_0^R - a_1 x_i^R - a_2 y_i^R)^2$$  

Equation 4-7

for $i = 1, 2, ..., N$

A solution for this problem can be found by calculating the associated normal equations for Equation 4-6, for the four cases: when $a_1 \geq 0$ and $a_2 \geq 0$, $a_1 < 0$ and $a_2 \geq 0$, $a_1 \geq 0$ and $a_2 < 0$, or $a_1 < 0$ and $a_2 < 0$. The derivation of the solution is trivial and is not shown here. A key difference between this method and the Tanaka method is that along with the output observations $\hat{Z}_i$, the input values $\hat{X}_i$ and $\hat{Y}_i$ are also fuzzy, though limited to triangular fuzzy numbers. However, this approach prevents the coefficients $a_1$ and $a_2$ from being fuzzy, due to the necessity of linearity of $\mu$ (Diamond, 1988). This method essentially defines a metric to express the deviation between two fuzzy numbers, $\hat{Z}$ and $\hat{A}_0 + a_1 \hat{X} + a_2 \hat{Y}$, and then minimises this deviation in a least-squares sense. The
implication of this method is that if the error between each corresponding element of the two fuzzy numbers can be simultaneously minimised, then the overall distance $D^2$ is also minimised. Kao & Chyu (2003) modified this method to allow the regressors and regressand to be non-triangular membership functions, but kept the coefficient crisp.

Key differences between this method and the Tanaka method is that when crisp data are used, the minimization problem reduces to a conventional least-squares regression problem (Kao & Chyu, 2003). By allowing the constant coefficient $A_0$ to be fuzzy, the random and fuzzy errors are both represented (to an extent). In addition, the h-certain factor does not need to be specified, as the method only relies on the support and the kernel, assuming $\mu$ to be triangular.

The majority of research focuses on fuzzy regression systems where not all parameters, i.e. regressor, regressand and coefficients are all fuzzy at once. Typically the model coefficients are crisp, and the input and output data are fuzzy (Yang & Lin 2002; Kao & Chyu 2003; Bargiela et al., 2007). However, as noted earlier, the vagueness in a system may be caused by a combination of both measurement error and the indefiniteness of model parameters (Lee & Chen, 2001). This combined fuzziness is common when observations are from uncontrollable or uninfluenced sources, as is the case in hydrological and environmental data (Lee & Chen, 2001). A consequence of using fuzzy numbers to represent each of the three parameters is that their multiplication results in an increase in the spread of predicted data. In some applications this is seen as an unfavourable, since specificity is highly favoured. In contrast, when the objective is to be able to define the maximum map of possibilities, this larger spread is favourable. Further, a possibility–probability transform can be used to convert the uncertain predicted data into a PDF that determines the chance of these outlying data to occur. Thus the “correct” form for a fuzzy regression equation is dependent on the nature of the system being modelled and the nature of the input and output data (Bisserier et al., 2009).

4.2.3.3 A proposed new method

A method for fuzzy linear regression is proposed, which is an extension of the Diamond method described above. This proposed method is designed to be suitable for the required application, i.e. predicting fuzzy DO, using fuzzy input parameters, and fuzzy regression coefficients. This application has additional requirements that have not
been addressed in current literature (Yang & Lin, 2002; Nasrabadi & Nasrabadi, 2004; Hojati et al., 2005; Chachi et al., 2011): specifically that these fuzzy numbers are not limited to the triangular L-R types, $\mu$ is not restricted to symmetric functions, and each variable may have distinctly shaped $\mu$. In addition, the method should be appropriate for less general fuzzy numbers with discrete membership functions directly created from observed data via a probability–possibility transformation. The proposed method is outlined below:

Given a set of fuzzy observations $\tilde{X}_i$ and $\tilde{Y}_i$, and their corresponding membership functions, $\mu(X_i)$ and $\mu(Y_i)$, for $(i = 1, 2, ..., N)$ a fuzzy equation is required to predict a fuzzy number $\tilde{Z}$ defined by $\mu(Z_i)$, where:

$$\tilde{Z} = \tilde{A}_0 + \tilde{A}_1\tilde{X}_i + \tilde{A}_2\tilde{Y}_i$$

Equation 4-8

where the coefficients $\tilde{A}_j$ (where $j = 0, 1, 2$) are fuzzy numbers. The objective is to solve the following least-squares problem:

$$\text{minimise } r(\tilde{A}_0, \tilde{A}_1, \tilde{A}_2) = \sum_i d^2(\tilde{Z}_i, \tilde{A}_0 + \tilde{A}_1\tilde{X}_i + \tilde{A}_2\tilde{Y}_i)$$

Equation 4-9

where

$$d^2(\tilde{Z}_i, \tilde{A}_0 + \tilde{A}_1\tilde{X}_i + \tilde{A}_2\tilde{Y}_i) = \bigcup \{(\tilde{Z}_i - \tilde{A}_0 - \tilde{A}_1\tilde{X}_i - \tilde{A}_2\tilde{Y}_i)^2\}_\alpha$$

$$= \sum_{\alpha=0}^{1} [(\tilde{Z}_i - \tilde{A}_0 - \tilde{A}_1\tilde{X}_i - \tilde{A}_2\tilde{Y}_i)^2]_\alpha$$

Equation 4-10

for $i = 1, 2, ..., N$

Note that $d^2$ in Equation 4-9 and Equation 4-10 is not the same as $D^2$ in Equation 4-7. This “distance” now measures the sum of the squared-deviations of the observed ($\tilde{A}_0 + \tilde{A}_1\tilde{X}_i + \tilde{A}_2\tilde{Y}_i$) and predicted ($\tilde{Z}$) crisp intervals $[(\tilde{Z}_i - \tilde{A}_0 - \tilde{A}_1\tilde{X}_i - \tilde{A}_2\tilde{Y}_i)^2]_\alpha$, for all $\alpha$-cuts between $\mu = 0$ and $\mu = 1$. To calculate $d^2$, the intervals at $\mu = 0, 0.25, 0.50$ and 0.75 are
decomposed into its left and right bounds, whereas the interval at \( \mu = 1 \) (which cannot be decomposed into left and right values) is left as:

\[
d^2(\tilde{Z}_i, \tilde{A}_0 + \tilde{A}_1 \tilde{X}_i + \tilde{A}_2 \tilde{Y}_i) = \sum_{\alpha=0}^{0.75} \left\{ \left( (\tilde{Z}_i^L - \tilde{A}_0^L - \tilde{A}_1^L \tilde{X}_i^L - \tilde{A}_2^L \tilde{Y}_i^L)^2 \right) \right. ...
\]

\[
... + (\tilde{Z}_i^R - \tilde{A}_0^R - \tilde{A}_1^R \tilde{X}_i^R - \tilde{A}_2^R \tilde{Y}_i^R)^2 \right\}_{\alpha} \right\} + \left\{ (\tilde{Z}_i - \tilde{A}_0 - \tilde{A}_1 \tilde{X}_i - \tilde{A}_2 \tilde{Y}_i)^2 \right\}_{\mu=1}
\]

for \( \alpha = 0, 0.25, 0.50, \) and 0.75, and \( i = 1, 2, \ldots, N \)

Note that in Equation 4-11, the values of \( \alpha \) used are not limited to 0, 0.25, 0.50 and 0.75, but can be any set of values, provided that the interval at \( \mu = 1 \) is separated. Using Equation 4-11, the minimization problem is then redefined as:

\[
\text{minimise } r(\tilde{A}_0, \tilde{A}_1, \tilde{A}_2) = \\
\sum_{i} N \left\{ \min \left\{ \sum_{\alpha=0}^{0.75} (\left( (\tilde{Z}_i^L - \tilde{A}_0^L - \tilde{A}_1^L \tilde{X}_i^L - \tilde{A}_2^L \tilde{Y}_i^L)^2 \right) \right. \right. \right.
\]

\[
... + (\tilde{Z}_i^R - \tilde{A}_0^R - \tilde{A}_1^R \tilde{X}_i^R - \tilde{A}_2^R \tilde{Y}_i^R)^2 \right\}_{\alpha} \right\} \right\} + \left\{ (\tilde{Z}_i - \tilde{A}_0 - \tilde{A}_1 \tilde{X}_i - \tilde{A}_2 \tilde{Y}_i)^2 \right\}_{\mu=1} \} \right\}
\]

\[
= \sum_{i} N \left\{ \min \left\{ \sum_{\alpha=0}^{0.75} (\left( (\tilde{Z}_i^L - \tilde{A}_0^L - \tilde{A}_1^L \tilde{X}_i^L - \tilde{A}_2^L \tilde{Y}_i^L)^2 \right) \right. \right. \right.
\]

\[
... + (\tilde{Z}_i^R - \tilde{A}_0^R - \tilde{A}_1^R \tilde{X}_i^R - \tilde{A}_2^R \tilde{Y}_i^R)^2 \right\}_{\alpha} \right\} \right\} + \left\{ (\tilde{Z}_i - \tilde{A}_0 - \tilde{A}_1 \tilde{X}_i - \tilde{A}_2 \tilde{Y}_i)^2 \right\}_{\mu=1} \} \right\}
\]

\[
= \sum_{i} N \left\{ \min \left\{ \sum_{\alpha=0}^{0.75} (\left( (\tilde{Z}_i^L - \tilde{A}_0^L - \tilde{A}_1^L \tilde{X}_i^L - \tilde{A}_2^L \tilde{Y}_i^L)^2 \right) \right. \right. \right.
\]

\[
... + (\tilde{Z}_i^R - \tilde{A}_0^R - \tilde{A}_1^R \tilde{X}_i^R - \tilde{A}_2^R \tilde{Y}_i^R)^2 \right\}_{\alpha} \right\} \right\} + \left\{ (\tilde{Z}_i - \tilde{A}_0 - \tilde{A}_1 \tilde{X}_i - \tilde{A}_2 \tilde{Y}_i)^2 \right\}_{\mu=1} \} \right\}
\]

for \( i = 1, 2, \ldots, N \)

Equation 4-12 is derived from the fact that the minimum total \( d^2 \) results when the square deviation is a minimum at each \( \alpha \)-level. Similarly, at each \( \alpha \)-level, \( d^2_{\alpha} \) is minimum.
when both the left \((d^L_\alpha)\) and right \((d^R_\alpha)\) values are minimum. Thus, Equation 4-12 can be solved by the usual procedure of deriving a solution from the normal equations, resulting in:

\[
\begin{align*}
(X^L_\alpha)^T(X^L_\alpha) \beta^L_\alpha &= X^L_\alpha^T y^L_\alpha & \text{(a)} \\
(X^\mu_\mu=1)^T(X^\mu_\mu=1) \beta^\mu_\mu=1 &= X^\mu_\mu=1^T y^\mu_\mu=1 & \text{(b)} \\
(X^R_\alpha)^T(X^R_\alpha) \beta^R_\alpha &= X^R_\alpha^T y^R_\alpha & \text{(c)}
\end{align*}
\]

where \(X^L_\alpha, X^\mu_\mu=1\) and \(X^R_\alpha\) are matrices of the left, kernel and right bound observations of the regressors, \(\tilde{X}\) and \(\tilde{Y}\), respectively at each \(\alpha\). \(\beta^L_\alpha, \beta^\mu_\mu=1\) and \(\beta^R_\alpha\) are vectors of coefficients \(\tilde{A}^L_\alpha, \tilde{A}^\mu_\mu=1\) and \(\tilde{A}^R_\alpha\), respectively, for \(j = 0, 1, 2\), at each \(\alpha\), and \(y^L_\alpha, y^\mu_\mu=1\) and \(y^R_\alpha\) are vectors of the left, kernel and right bound observations of the regressand \(\tilde{Z}\), at each \(\alpha\). These equations result in the following system of equations:

\[
(X^T X) \beta = X^T y
\]

where \(X\) is a matrix of each \(X^L_\alpha, X^\mu_\mu=1\) and \(X^R_\alpha\) along the diagonal and zeros everywhere else, and \(\beta\) and \(y\) are vectors of the corresponding \(\beta_\alpha\) and \(y_\alpha\). Equation 4-14 is thus solved for \(\beta\), which gives the values at required membership levels for each fuzzy coefficient \(\tilde{A}_0, \tilde{A}_1\) and \(\tilde{A}_2\).

**4.2.4 Model implementation and comparison**

The 2006 and 2007 datasets of \(\tilde{T}, \tilde{Q}\), and \(\tilde{D}\) (the fuzzy versions of observed \(T, Q\), and \(DO\)), were used to construct each of the three fuzzy linear regression models. The 2008 dataset was used for model verification. It should be noted here that the logarithm of \(Q\) was used in each case. The models were implemented in MATLAB (version 2012b).

Since the Tanaka method requires crisp values for the regressors \(T\) and \(Q\) (see Equation 4-2), values of \(\tilde{T}\) and \(\tilde{Q}\) at \(\mu = 1\) were used. Also, the method required the regressand \(\tilde{D}\) to be a symmetrical, triangular fuzzy number; thus, values of \(\tilde{D}\) at \(\mu = 0^L, \mu = 0^R\) and \(\mu = 1\) were used to define \(\tilde{D}\) as \(\{DO_\mu=1, DO_\mu=0^L, DO_\mu=0^R\}\). Note that for symmetry \(DO_\mu=0^L\) must equal \(DO_\mu=0^R\); thus, the larger of the distances between each support and the kernel was used. Further analysis was conducted using four different values of \(H\): 0, 0.25,
0.50, and 0.75. Results are only presented for H = 0.25, as higher values do not capture enough of the observed variability, while lower values (i.e. H = 0) are highly uncertain. The Diamond method requires the regressors and regressand to be triangular fuzzy numbers, thus values of $\tilde{T}$, $\tilde{Q}$, and $D\tilde{O}$ at $\mu = 0^L$, $\mu = 0^R$ and $\mu = 1$ are used to define a three element fuzzy number for each variable. For the proposed method, the fuzzy numbers for each variable are represented at five $\alpha$-levels, corresponding to $\mu = 0$, 0.25, 0.50, 0.75, and 1; these values were selected to give a complete representation of the observed fuzzy number. In each model (Equation 4-2, Equation 4-5 and Equation 4-8) the coefficient $\tilde{A}_1$ was associated with $\tilde{T}$, and $\tilde{A}_2$ was associated with $\tilde{Q}$.

The results from model construction and verification were evaluated using the Nash-Sutcliffe model efficiency coefficient, $E$, defined as:

$$E = 1 - \frac{\sum_i (z_i - \hat{z}_i)^2}{\sum_i (z_i - z_m)^2}$$  
Equation 4-15

where $z_i$ is the observed value, $\hat{z}_i$ is the predicted value, $z_m$ is the mean of the observed values, of the regressand, and $N$ is the total number of observations. The metric $E$ was first calculated by grouping $z_i$ at each $\mu$ together and grouping $\hat{z}_i$ at each $\mu$ together. This meant that for the Tanaka and Diamond methods, the values of $z_i$ and $\hat{z}_i$ at $\mu = 0^L$, 0$^R$ and 1 were combined into one vector, while for the proposed method, their values at $\mu = 0$, 0.25, 0.50, 0.75 and 1 were combined into one vector. In addition to this, for the Tanaka and Diamond methods, values of $z_i$ and $\hat{z}_i$ at $\mu = 0.25$, 0.50, and 0.75 were interpolated (treating the membership function as a linear relationship). These data were then used to estimate $E$ individually at each $\mu = 0$, 0.25, 0.50, 0.75 and 1, and compared with the results from the proposed method at the corresponding membership levels. Similarly, the root mean square error (RMSE, defined in Equation 4-16) was calculated for each method, with the data at each membership level combined together. In addition to this, RMSE was calculated at $\mu = 0$, 0.25, 0.50, 0.75 and 1, using the interpolated data described above, for the Tanaka and Diamond methods, and then summed to give an
overall total error for each method. The objective of doing this was to demonstrate the
difference between errors when using triangular membership functions and when using
higher resolution membership functions (as is done in the proposed method).

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N}(z_i - \hat{z}_i)^2}{N}} \quad \text{Equation 4-16}
\]

The Akaike Information Criterion (AIC, defined in Equation 4-17) is also calculated
for each method, grouping all the data together, and as a sum at each membership level,
as described for RMSE.

\[
AIC = N[\ln (\text{RMSE})] + 2(k) \quad \text{Equation 4-17}
\]

where \( N \) is the total number of observations, RMSE is calculated by Equation 4-16, and
\( k \) is the number of parameters to be estimated in the models. Lastly, the metric \( D^2 \)
developed by Diamond (1988) and defined in Equation 4-7, along with the proposed new
metric, \( d^2 \) (defined in Equation 4-11), were also calculated for each method. These
metrics were used to define the objective function of the minimization problem and thus,
are evaluated to compare the efficacy of each method.

**4.3 Chapter results and discussion**

The coefficients for the three regression models were calculated and their values are
illustrated in Figure 4-4. The Tanaka model, which used crisp input observations (T and
Q), and fuzzy output observations (DÔ), calculated a crisp value for \( \tilde{A}_0 \), i.e. \( c_0 = 0 \), and
symmetric triangular membership functions for the fuzzy coefficients \( \tilde{A}_1 \) and \( \tilde{A}_2 \). This
implies that the best fit of the regressors to the fuzzy regressand, as defined by the
criterion \( J \), did not require any uncertainty to be present in \( \tilde{A}_0 \). The Diamond method used
fuzzy input and output observations and had a fuzzy \( \tilde{A}_0 \) and crisp \( a_1 \) and \( a_2 \). This resulted
in completely different values for the coefficients compared to the Tanaka method. The
Tanaka method predicted crisp \( \tilde{A}_0 \) and fuzzy \( \tilde{A}_1 \) and \( \tilde{A}_2 \), while the Diamond method
predicted fuzzy \( \tilde{A}_0 \) and crisp \( \tilde{A}_1 \) and \( \tilde{A}_2 \).
The proposed method, which did not have constraints on whether the coefficients, the regressors or the regressand were fuzzy or crisp, resulted in each coefficient being fuzzy, each with distinct, non-linear membership functions. This highlights the fact that assuming or selecting linear, or even a symmetric representation, of \( \mu \) may not be realistic. Also, by removing the crispness constraints on the coefficients, the proposed method produces completely different results compared to both the Tanaka and Diamond methods. The uncertainty, which was represented only in one coefficient in the Diamond method, is now transferred to all three coefficients, with drastically different values. Also, the new approach shows that without having predefined constraints on fuzziness, it still has the ability to generalise to a crisp number if necessary, as is the case for \( \tilde{A}_1 \), which is nearly fuzzy, with a small support. This shows that the regression model can generalise to crisp linear regression, if crisp regressors are used. By incorporating full system fuzziness, the regression model can provide insight into the physical-system, e.g. at low membership levels, when \( \mu = 0^L \), i.e. at low Q, the influence of the constant \( \tilde{A}_0 \) is much lower compared to \( \mu = 0^R \) when the constant is high (approximately 30, as illustrated in Figure 4-4).

The impacts of the different values of the coefficients are illustrated in Figure 4-5, Figure 4-6 and Figure 4-7. Each figure shows a time series plot of observed and predicted DÔ for each method. The observed DÔ is only shown at \( \mu = 0^L \) and \( 0^R \), whereas the predicted DÔ is shown at \( \mu = 0^L, 0^R, \) and 1. Data at these membership levels was sufficient to illustrate the overall results of each approach, thus data at other \( \mu \) is not shown. The Tanaka method was not able to capture the observed DÔ at any membership level. This is illustrated by the fact that in each plot, the predicted values at \( \mu = 0^L \) and \( 0^R \) (black crosses) are three to four times larger than the observed values. On top of this, the predicted DÔ at \( \mu = 1 \) (white circle) falls outside the observed DÔ interval at \( \mu = 0 \) (black circles). Generally speaking the Tanaka method is not appropriate for this application as the results from the regression equation do not provide meaningful predictions of DÔ.
In contrast to this, both the Diamond and the new method provide more realistic predictions of DÕ. A major difference between these two methods is immediately evident: though the Diamond method generally follows the observed trend, it cannot replicate the fluctuations in daily observed DÕ. This is illustrated by the band of black crosses with a near constant width throughout the entire dataset. In contrast to this, in the new method, the predicted data follow the transformed observed data much more closely. The predicted supports (black crosses) follow the observed supports (black circles) as the size of the fuzzy interval changes on a daily basis. The reason for this improvement can
be traced back to the formulation of the fuzzy regression technique. In the Diamond method, the coefficient $a_1$ and $a_2$ are crisp, essentially scaling the observed $\tilde{T}$ and $\tilde{Q}$, and thus limiting the propagation of uncertainty. In contrast, in the new method, the product of the fuzzy coefficients and fuzzy variable, produce values that are able to mimic the fluctuations. In addition to this, the regression problem defined by the Diamond method is not well equipped to predict this type of uncertainty: it is designed to minimise the overall uncertainty (i.e. at $\mu = 0^L$, $0^R$ and 1 combined). However, the approach of the new method is to minimise the distance on an $\alpha$-cut interval basis. This ensures that fluctuations at each membership level are matched by the predictions from the regression model. Of note is that the values of $D\tilde{O}$ at $\mu = 1$ are skewed closer to values at $\mu = 0^L$ than $\mu = 0^R$ in both the Diamond and the new method. This mimics the observed data where $D\tilde{O}$ is skewed to lower values (see Figure 4-3 for an example). The Tanaka method cannot capture this skewness since it is limited to symmetric fuzzy numbers.
Figure 4-5 Comparison of observed and predicted DO for model construction (2006 data), using the (a) the Tanaka method, (b) the Diamond method, and (c) the proposed method. Note that the observed data (black circles) correspond to $\mu = 0^L$ and $0^R$, while the predicted data are shown for $\mu = 0^L$ and $0^R$ (black crosses), and at $\mu = 1$ (white circles)
Figure 4-6 Comparison of observed and predicted DO for model construction (2007 data), using the (a) the Tanaka method, (b) the Diamond method, and (c) the proposed method. Note that the observed data (black circles) correspond to $\mu = 0_L$ and $0_R$, while the predicted data are shown for $\mu = 0^L$ and $0^R$ (black crosses), and at $\mu = 1$ (white circles).
Figure 4-7 Comparison of observed and predicted DÔ for model verification (2008 data), using the (a) the Tanaka method, (b) the Diamond method, and (c) the proposed method. Note that the observed data (black circles) correspond to μ = 0L and 0R, while the predicted data are shown for μ = 0L and 0R (black crosses), and at μ = 1 (white circles)
Figure 4-8 shows plots of observed versus predicted DO for each method, for both the construction and verification datasets. Note that these plots include data from all calculated membership levels. The Nash-Sutcliffe coefficient, E, is listed for each method on the figure. The new method has higher values of E compared to the other methods, with E equal to 0.75 and 0.64 for the construction and verification datasets, respectively. The Tanaka method has negative E values for both datasets, confirming that the method is not suitable for this application. The Diamond method fares better than the Tanaka method, with E equal to 0.67 and 0.45, for the construction and verification dataset, respectively. Of note is that the proposed method slightly under-predicts DO (Figure 4-8c and f), whereas the Diamond method both under and over-predicts DO (Figure 4-8b and e). Another important point to note is that in the Diamond method (Figure 4-8b and e) the predicted DO has a gap in values between approximately 10 to 12 mg/L, in both datasets. This gap represents the data that the method is unable to predict because it cannot capture the daily variability. This gap, and the spread of the data about the 1:1 line, is symptomatic of the shape of the membership functions used (i.e. assumed to be linear in the Diamond method and unconstrained in the proposed method), as well as the representation of uncertainty in each method. It highlights the importance of incorporating system uncertainty (via fuzzy regression coefficients), without which the full spectrum of data cannot be predicted (as seen in the Diamond method).
Figure 4-8 A comparison of observed and predicted DÕ, for the construction and verification datasets: (a) and (d) the Tanaka method, (b) and (e) the Diamond method, and (c) and (f) the proposed method
Figure 4-9a shows the values of E calculated for the Diamond and the proposed method at each membership level. Results for the Tanaka method are not included since the method substantially underperformed compared to the other two, and would plot outside the axis limits presented in Figure 4-9. Note that for the Diamond method, values at $\mu = 0.25$, $0.50$, and $0.75$ were calculated using the linearly interpolated predicted values. The figure illustrates that both methods have higher E when predicting the lower values (i.e. left bound) of DO, which are the values of interest in risk analysis. Both models are less accurate when predicting high DO concentration. Note that in all but two cases (at $\mu = 0.75^R$ and $0.50^R$ for the verification dataset), the proposed method has higher E values compared to the Diamond method, and this trend is consistent in both the construction and verification datasets. For the proposed method, E was calculated to be less than 0.2 on two occasions: at $\mu = 0.75^R$ and $0.50^R$ (for the verification dataset). In comparison, the Diamond method, E was below 0.2 on 12 occasions: at $\mu = 0.50^R$, $0.25^R$ and $0^R$ (for the construction dataset) and at all membership levels for the verification dataset. This shows that the proposed method is superior to the Diamond method, even at lower performance level (i.e. when E is less than 0.2). The lower E values of the proposed method, at $\mu = 0.75^R$ and $0.50^R$ for the verification data, are a consequence of the method used: it does not suggest a flaw in the new approach, in fact the opposite. Since each $\alpha$-cut interval is calculated independently for the proposed method, the predicted values reflect the highly variable nature of the daily data at each $\alpha$-cut, which the Diamond method does not attempt to reproduce. This is also reflected in the RMSE calculations discussed below. Similarly, the low E (less than 0.2) of the proposed method at the two aforementioned membership levels are not a reflection of overall model performance, nor related to the performance at other membership levels. Note that the Nash-Sutcliffe coefficient was not developed to measure the efficiency of models using fuzzy numbers; nonetheless, it has been presented here for comparison purposes though it may not be a suitable performance measure. Since a fuzzy number is a set of values, all analyses should be conducted using the entire set rather than values at individual membership levels. Thus, the “mean” value required to calculate E at individual membership levels, does not represent the “mean” of a series of fuzzy numbers, as it would for crisp numbers. Thus additional error terms are explored below for comparing model performance.
Figure 4-9b shows the RMSE values calculated at each membership level for the Diamond and proposed method. Again, for calculating the RMSE for the Diamond method, interpolated values are used for $\mu = 0.25$, 0.50, and 0.75. The RMSE values increase from the left bound to the right bound (i.e. from $\mu = 0^L$ to $\mu = 0^R$). However, this is not reflective of a decrease in performance, but is due to the fact that right bound values are associated with higher values of DO, due to the skewed nature of the data. In general, the new method has lower RMSE than the Diamond method for both datasets, particularly for the left bound (or low DO) data.

![Figure 4-9](image)

**Figure 4-9** A comparison of (a) Nash-Sutcliffe coefficient and (b) RMSE values for the Diamond and proposed method calculated at $\mu = 0$, 0.25, 0.50, 0.75 and 1

Table 4-1 lists the results of other model comparison criteria, including RMSE, AIC and $D^2$. RMSE was first calculated by combining data at all membership levels for each method. The proposed method had lower RMSE in each case. The total RMSE, which is the sum of RMSE presented in Figure 4-9b is also presented in Table 4-1 and is lower for the proposed method compared to the other two. Similarly, the results for AIC and total AIC are listed. In both of these cases, the proposed method has lower AIC. The significance of these results is that when non-linear membership functions are used to
represent the predicted DÕ, then the error between it and the observed DÕ at the intermediate membership levels are lower compared to assuming linear membership functions.

The bottom row of Table 4-1 includes results of calculating the metric $D^2$ (using Equation 4-7). $D^2$ was calculated for each predicted DÕ, summed, then divided by the length of the datasets. The results show that the proposed method outperforms the Diamond method, even when using the latter’s minimization criteria. The proposed method had lower values of $D^2$ for both the construction and verification datasets. Similarly $d^2$ was calculated using Equation 4-11, at $\mu = 0, 0.25, 0.50, 0.75,$ and 1. The results of these calculations are shown in Figure 4-10, which illustrates the fact that the new method outperforms the Diamond method. The $d^2$ values of the proposed method are lower at each membership level for both the construction and verification datasets. In fact, the difference between the two is greatest at $\mu = 0$, suggesting that the Diamond method is not suitable to model low or high DO. In addition to this, the results confirm assuming a linear membership function to determine the values of DÕ at $\mu = 0.25, 0.50$ and 0.75 is not appropriate, as this produces high errors compared to the non-linear functions used in the new method.

<table>
<thead>
<tr>
<th></th>
<th>Tanaka method</th>
<th>Diamond method</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>mg/L</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27.05</td>
<td>26.40</td>
<td>1.85</td>
</tr>
<tr>
<td>Total RMSE</td>
<td>mg/L</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>170.57</td>
<td>165.84</td>
<td>14.43</td>
</tr>
<tr>
<td>AIC</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1169</td>
<td>584</td>
<td>220</td>
</tr>
<tr>
<td>Total AIC</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8531</td>
<td>4261</td>
<td>1437</td>
</tr>
<tr>
<td>Total $D^2$</td>
<td>(mg/L)$^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2195</td>
<td>2090</td>
<td>10.29</td>
</tr>
</tbody>
</table>
Figure 4-10 A comparison of $d^2$ for the Diamond and proposed method calculated at $\mu = 0, 0.25, 0.50, 0.75$ and 1

An implication of the differences in results from the methods is illustrated in Figure 4-11. The figure shows observed versus predicted data points for the Diamond and the proposed method, concentrating specifically on the days where there is a possibility of low DO. As previously shown, the Diamond method is unable to capture the day-to-day variability in DO (Figure 4-5 to Figure 4-7), and has higher errors corresponding to low DO values. A result of this inflexibility is that the method predicts fewer days with a possibility of low DO compared to the proposed method. In the example shown in Figure 4-11, at $\mu = 0.5$, the Diamond method predicts that there is a possibility of low DO (specifically that DO is less than 5 mg/L) on 7 days, compared to 28 occurrences with the
new method. This difference is pronounced at higher $\mu$: at $\mu = 0.25^L$ and $0.5^L$, the new method predicts 16 and 4 days, respectively, where DO is less than 5 mg/L, compared to zero occurrences at both $\mu$ for the Diamond method. In addition to this, the Diamond method overestimates DO at $\mu = 0^L$, when the DÔ is observed to be less than 5 mg/L, as illustrated in Figure 4-11. These differences highlight two major results of this research: the proposed fuzzy regression method can predict DÔ with lower error and that its non-linear membership functions improve low DO predictions. In addition to this, the advantage of using fuzzy numbers rather than crisp numbers can be highlighted here as well. Had only crisp observations ($\mu = 1$) been used for this analysis, the linear regression equation would not have predicted any days to have a possibility, or probability, of low DO for the same dataset illustrated in Figure 4-11.

Lastly, to illustrate the difference between the results of the two methods, the membership functions of predicted DÔ are shown in Figure 4-12 for one day in 2006 and another in 2007. The functions clearly show that the proposed method does not follow the linear or triangular pattern, typically used to represent fuzzy numbers in environmental data. It is worth highlighting again that for the proposed method though the values at each membership level are connected with a straight, dashed line, this does not assume that the values between the selected $\mu$ follow the line (i.e. a linear relationship), as in assumed in the Diamond method. Lastly, the figure also shows that the support of DÔ for the proposed method is more flexible, in that it can match the observed changes in daily DO variation, whereas the Diamond method is less sensitive to change. This is a direct result of how the proposed regression method was designed.
Figure 4-11 A comparison of observed and predicted DO for the Diamond and proposed method, using the model construction dataset. The Alberta guideline for low DO (5 mg/L) has been highlighted.
A new fuzzy linear regression method is proposed in this chapter. The method builds on two previously defined fuzzy linear regression methods (namely, the Tanaka and Diamond methods) and compared to evaluate its performance. DO, Q and T data from the Bow River in Calgary, Alberta, Canada was used to demonstrate the performance of the model. First, an algorithm was developed to apply a probability-possibility transformation that converts the observed data into fuzzy numbers. The transformation was able to represent the skewness in the observed data. The fuzzy data were then used to test the three fuzzy linear regression methods, to predict fuzzy DO using fuzzy Q and T as the regressors.

In general, the new method was superior to the existing methods since it was able to represent both random and non-random uncertainty in the system, by using fuzzy regressors, regressand, and fuzzy coefficients. The new method was able to generate distinct, non-linear membership functions for each regression coefficient. The higher performance of the new method compared to the other two, suggests that linear or symmetric representation of uncertainty (i.e. \( \mu \)) in the coefficients may not be realistic. The new method also showed that it has the ability to reduce to crisp regression if necessary, a key requirement for fuzzy linear regression methods. This type of fuzzy linear regression approach has the ability to provide insight into physical system.

Figure 4-12 A comparison of membership function shapes for the predicted D\( \text{O} \) for (a) July 23, 2006 and (b) May 22, 2007
Comparing the three methods, the proposed method can model the daily fluctuations in DO variability more precisely and accurately compared to the other two methods. This is due to the way the regression system is set up, where the residuals between the observed and modelled data are minimised on an $\alpha$-cut interval basis collectively. One of the major advantages of this approach is that the skewness of the observed DO can be modelled effectively, giving a tool to predict low DO and other environmental data that are typically skewed. This advantage is highlighted by the fact that the Nash-Sutcliffe coefficient for the new method is higher than for the other two methods, and this difference is greater at low DO values. The proposed method outperforms the other two methods with lower RMSE and AIC at each $\mu$ and also has lower total RMSE and total AIC. The proposed method outperforms the Diamond method when using the Diamond $D^2$ and the proposed $d^2$ metric. The difference in performance reflected by $d^2$ is greatest at $\mu = 0$, which again suggests that the proposed method is better suited to predict low DO. The significance of these results is that the new method does not over predict low DO, as does the Diamond case. In fact, it highlights a larger number of days, where there is a possibility of low DO to occur, compared to the Diamond method. These results can be used by water resource managers to conduct risk assessments for the occurrence of low DO.
5. Comparing fuzzy and Bayesian linear regression methods

5.1 Chapter introduction
The dissolved oxygen (DO) concentration of a waterbody is a key indicator of overall aquatic ecosystem health (Dorfman & Jacoby, 1972; Hall, 1984). Low DO concentration, and the daily variability in DO, can increase the risk of adverse effects to the aquatic environment. While the impact of long term effects are largely unknown, low oxygen events can have devastating effects on biological communities (Adams et al., 2013). For this reason, it is widely measured and modelled in various jurisdictions as part of broader water quality assessment programs. A number of different biotic and abiotic factors are known to govern and impact the magnitude and diurnal fluctuation of DO concentrations in riverine environments (He et al., 2011). This includes the amount of aquatic organisms such as macrophyte and algae, the concentration of nutrients in the water column, the oxygen demand exerted by sediment in the riverbed, or the temperature of the water (Hauer & Hill, 2007; Pogue & Anderson, 1995). However, understanding DO variability in large urban environments is considerably more challenging. The added complexity resulting from the interactions of numerous factors, over a relatively small area and across different temporal scales means that DO trends in urban areas are more difficult to predict (Hall, 1984; Niemczynowicz, 1999). Secondly, rapid changes in the urban environment means that the factors and regimes influencing water quality and DO in the riverine environment might also change rapidly.

These issues highlight the limitations of physically–based models, which are typically used to model DO, and indicate that these types of models may be unsuitable in urban watersheds (He et al., 2011; Khan et al., 2013; Khan & Valeo, 2015a). This is because these models cannot capture the added complexity of the urban aquatic ecosystem; are typically calibrated at a scale (spatial and temporal) that is not representative of the entire system; require a complete, or near–complete understanding of all the factors affecting DO; or that the underlying processes governing DO are unknown (Cox, 2003; Khan & Valeo, 2015a). Secondly, given the rapid changes in urban areas, any long–term inferences made using static (i.e. non–updating) physically–based or conceptual models,

---

7 Material from this chapter was accepted for publication in the Journal of Environmental Informatics.
do not capture the impacts of these short–term changes (Radwan et al., 2009). Uncertainty in measured data and temporal and spatial variability compound these problems (Robinson et al., 2009; Khan & Valeo, 2015a) and highlight the need for a modelling approach that is: sensitive to the short–term changes in the urban environment; not limited to calibration at specific locations; and can capture the observed uncertainty within the model. In this chapter, we propose that a data–driven approach to predict DO concentration in an urban riverine environment is an attractive alternative to existing methods.

Data–driven models are a class of numerical models that create generalised links between input and output datasets (Solomatine & Ostfeld, 2008). Examples of data–driven models include linear regression, autoregressive models, neural networks, fuzzy regression and fuzzy rule–based systems, model trees, and genetic programming (Shrestha & Solomatine, 2008; Solomatine et al., 2008b; Elshorbagy et al., 2010). These methods have been widely used in environmental and hydro–informatics, because they provide good agreement between observed and modelled data, are generally easier to calibrate, are based on objective information and require limited assumptions about the physical process being modelled (Solomatine et al., 2008a; Solomatine et al., 2008b; Elshorbagy et al., 2010). In addition to this, data–driven models are useful in solving practical problems, especially when knowledge driven simulation models cannot be constructed due to lack of understanding of the processes or when existing models are inadequate (Solomatine et al., 2008b). Often it is assumed that data–driven models have higher data requirements to calibrate and validate the model, however, physically–based models also require extremely specific data to calibrate conceptual models, which are often very difficult to measure (Solomatine et al., 2008a; Antanasijević et al., 2014). Some have argued that the data requirements for data–driven models are in fact lower than deterministic models (Antanasijević et al., 2014). In addition to this, with increased use of real–time water quality monitoring stations, there exists a great opportunity to utilise high resolution data that is continuously being collected at numerous locations in many jurisdictions. Thus, the availability of data – which is often cited as a limiting factor in the success of data–driven models (e.g. Solomatine et al., 2008b) – is no longer
a major issue, particularly for major urban areas, where high resolution datasets are readily available.

Data–driven models have been used in a large number of studies to predict water quality parameters, including DO in many rivers across the world. A recent example of using these types of models to predict DO concentration in riverine environments include Wen et al. (2011) who used artificial neural networks (ANN) to predict DO in a river in China using ion concentration as the predictors. Similarly, Antanasijević et al. (2014) used ANNs to predict DO in a river in Serbia using a Monte Carlo approach to quantify the uncertainty in model predictions and temperature as a predictor. Chang et al. (2015) also used ANNs coupled with hydrological factors (such as precipitation and discharge) to predict DO in a river in Taiwan. Singh et al. (2009) used water quality parameters to predict DO and BOD in a river in India. Other studies have used multiple linear regression to predict DO in rivers using factors such water temperature, or electrical conductivity, amongst others: e.g. Heddam (2014) for a river in Oregon, USA, Ay & Kisi (2012) for a river in Colorado, USA and He et al. (2011) for a river in Calgary, Canada. These studies have found that data–driven model performance for DO prediction is suitable, and model performance statistics have been high. This shows that modelling DO using data–driven techniques is growing in popularity, likely owing to the difficulty in fully defining and understating the physical mechanisms that govern DO.

Recent research on improving DO prediction in the Bow River in Calgary, Canada (the study site for the present research) has focused on including uncertainty in the data–driven models (He et al., 2009; Khan et al., 2013; Khan & Valeo, 2015a; Khan & Valeo, 2014a). In Calgary, rapid population growth and expansion has led to increased urbanisation, resulting in detrimentally low DO events in the Bow River. The river is of extreme importance for the City, given that it provides more than 60% of the population with potable drinking water, and is also used for irrigation, industrial and recreational activities (Robinson et al., 2009; BRBC, 2010a). The river has an average annual discharge of 90 m$^3$/s, an average width and depth of 100 m and 1.5 m, respectively (Khan & Valeo, 2015a). Given its size, and the rate of development in the watershed – the river is one of the most regulated rivers in Alberta (BRBC, 2010a) – the Bow River is approaching is assimilative capacity. The City of Calgary is mandated to control the total
loading of sediment and nutrients in the river to prevent detrimental impacts on the waterbody. The City has implemented numerous strategies to limit loadings into the river, and use numerical modelling as part of its strategy to predict not only the impacts of strategies on water quality, but also to forecast the future state of the river under different scenarios. Currently, the Bow River Water Quality Model (see Tetra Tech, 2013 and Golder, 2004 for details) a physically–based conceptual model is used to predict DO and other water quality parameters. However, research by He et al. (2009), Khan et al. (2013) and Khan & Valeo (2015a) has shown that data–driven models, particularly those that use abiotic factors (such as river discharge, or water temperature) as inputs, have promising results to predict DO concentration in the Bow River. The advantage of using readily available data in these studies was that if a suitable relationship between these factors and DO could be found, changing the factors (e.g. increasing the flow rate downstream of a treatment plant) could improve DO at times when the risk of low DO was high.

However, a general drawback of numerical representation of physical systems is uncertainty (in its various forms, such as in the input and output data, in the model structure, and in the parameters). While in physically–based models, the structure of the model is assumed to be true (e.g. the empirical or theoretical relationship), uncertainty is typically only thought to be present in the model parameters. However, this does not take into account that these process models may be incomplete, e.g. a certain process is missing. In contrast to this, there might exist structural uncertainty in data–driven models, i.e. whether or not the best model structure has been selected. In both cases, this type of uncertainty is due to imperfect model structure (Shrestha & Solomatine, 2008). Scarcity in data, measurement errors and variability over time scales can contribute to uncertainty in modelling projections (Zhang et al., 2009; El–Baroudy & Simonovic 2006). Additionally, uncertainty may be increased when data from multiple sources are used, integrated and propagated (Porter et al., 2000; Shrestha & Nestmann 2009; Shrestha & Simonovic 2010; Suo et al., 2013). The proper identification and propagation of this uncertainty is critical for understanding and evaluating model prediction (Shrestha & Nestmann, 2009; Li et al., 2009).

Typically, to address uncertainty, probability based methods, such as Bayesian inference have been used to describe the uncertainty in models (Vrugt et al., 2009; Freni
& Mannina et al., 2010; Tyralis & Koutsoyiannis, 2013; Gelman et al., 2014). The use of Bayesian based methods has increased significantly in recent years, due to the fact that computing speed has increased significantly (Koop, 2003). Thus, numerical solutions (such as Monte Carlo simulations, which are almost always required) for Bayesian problems can now easily and quickly be found for cases where no analytical solutions exist (Greenberg, 2008). The basic principle of Bayesian applications in uncertainty analysis in numerical models is that any prior information (i.e. information gleaned from previous research or experience) can be included into the current analysis to estimate or update current parameters (Birkes & Dodge, 1993; Shrestha & Solomatine, 2009; Gelman et al., 2014). This essentially weighs the results from current data (referred to as the likelihood in Bayesian terminology) against prior data. The posterior then, according to Bayes rule, is proportional to the product of the likelihood and prior. This implies, for example, that any calculated model parameters are a function of the likelihood function (calculated from the current data) and the prior (from any previous data). The advantages of using a prior and Bayesian analysis in general, is that it prevents model “over–learning”. More generally, Bayesian analysis provides a statistical or probabilistic representation of a system rather than the typical deterministic representation, thus, providing confidence in predictions (Thiemann et al., 2001; Kingston et al., 2005). This means that the uncertainty in a model is represented by including probability distributions of various parameters in a model. It also allows the inclusion of subjective or expert opinion (through the prior).

Since Bayesian inference of model parameters is a probabilistic technique, it requires strong assumptions regarding the probability distribution functions for parameters, which may not always be suitable. Assuming or substituting an approximate distribution in place of the true or exact distribution might lead to large errors (He & Valeo, 2009). In addition to this, there is an on–going debate on the sensitivity of Bayesian analyses on the selection of a suitable prior: in many cases a subjective prior may give vastly different results than an objective prior (Ordaz et al., 1994; Freni & Mannina, 2010). In general, in many cases where uncertainty in model parameters is important, a Bayesian analysis is preferred over a simple deterministic analysis (Vicens et al., 1975). It should be noted, however, that uncertainty in data–driven modelling techniques is not purely random or
probabilistic in nature (Dubois & Prade, 1997; Ozbek & Pinder, 2006), as is implied by adopting a Bayesian framework. An alternative to the probability based representation of uncertainty is through the use of fuzzy set theory, particularly in relation to possibility theory and fuzzy numbers.

A fuzzy number is a specific type of quantity that expresses uncertain or imprecise quantities, measurements or observations (Khan & Valeo, 2015a; Huang, et al., 2010). They are more suitable when data is missing, incomplete, combined from multiple sites, and to represent uncertainty that is not purely random in nature (Bárdossy et al., 1990; Guyonnet et al., 2003; Zhang & Achari, 2010a; Huang et al., 2010). A fuzzy number is a convex and normal fuzzy set (Zadeh, 1965). A fuzzy number \( A \), with elements \([a_1, a_2, ..., a_n]\) is defined by its membership function, \( \mu \), which assigns a membership level for each element of \( A \). For a fuzzy number, \( \mu \) is bounded in the interval \([0, 1]\). A linear (or triangular) membership function is typically used since it represents the highest form of uncertainty (akin to a uniform probability distribution), however this form may not be the most suitable for environmental data (Khan et al., 2013, Khan & Valeo, 2015a). The consistency principle (Zadeh, 1978; Dubois et al., 1993) links possibility and probability, implying that something has to be possible before it is probable. A number of methods have been created that utilise this principle to convert observed data into fuzzy numbers. A summary of these methods is available in Oussalah (2000).

As in the Bayesian inference case, fuzzy numbers can incorporate expert opinion, and also provide more meaningful information than traditional, non–fuzzy (referred to as ‘crisp’) numbers (Khan et al., 2010; Khan & Valeo, 2015a). Since fuzzy numbers are based on possibility theory, it means that some of the strict assumptions in many probability models can be relaxed when dealing with fuzzy numbers, making it useful for environmental systems (Peters, 1994; Kim et al., 1996; Kahraman et al., 2006). The nature of fuzzy numbers lends itself well to risk analyses, since the risk of a fuzzy number to cross a given threshold (e.g. water quality guidelines) and can be directly inferred from the fuzzy number. Fuzzy numbers have been widely used in countless environmental and hydrological applications. The literature demonstrates the utility and advantage of using fuzzy numbers and a summary of some of these applications can be found in Khan & Valeo (2015a).
5.1.1 Chapter objectives

The above discussion shows that there is considerable opportunity to use both Bayesian and fuzzy number based methods, along with data–driven models to improve the prediction of environmental parameters such as DO. In this chapter, we formulate a Bayesian linear regression and a fuzzy linear regression approach to predict DO in the Bow River in Calgary. To do this, we rely on real–time DO data that is already being collected (in this case by the City of Calgary, as part of its on–going long term water quality assessment program), and we eliminate the need for other data (hence reducing uncertainty) by using an auto–regressive approach. Finally, we propose a recursive algorithm that is updated as more data is added into the modelling system. The advantage of this approach would be to capture the rapid changes in the urban environment reflected in the new data. This is contrast to typical modelling structures where part of the data is allocated for calibration, and the remainder for validation. The use of Bayesian and fuzzy methods ensures that the uncertainty is captured in the respective models. A comparison between the two methods will be conducted to determine the efficacy of each, and the conditions when one method outperforms the other.

5.2 Methods

5.2.1 Site description and data collection

The Bow River basin, located in southern Alberta, Canada, is approximately 25,123 km$^2$ in size and the river is 645 km long (BRBC, 2010a). The headwaters are located at Bow Lake, in the Rocky Mountains, from where it flows south–easterly to Calgary, before meeting the Oldman River south–east of the City, and eventually draining into Hudson Bay (Robinson et al., 2009; Environment Canada, 2014). The Bow River is supplied by precipitation accumulated in the snowpack in the Rocky Mountains, rainfall and discharge from shallow groundwater.

The City of Calgary routinely samples a variety of water quality parameters along the Bow River within the City limits, primarily to measure the effect of urbanization, including the impacts of wastewater treatment plant effluent and stormwater runoff on the Bow River. Real–time water quality monitoring systems are stationed at the upstream (at

---

8 The MATLAB code associated with this section is included in Appendix G.
the Bearspaw reservoir) and downstream (currently at Highwood, see Figure 5-1) ends of the City. Comparing water quality results from these two stations shows the direct impact of the City on the water quality, and thus the health of the Bow River (Khan & Valeo, 2014a). DO concentration measured at the upstream site is generally high throughout the year, with little diurnal variation (He et al., 2011; Khan et al., 2013; Khan & Valeo, 2015a). However, DO concentration downstream of the City limits is typically lower, and experiences much higher diurnal fluctuations. The three wastewater treatment plants (Bonnybrook, Fish Creek and Pine Creek, see ) located upstream of this monitoring site, and other impacts of urbanization are thought to be responsible for the degradation of water quality at the Highwood monitoring station.

For this research, real–time DO concentration data was collected for this downstream station for the period of 2004 to 2013. Between 2004 and 2007 the monitoring station was located at Pine Creek and sampled DO (along with a suite of other standard water quality parameters) every 30 minutes (for 2004 and 2005), and every 15 minutes (for 2006 and 2007). In 2008 the station was moved to Stier's Ranch, where it remained until 2011, and sampled data every hour (in 2008) and every 15 minutes from there on. The site was moved further downstream to its current location (at Highwood) in 2012 where it sampled every 15 minutes until 2013.

A YSI sonde is used to monitor DO (along with the other water quality parameters. The accuracy of DO concentration measurement using this sonde is listed by the manufacturer as ±2% or ±2 mg/L, whichever is greater (YSI Inc., 2013). The sonde is not accurate in freezing water, thus only data from the ice free period was considered, which is approximately from April to October for most years. Since low DO usually occurs in the summer (corresponding to high water temperature and lower discharge), the ice–free period dataset still contains the dates that are of interest for low DO modelling.
Figure 5-1 Aerial image of the City of Calgary showing the locations of (a) Bearspaw water treatment plant, (b) Bonnybrook, (c) Fish Creek, and (d) Pine Creek wastewater treatment plants, and two sampling locations (e) Stier’s Ranch and (f) Highwood. Note that the Bow River flows from the northwest to the southeast, through the centre of the City

5.2.2 Model implementation

A preliminary analysis of the DO concentration (Khan & Valeo, 2014a) shows that the daily mean DO concentration is highly serially correlated with lagged daily mean DO concentration (the correlation was high for 1, 2 and 3 day lags). This correlation was
stronger than any other relationship explored (with other water quality parameters such as water temperature, or other environmental parameters such as river discharge, or climatic data such as solar radiation). Thus, for this research, a model was sought in the following form:

$$DO(t) = f[DO(t - d)]$$  \hspace{1cm} \text{Equation 5-1}$$

where \(DO(t)\) is the mean daily DO for day \(t\), and \(DO(t - d)\) is the mean daily DO for day \((t-d)\), and \(d\) is in units of days. Values of \(d\) were selected as 1, 2, 3, and 7 days – this was to see the applicability of data–driven models for short–term model predictions. As mentioned above, two regression techniques were explored for this research: a Bayesian and a fuzzy linear regression approach. For both cases, the form of the model was:

$$DO = \beta_0 + \beta_1 DO_L$$  \hspace{1cm} \text{Equation 5-2}$$

where \(DO\) is the observed mean daily DO concentration, \(DO_L\) is the lagged mean DO concentration, and \(\beta_0\) and \(\beta_1\) are the regression coefficients. Typically, in a numerical modelling set–up, a portion of available data is used for model calibration, and the remainder is used for model validation. For this research, a different algorithm to calibrate and validate the model was implemented, a quasi–real–time, model updating system was used. This means, initially, one years’ worth of data (2004) was used to calibrate the regression model (either Bayesian or fuzzy), while data from the following year (2005) was use to validate the model. Then, two years of data were used to calibrate the model (2004 and 2005) and the next year (2006) for validation, and so on. This structure is used to simulate a real–time recursive algorithm, where when more data is available, the model updates its parameters (\(n.b\). a similar approach for hydrologic parameter estimation was adopted by Thiemann et al., 2001). By doing so, any changes in the system are implicitly captured by the data–driven model as each subsequent year is added to the dataset. The table below lists the nine models that were created using this approach and the associated datasets. These models were created for four different cases (lags of 1, 2, 3 and 7 days) for both the Bayesian and fuzzy regression cases.
Table 5-1 Details of data used to calibrate and validate each model in the recursive modelling

<table>
<thead>
<tr>
<th>Model no.</th>
<th>Calibration data</th>
<th>Validation data</th>
</tr>
</thead>
<tbody>
<tr>
<td>M01</td>
<td>2004</td>
<td>2005</td>
</tr>
<tr>
<td>M02</td>
<td>2004–05</td>
<td>2006</td>
</tr>
<tr>
<td>M03</td>
<td>2004–06</td>
<td>2007</td>
</tr>
<tr>
<td>M04</td>
<td>2004–07</td>
<td>2008</td>
</tr>
<tr>
<td>M05</td>
<td>2004–08</td>
<td>2009</td>
</tr>
<tr>
<td>M06</td>
<td>2004–09</td>
<td>2010</td>
</tr>
<tr>
<td>M07</td>
<td>2004–10</td>
<td>2011</td>
</tr>
<tr>
<td>M08</td>
<td>2004–11</td>
<td>2012</td>
</tr>
<tr>
<td>M09</td>
<td>2004–12</td>
<td>2013</td>
</tr>
</tbody>
</table>

An added utility of this method is that in Bayesian modelling techniques, a prior value is often needed to calibrate the model. Thus, with this approach, the model parameters obtained from M01 were used as priors for calibrating M02, and then these updated parameters used as priors for M03, and so on. For the first case (i.e. M01), a type of prior known as a non-informative prior was used. This prior assumes complete ignorance of prior values of the parameters, essentially a uniform distribution with an extremely large variance. Details of the informative and non-informative priors are described in detail in the proceeding section.

As mentioned earlier, the efficacy of data-driven modelling may be dependent on the amount and quality of data available. The available dataset has a resolution of 96 samples per day (corresponding to sampling every 15 minutes) for most years. As an additional test to compare the two data-driven models, the resolution of the input dataset was reduced from 96 samples, first to 24 samples per day (sampling every hour), and then to 6 samples per day (sampling every 4 hours). Then, the entire analysis (i.e. calibrating and validating the nine recursive models, M01–M09) was repeated using the lower resolution datasets, at each of the four lags. In other words, for the initial case, mean daily DO was calculated using 96 samples for each day, which reduced to using 24 samples for the second case, and only 6 samples in the last case. This was done to compare the change in
performance of both methods as the available data was reduced. Results from this component of the research can assist in determining an optimal data sampling scheme.

5.2.3 Bayesian linear regression

Bayesian linear regression (BLR) is an approach to ordinary linear regression (OLR; one of the most widely used data–driven models in environmental informatics) within a Bayesian framework. There are several advantages of using Bayesian regression over the ordinary case including: the use of prior information, the ability to update on existing evidence, improved performance for small samples, the ability to use data from different sources (through a strong prior), easier formulation in complex models, higher flexibility and greater stability, lower influence of outliers, and directly estimating uncertainty in the parameters and predicted values (Chen & Martin, 2009; Wakefield, 2013). This section describes the two Bayesian frameworks adopted for this research to prediction DO concentration. Details of the development of these methods are included in Appendix E. The typical OLR model can also be represented as a Normal probability model:

$$y \sim N(X\beta, \sigma^2 I)$$  \hspace{1cm} \text{Equation 5-3}

where $X$ is the matrix of dependent variables, limited to $[x_1 \ x_2]$ here with $x_1 = I$ to allow for an intercept, $x_2$ are the observations, the $\beta$ are the corresponding regression coefficients, $\sigma^2$ is the variance of the dependent variable $y$ given the observations (i.e. $y \mid \beta, X$), and lastly, $I$ is the identity matrix. Using Bayes rule, the posterior distribution of the OLR model can be written as:

$$p(\beta, \sigma^2 \mid y, X) \propto p(y \mid \beta, \sigma^2, X)p(\beta, \sigma^2)$$  \hspace{1cm} \text{Equation 5-4}

where the term on the left hand side represents the posterior, and the right hand side is the product of the likelihood and the prior. In many cases, prior information may not be available, making it difficult to describe the prior. Thus, an objective prior (i.e. a non–informative prior) is typically proposed. Thus, in a Bayesian framework, the regression model can then be parameterised using the following posterior hyper–parameters (which are identical to the OLR estimates):
\[
\mathbf{V} = (\mathbf{X}^T \mathbf{X})^{-1}
\]

\[
\mathbf{\bar{\beta}} = \mathbf{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}
\]

\[
s^2 = (y - \mathbf{X} \hat{\beta})^T (y - \mathbf{X} \hat{\beta})
\]

Using this formulation the predictive distribution of \(y\) can be found analytically and is multivariate \(t\):

\[
(y^*|y) \sim t(X^* \mathbf{\bar{\beta}}, s^2[I + X^* \mathbf{V} X^{*T}], \nu)
\]

where \(X^*\) are new observations used for prediction, \(y^*\) are the predicted values, and \(\nu\) are the posterior degrees of freedom. This formulation was used to model daily mean \(DO\) (i.e. \(y^*\)) using \(DO_t\) (i.e. \(X\)) as the regressor for the case where the first year of available data (2004) was used to calibrate the model and data from 2005 was used to validate the model (labelled as M01 in Table 5-1). This is because for this case, in 2004, there is no information available to choose a so-called informative, natural conjugate prior.

For subsequent models (i.e. M02 onwards), independent and informative priors were used. For example, in M02, two years of data are used to calibrated the Bayesian regression model (2004 and 2005) while data from 2006 is used to validate the model. For the calibration procedure, the posterior estimates of \(\beta\) and \(\tau\) (where \(\tau = 1/\sigma^2\)) from M01 are used as the priors for the estimates \(\beta\) and \(\tau\) for the M02 model. For the independent and informative priors, we assume \(p(\beta, \tau) = p(\beta) p(\tau)\), where \(\beta \sim N(\mathbf{\bar{\beta}}, \mathbf{V})\) and \(\tau \sim \Gamma(s^{-2}, \nu)\). However, using these priors an analytical solution for the posterior is not possible, so a numerical solution using a Gibb’s sampling algorithm was used (details of this algorithm are provided in Appendix E. In this case, the hyper-parameters for the posterior distribution are given as:

\[
\mathbf{\bar{V}} = (\mathbf{V}^{-1} + \tau \mathbf{X}^T \mathbf{X})^{-1}
\]

\[
\mathbf{\bar{\beta}} = \mathbf{\bar{V}} (\mathbf{V}^{-1} \mathbf{\bar{\beta}} + \tau \mathbf{X}^T \mathbf{y})
\]

\[
\nu = \nu + N
\]
\[
\bar{s}^{-2} = \frac{(y - X\beta)^T(y - X\beta) + \nu s^{-2})/\nu}{\nu}
\]

Equation 5-12

The predictive distribution of \( y \) can be estimated numerically: for each pair of \( \beta \) and \( \tau \) (from the Gibb’s sampler), a sample for \( y^* \) can be drawn from the predictive distribution:

\[
(y^*) \sim N(X^*\beta, 1/\tau)
\]

Equation 5-13

where \( X^* \) (DOI) are new observations used for prediction and \( y^* \) (DO) are the predicted values. The posterior conditional distributions from the Gibb’s sampling routine are then used as the priors for the M03 model, and so on. Thus, in this way data from each quasi–real–time model is used to add information into subsequent forms of the model.

5.2.4 **Fuzzy linear regression**

Fuzzy linear regression (FLR) is a method used to extend simple linear regression for applications involving fuzzy numbers, i.e. for uncertain or imprecise systems (Khan & Valeo, 2015a). This property makes them ideally suited for data–driven techniques that model environmental systems. It provides an alternative method when simple linear regression may not be possible, e.g. when assumptions of simple linear regression are not met, or if there is obvious fuzziness or uncertainty 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 & Ayyub, 2001). In simple linear regression models, the independent variable \( x \) predicts dependent variable \( y \) and it is assumed that \( x \) are observed without error (i.e. they are fixed). This assumption is typically only true when the independent variables are controlled and the effect on the dependent variable is measured. Often in environmental informatics applications, the dependent data are observed with an error, or bias, or are random rather than fixed (Madalla, 1988). This introduces a bias in the estimation of the model coefficients making an ordinary least square estimate of regression coefficients to be underestimated even for very large samples (Fuller, 1987; Madalla, 1988). Unlike a traditional simple linear regression model, FLR allows uncertainty in the input parameters to be included in the analysis.
The FLR used for this research was developed by Khan & Valeo (2015a) and is unique in that it uses fuzzy inputs, outputs and fuzzy coefficients to parameterise the regression. By doing so, the uncertainty in the data and in the structure of the model are all represented. In addition to this, the proposed method uses non–linear membership functions to define fuzzy numbers rather than the typical linear (or triangular) representation. This is more suitable in environmental applications, such as predicting water quality parameters (Khan et al., 2013; Khan & Valeo, 2015a).

Instead of minimising the residual between an observed and regressed value, the distance between two fuzzy numbers is minimised instead. Given a set of fuzzy observations \( \tilde{x} \) and \( \tilde{y} \), and their corresponding membership functions, \( \mu(\tilde{x}) \) and \( \mu(\tilde{y}) \), for \( (i = 1, 2, \ldots, n) \) a regression model is defined as:

\[
\tilde{y} = \beta_0 + \beta_1 \tilde{x} + \tilde{e}
\]

Equation 5-14

where the coefficients \( \beta_0 \) and \( \beta_1 \) are fuzzy numbers. The objective is to solve the following least–squares optimization problem:

\[
\min r(\beta_0, \beta_1) = \sum_{i=1}^{N} d^2(\tilde{y}_i, \beta_0 + \beta_1 \tilde{x}_i)
\]

Equation 5-15

where \( d^2(\tilde{y}_i, \beta_0 + \beta_1 \tilde{x}_i) = \mu[\big((\tilde{y}_i - \beta_0 - \beta_1 \tilde{x}_i)^2\big] \) for \( i = 1, 2, \ldots, n \) and \( \mu = 0 \) to 1. The metric \( d \) measures the sum of the squared–deviations of the observed (\( \tilde{y}_i \)) and predicted (\( \beta_0 + \beta_1 \tilde{x}_i \)) intervals \([\ldots]_\mu \) for all \( \alpha–cuts \) between \( \mu = 0 \) and \( \mu = 1 \). Using fuzzy arithmetic ensures that the coefficients \( \beta_0 \) and \( \beta_1 \) are normal and convex, a requirement of fuzzy numbers. Using this FLR method means that the output of the method is also a fuzzy number, in this case a set of values corresponding to the upper and lower limits of \( \alpha–cuts \) at 0, 0.25, 0.5, 0.75 and 1. These five levels were selected to give a full spectrum of possible values of the fuzzy number.
5.2.4.1 Creating fuzzy numbers and risk analysis

A number of different methods exist to create fuzzy numbers from observed data; these methods are known as probability–to–possibility transformations. A recent summary of different conceptual approaches to these transformations is provided in Mauris (2013). For this research, a method by Dubois et al. (2004) is adapted and implemented to convert the sub–daily DO concentration to daily DO fuzzy numbers. This method was designed to convert a uni–modal discrete probability distribution to a triangular fuzzy set. It was adapted by Khan & Valeo (2015a) to create non–linear discrete fuzzy numbers.

The basic premise of this transformation is to convert a non–specific probability distribution (based on the observations) to a membership function where the modal value has a \( \mu = 1 \), and the support (the limits of the \( \alpha \)–cut interval) is calculated using an uncertainty value \( e \). The value of \( e \) for the present research was selected as 10% representing a cumulative uncertainty of the measurement and scale effects (Struve & Zhou, 2011). The values of the fuzzy number at other membership levels are calculated using a relationship between the probability and possibility that matches the area under the pdf to the highest membership level.

The inverse of a probability–possibility transformation (sometimes referred to as “defuzzification” can be used to create the probabilistic versions of fuzzy numbers. This is useful when model outputs (which are fuzzy numbers) need to be converted back to probability based data - typically for communicating fuzzy results to water resource managers. For example, a fuzzy model output might state that the future value of DO will likely be within the interval \([4.5 \, 10]_{\mu=0}\). A decision maker however, would like to know the probability \( P(X < 5 \text{ mg/L}) \), i.e. the probability that the predicted DO will be below some limit (5 mg/L in this case). Inverse transformations have been proposed by Oussalah (2000) and Dubois et al. (2004), but produce a point estimate of probability for any given value. However, in most environmental informatics applications, we are interested in the cumulative probability rather than probability of a single discrete event. Thus, an inverse transformation to calculate the cumulative probability was developed specifically for non–symmetric fuzzy numbers for discrete systems.

For any \( x \) in \( X \) in the support of a fuzzy number \([a, b]\), we have the corresponding membership level \( \mu(x) \) and the paired values \( x' \) which also shares the membership level.
The value $\mu(x)$ is the sum of the cumulative probability distribution between $[a, x]$ and $[x', b]$, labelled $A_L$ and $A_R$, respectively:

$$\mu(x) = A_L + A_R$$  \hspace{1cm} \text{Equation 5-16}

where $A_L$ represents the cumulative probability between $a$ and $x$ which is equal to the probability $P(x < X)$. Given the fact that the fuzzy number is not symmetrical, we use the lengths of the two intervals $[a, x]$ and $[x', b]$ to establish a relationship between $A_L$ and $A_R$. We can then estimate $A_L$ as:

$$A_L = P(x < X) = \frac{\mu(x)}{1 + \frac{b - x'}{x - a}}$$  \hspace{1cm} \text{Equation 5-17}

Thus, this predicted value $P(x < X)$ can be used by water resource managers to determine if there is a sufficient risk of low DO concentration and can take necessary steps to prevent it, if possible. For this research, a low DO “warning level” was established as 6.5 mg/L, which represents the lowest acceptable DO concentrations for the protection of aquatic life (for non–early life stages) in cold water ecosystems (Canadian Council of Ministers of the Environment, 1999). In this chapter, this inverse transformation was used to calculate the risk of low DO (below the warning level) if there was a possibility (i.e. when the fuzzy prediction was below the warning level, at any $\mu$) of low DO.

### 5.2.5 Quantifying model performance

Three common model evaluation metrics were used to analyse and compare the two regression techniques. The first is the Nash–Sutcliffe efficiency (NSE), arguably the most widely used performance metric in hydrology and environmental informatics; secondly the RSR (defined by Moriasi et al., 2007); and the percentage bias, (PBIAS). These metrics were calculated, respectively, as follows:
$$NSE = 1 - \frac{\sum_l(y_i - \hat{y}_i)}{\sum_l(y_i - \bar{y})}$$ \hspace{0.5cm} \text{Equation 5-18}$$

$$RSR = \sqrt{\frac{\sum_l(y_i - \hat{y}_i)^2}{\sum_l(y_i - \bar{y})^2}}$$ \hspace{0.5cm} \text{Equation 5-19}$$

$$PBIAS = \frac{\sum_l(y_i - \hat{y}_i)}{\sum_l y_i} \times 100\%$$ \hspace{0.5cm} \text{Equation 5-20}$$

where $y_i$ are the observed values and $\hat{y}_i$ are the predicted values. Each metric was calculated at each membership level for the FLR results, but only calculated for the mean values of the sampled Bayesian results. This means that the calculated metrics for the Bayesian results will have the best values compared to values calculated using the minima or maxima sampled values. It should be noted that these metrics are only a rough estimate for the fuzzy numbers. The fuzzy numbers require a fuzzy based performance metric to evaluate the actual efficacy of the model – but this is beyond the scope of the current research.

In addition to the aforementioned metrics, another metric, the mean squared error (MSE), was calculated and is defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ \hspace{0.5cm} \text{Equation 5-21}$$

The MSE was calculated between the observed minimum DO for a particular day, and the minimum value of the sampled Bayesian results or the fuzzy interval (i.e. when $\mu = 0$). This is done specifically to assess the efficacy of each model to predict extreme values.

5.3 Chapter results and discussion

5.3.1 Model performance comparison

Both the BLR and FLR models were sequentially calibrated using data from 2004 to 2013 (i.e. models M01 to M09) at four different lags ($d = 1, 2, 3$ and 7 days). Figure 5-2a shows the evolution of the regression coefficients calculated using the BLR method, as more data is added into the system for the 1 day lag case. The value of $\sigma^2$ (i.e. $1/\tau$)
decreases with each subsequent year of added data, while $\beta_0$ increases and $\beta_1$ remains relatively unchanged. Figure 5-2b, c and d show the evolution of the approximate density function of the regression coefficients and variance of the Normal model. As more data is added, the spread of each parameter decreases: this clearly demonstrates the functionality of a Bayesian approach to data analysis. The uncertainty in parameter estimates is decreasing as more data is added to the system. This figure illustrates the utility of using a Bayesian method: the non-informative prior used influences the estimates of the coefficients (e.g. M02) since the variance is high.

Figure 5-2 shows the change in mean values of the two coefficients and the variance; the bottom figure shows the change in the approximate pdf of the variables for the M02, M04 and M09 cases.

Figure 5-3 shows the fuzzy regression coefficients calculated using FLR. The results are similar to the BLR coefficient results: $\beta_0$ increases with more data (Figure 5-3a), while $\beta_1$ decreases (Figure 5-3b). However, a notable difference between the two is that in the fuzzy case, the spread of the regression coefficient increases as data is added into
the system, whereas in the BLR case, the variance of the coefficients reduces. For example, the range of $\beta_0$ for the FLR M09 case is between 0 and 3, and is between 0.15 and 0.4 for the BLR case. This illustrates that as more data is included in the BLR model the coefficients tend to get more "precise", whereas the FLR coefficients increase in width to accommodate the full extent of variability seen in the data used to calibrate the model. The impact of this major structural difference between the two modelling approaches is discussed below.

Figure 5-4 shows selected results for the validation data for the 1 day lag case, for model M01 (2005 data used for validation), M04 (2008 data used for validation) and M09 (2013 data used for validation). Each subplot shows the trend of the observed daily mean DO, and the daily minimum and maximum concentration. In addition to this, the figures shows the minimum and maximum sampled data for the BLR model results, and the fuzzy interval at $\mu = 0$ for the FLR model results. For both cases, the mean observed DO generally falls within the predicted intervals for each case. The figure shows that both models can capture the observed trend with only one year's worth of data. This has an important implication for numerical modelling, namely, that using data–driven methods means that a physical system can be adequately characterised using only one year’s worth of data.

Another notable difference between the two methods is that while the variance of the BLR predictions is generally constant, the fuzzy interval increases and decreases along with changes to the observed variance. For example, starting in June 2005 for M01, the observed minimum and maximum interval decrease and the fuzzy interval reduces to match this change, but the BLR interval is not as flexible. In M04, the observed variability increases at the end of June 2004 and the fuzzy interval increases to match this change, whereas the BLR interval stays more–or–less constant throughout this change. Similar trends can be seen for M09, especially at the end of the season when the observed variance increases in October 2013, with the fuzzy intervals matching this change.
Figure 5-3 Evolution of the FLR coefficients for the 1–day lag and resolution = 96 case, shown here for the M02, M04 and M09 cases.
Figure 5-4 Trend plots for validation results for M01, M04 and M09 for the 1 day lag, resolution = 96 case
The impact of selecting a non-informative prior for M01 and independent priors for M04 and M09 is also demonstrated in Figure 5-4. In M01, generally speaking, the size of the BLR interval is larger, reflecting the relative "lack of knowledge" after implementing only one year of data. This interval decreases as more data is added, as seen for the M04 and M09 cases. Similar results were seen for the models constructed at other lags – results for these models are summarised via the performance metrics discussed later.

Figure 5-5 shows a selection of observed versus predicted DO results for the three models discussed above for the 1 day lag case. These figures highlight that in general, the fuzzy predictions (represented by the grey boxes) overlap the 1:1 line for more cases than the BLR intervals (plotted as black lines) (see Figure 5-5a –c). This figure also shows a plot of lagged DO versus DO for the observed, BLR and FLR predicted datasets (see Figure 5-5d –f). These figures again highlight clearly that unlike the BLR results, the FLR results (shown as intervals) increase or decrease to reflect the change in observed variability, and are thus, better at capturing both the lower or upper daily DO values, along with the daily mean DO. The BLR results meanwhile show a narrowing of the predicted intervals as the size of the dataset increases. In addition to this, unlike the typical OLR case, the FLR intervals are independent of the magnitude of dependent variable.
Figure 5-5 (top) a comparison of observed vs. predicted DO for (a) M01, (b) M04 and (c) M09; the black circles and lines represent the Bayesian predictions, the grey boxes represent the fuzzy predictions at $\mu = 0$; and (bottom) a comparison of observed data (dots), Bayesian (black lines) and fuzzy predictions (triangle at $\mu = 0$) for (d) M01, (e) M04 and (f) M09.
The error metrics for each model are summarised in Figure 5-6 to assess model performance. Figure 5-6a, d, and g show the NSE, RSR and PBIAS results for all models, M01 – M09 for the 1 day lag and 96 point resolution case. These metrics were only calculated at the mean value in the BLR intervals (where the performance was highest and decreased significantly at the extreme values of the intervals), and at each of the five membership levels considered for the FLR case. In general, the performance is high for each of the nine models, using both methods, with NSE values above 0.65, RSR values below 0.6 and PBIAS below 5%. Using the rating system developed by Moriasi et al. (2007), these values fall under “good”, “good” and “very good”, respectively. As more data is added, the performance improves slightly (M05 onwards). In all cases, the BLR metric falls within the fuzzy interval, meaning comparable efficacy with respect to these metrics.

Similar trends are seen as the lag is increased (only shown for the M09 96 point resolution case, see Figure 5-6b, e, and h). The NSE is “good” for the first 3 lags, but falls to “satisfactory” (fuzzy) and “non–satisfactory” (Bayesian) for the 7 day lag case. The RSR slightly increases (i.e. lower performance) as the lag is increased from ~0.5 (“good”) to ~0.6 (“satisfactory”); with the mean value of the Bayes results performing slightly better than the fuzzy results. The change in PBIAS for the M09 case was not significant; with performance classified as “very good” (< 10%) for all cases.
Figure 5-6 A summary of the performance metrics for each model; as more data is added (a, d and g); as the lag is increased (b, e and h); as the resolution is reduced (c, f and i) for the 1 day lag case; the black markers are the mean Bayesian results, and the lines are the limits of the fuzzy results.

Lastly, Figure 5-6c, f and i shows the results of the final evaluation of both BLR and FLR methods: the amount of data used to construct each model was decreased from 96 samples per day (i.e. sampling every 15 minutes), to 24 samples per day, to finally 6 samples per day (or every 4 hours). The results are shown only for the M09 1 day lag case. The primary objective of this test was to see the impact of input data resolution on performance for both methods. The NSE does not show a marked change as the resolution is reduced for either method, with a value at 0.75 (“very good”). The change in RSR is also minimal: it stays at 0.5 “good” for the mean Bayes results and at 0.65 (“satisfactory”) for the best fuzzy results. However, a decrease in the fuzzy interval is
seen for the RSR results, meaning the variance in results decreases: a result of using less data. The value of PBIAS does not changed markedly as the resolution is decreased, generally less than 1% (“very good”) for both methods. The PBIAS interval for the fuzzy results decreases as the resolution is decreased, again reflecting lower variance with the decreased data set. The results in Figure 5-6 are only shown for one case – results for all other model cases follow the same pattern and are included in Appendix E.

The significance of the results pertaining to the data resolution is the limited effectiveness or utility in increasing sampling rates for real-time, water quality monitoring stations for water quality prediction. Similar results are seen for the 96 samples per day and 6 samples per day case. This suggests that it might be more prudent to sample a number of locations along a river, rather than sampling at high resolutions at one location. The difference between the two sampling rates means that for the same computer storage capacity, 16 locations can be sampled at 6 samples per day rather than one location at 96 samples per day.

In terms of changes to the lag time, while performance decreased slightly as lag time was increased, both modelling approaches were still able to provide meaningful predictions of daily DO. This suggests, that an approach that uses multiple lags as predictors can be used by water resource managers in Calgary, to determine the risk of low DO up to a week in advance. Both approaches provide similar performance levels with respect to these metrics; the important differences between the two are discussed in the next section.

5.3.2 Low DO analysis

One of the reason for predicting daily DO in the Bow River is to focus on the risk of low DO. For the given dataset, days where observed DO was measured to be below 6.5 mg/L (used as a reference for low DO in this chapter) were isolated; a total of 142 occurrences of low DO were recorded between 2005 and 2013 (with the most occurrences in 2006, which had 67 of these days). Using only the validation results from the 1 day lag, 96 resolution model, the ability of each model to capture the observed minima within the predicted intervals was calculated.

For 80 out of the 142 cases (about 56%), the minima fell within both the Bayesian (minimum and maximum sampled values) and fuzzy (at $\mu = 0$) intervals. This is
illustrated in Figure 5-7a, where the observed minimum DO is plotted along with the lower bounds of the predicted intervals. For these cases, the MSE between the lower limit of the predicted interval and the observed minimum was calculated. The MSE for the fuzzy case was much lower (0.55 vs. 8.75) than the Bayesian MSE. This means that the FLR method is better at predicting low DO. In addition to this, for 54 out of the 142 cases (~38%), the observed minimum fell within the fuzzy interval but outside the Bayesian interval, i.e. the Bayesian case over-predicted the minimum. For these cases (as seen in Figure 5-7b) the MSE for the fuzzy case was 0.32 compared to an MSE of 23.28 for the Bayesian case. This shows that for this particular river, the FLR method can capture more of the low DO events than the BLR case, and predicts them with higher accuracy.

There were also three cases where both methods over-predicted the observed minimum (see Figure 5-7c); however, even in this case, the fuzzy results were much closer to the observed minima than the BLR case (as reflected by the lower MSE of 5.02 versus 36.53 for the Bayesian case). Lastly, for four cases the Bayesian interval captured the low DO within its interval, while the fuzzy interval over-predicted the minimum value. However, even for these cases where the fuzzy interval is unable to capture the low DO, the MSE value is still much lower, about four times compared to the Bayesian case (see Figure 5-7d): 1.12 for FLR compared to 4.52 for BLR.
Figure 5-7 Observed minimum DO compared to the lower limits of the predicted Bayesian and fuzzy (at $\mu = 0$) intervals: (a) when both methods capture the minimum value, (b) when only the fuzzy interval captures the minimum; (c) when neither method captures the minimum; (d) when only the Bayesian interval captures the minimum.
Thus, in all four possible scenarios, the FLR models are more accurate (i.e. much lower MSE) compared to the BLR models. The FLR is also able to capture more of the low DO events, 134 out of 142 events, whereas the BLR only captured 84.

The ability of the fuzzy method to capture the observed minimum within its prediction intervals can be further highlighted using an inverse transformation. This transformation was used to calculate the probability $P(\bar{DO} < \min(DO_{obs}))$, i.e. the probability that the FLR predicted DO will have a possibility of being lower than the observed minimum for that day. This probability highlights the conservative nature of the fuzzy predictions in contrast to the Bayesian results which indicate zero probability of DO being lower than the observed minimum. For these 54 cases, the average probability is 4.06% that the predicted DO will be below the observed, whereas the BLR results indicate that the predicted DO will be exclusively above the observed minimum. Figure 5-8 shows three examples of this inverse transformation for days in 2005 and 2006. These figures show that the Bayesian interval does not capture some of the low DO observations, whereas the fuzzy number results fully encompasses the observations, and in addition to this provides estimates of being lower than the observed values (between 7.8 and 11% for these three cases).

![Figure 5-8 Sample results of inverse transformation showing the probability of the fuzzy method to predict the minimum observed DO.](image)

The grey area represents the interval of observed DO values, the dashed lines show the upper and lower limits of Bayesian predictions, and the solid black line with dots shows the fuzzy number
The same inverse transformation along with the predicted fuzzy numbers can be used to highlight the generic risk of low DO. For example, if a “warning level” is set at 6.5 mg/L, then the inverse transformation can determine the probability: $P(\bar{DO} < 6.5 \text{ mg/L})$. This probability can be used by water resource managers to initiate risk management systems. For example, if the numerical model predicts a 15% chance of low DO, with a one–day lead, the water resource manager can start to take precautionary steps to limit this possibility. This inverse transformation was conducted on a subset of the 142 cases where low DO occurred: days where the Bayesian model predicted DO to be above 6.5 mg/L and where the fuzzy model predicted DO to be below 6.5 mg/L, a total of 24 cases (see Figure 5-9). On average for this subset, the fuzzy predictions showed an 8% chance that the predicted DO on a particular day would be below the warning level. And in comparison, the Bayes method did not predict DO to be below 6.5 mg/L. Thus, apart from providing low DO predictions with a much lower MSE, the fuzzy number method has another advantage that it can provide the probability (risk) of low DO events as part of its predictions. The figure below shows three examples of these cases, where the probability of low DO was between 10 and 13%.

Figure 5-9 Sample results of inverse transformation, showing the probability of the fuzzy method to predict DO to be below 6.5 mg/L. The grey area represents the interval of observed DO values, the dashed lines show the upper and lower limits of Bayesian predictions, and the solid black line shows the 6.5 mg/L warning level, and the black line with dots represent the fuzzy number.
5.4 Chapter summary

In this chapter, an auto-regressive type approach was taken to predict DO concentration in the Bow River in Calgary, Canada. Two regression models, a Bayesian and a fuzzy linear regression model were constructed for a number of different scenarios (four different lags, three different input resolutions). In addition to this, the modelling approach adopted a recursive algorithm to mimic a real-time prediction model. In general, the results show that both approaches provide “good” or “very good” predictions of daily DO, showing the utility of using data-driven modelling in a complex urban environment. In fact, only using one year of data to calibrate the model (M01) showed very good validation results, thus, illustrating that the data-driven approach can represent the physical system with even a relatively small dataset. This is a useful approach for water resource managers in many jurisdictions, given the simplicity of the model and ease of collecting real-time data, and can be easily extended to other water quality parameters.

The use of different lag times to predict DO showed that very good short term predictions are possible using both regression approaches, though performance decreases as the lag is increased to seven days. The influence of changing the input resolution showed that the performance of either model did not change significantly. This suggests that high resolutions (96 samples per day) are not necessary to calibrate the system when equivalent results can be achieved at lower resolutions (6 samples per day).

The utility of the fuzzy method was demonstrated by analysing the ability of each model to predict low DO events. The fuzzy method captured more low DO events within its predicted interval compared to the Bayesian method, and did so with much lower MSE values. Inverse transformations to convert fuzzy predictions to probability based method showed that the fuzzy method can predict risk of low DO when the Bayesian method could not, thus, showing a useful application for water resource managers.
6. Fuzzy linear regression for flood prediction and risk assessment

6.1 Chapter introduction

The June 2013 floods in southern Alberta were one of the worst natural disasters to occur in Canada. The floods were responsible for four deaths, displaced more than 100,000 Albertans in over 30 communities (Alberta Government, 2014), and caused approximately $6 billion in damage (Environment Canada, 2013). Also, the floods contributed to the transport of large amounts of sediment and the destruction of river banks, channels and aquatic ecosystems (Environment Canada, 2013). This event highlighted the importance and necessity of better flood protection, effective and timely flood mitigation strategies, including improved flood prediction.

The mechanisms behind extreme events in southern Alberta are generally understood and documented (Valeo et al., 2007; Ardell, 2013a); however, predicting and assessing the risk of floods in the future remains a challenge, in part due to the uncertainty in the numerical models that are used. Flood prediction using physically-based, deterministic hydrologic models (such as a rainfall-runoff routing model) rely on simplified, conceptual representation of highly complex, correlated and spatially distributed processes that occur in a watershed (Wijiesekera et al., 2012; Vrugt et al., 2005; Cox, 2003). This leads to considerable uncertainty. Furthermore, uncertainty in the data used to calibrate model parameters and errors in model structure, compounds this issue (Vrugt et al., 2005).

An alternative to physically-based models is to use data-driven models, which are based on generalised relationships, links or connections between input and output datasets (Solomantine & Ostfeld, 2008). The models can characterise a system with limited assumptions and have similar, if not better performance than physically-based models. In addition, a simpler model structure means that the propagation of uncertainty from different sources is easier to assess. The use of data-driven models, such as neural networks, statistical methods or regression-based techniques, has been widespread in hydrology, particularly for short term daily flow rate forecasts, using a variety of input

---

9 Material from this chapter was accepted for publication in the Journal of Environmental Informatics.
variables (Garen, 1992; Zealand et al., 1999; Campolo et al., 1999; Schilling & Wolter, 2005; Adamowski & Sun, 2010; Duncan et al., 2011; Li et al., 2015). A recent regression based study predicted flow in the Bow River in Calgary, using a base difference regression model (Veiga et al., 2014). Two-day lead times were used to predict daily flow rates at one station, using data from upstream locations, with promising results. Another advantage of using a data-driven approach is that data collected from on-going monitoring systems, e.g. real-time flow rate data that is routinely collected by Environment Canada (Environment Canada, 2014) can be used to calibrate and validate the model, rather than site specific surveys that are required for many physically-based methods.

Data-driven modelling have intrinsic uncertainties associated with it that are not random or probabilistic in nature, thus, making it well suited for the use of fuzzy number theory (Dubois & Prade, 1997; Ozbek & Pinder, 2006). Fuzzy numbers use fuzzy sets (Zadeh, 1965) and possibility theory to describe uncertain or imprecise quantities, measurements or observations (Khan and Valeo, 2015; Huang, et al., 2010; Zhang & Achari, 2010a). They are more suitable when data is missing, incomplete or vague, combined from multiple sites, and to represent uncertainty that is not purely of a random nature (Bárdossy, 1990; Bárdossy et al., 1990; Bogardi & Duckstein, 2003; Guyonnet et al., 2003; Zhang & Achari, 2010a; 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, and the risk of certain events can be directly estimated from the fuzzy number itself (Khan et al., 2013; Khan & Valeo, 2015a). Fuzzy numbers are based on possibility theory, which suggests that some of the strict assumptions in many probability models can be relaxed, thus, making it useful for hydrological systems (Peters, 1994; Kim et al., 1996; Kahraman et al., 2006). The consistency principle (Zadeh, 1978; Dubois et al., 1993) links possibility and probability, implying that something has to be possible before it is probable. A number of methods have been created that utilise this principle to convert observed data into fuzzy numbers (Oussalah, 2000).

Fuzzy numbers have been widely used in hydrology to represent uncertainty in the parameters of numerical models (Khan and Valeo, 2015a; Khan et al., 2013). The
literature demonstrates the utility and advantage of using fuzzy numbers and a summary of some of these applications can be found in Khan and Valeo (2015). Fuzzy numbers have also been used in conjunction with data-driven models to predict flow rate in rivers. Alvisi & Franchini (2011) developed a method to use fuzzy numbers for the weights in an artificial neural network algorithm to predict water level and discharge under uncertainty. Corani & Guariso (2005) proposed an artificial neural network and fuzzy logic based method for river flood predictions that used a weighted least squares training algorithm, rather than the use of fuzzy numbers. Nguyen & Chua (2012) used a recursive adaptive network-based fuzzy inference system for real-time flood forecasts, with lead times between 1 to 5 days. Results from this model were better compared to a physically-based model used for comparison. Also, Wang & Huang (2013b) used a two-stage mixed-integer linear programming approach to develop a model for flood-diversion planning. They represented uncertainties in the system using a fuzzy based approach. Similarly, Wang et al. (2014) used a possibility-probability hybrid approach to represent multiple uncertainties in a flood management system. In addition to this, Ahmad & Simonovic (2011, 2013, and 2015) use a 3D fuzzy set approach to assess spatial and temporal variability of urban flood damage and risk assessment. Abdalla et al. (2014) used a fuzzy based method to quantify the uncertainty in flood risk assessment. Lohani et al. (2014) proposed a Takagi-Sugeno fuzzy inference system that uses cluster analysis, to predict floods using hourly data (rainfall and flow rate). These type of fuzzy rule based analysis has been quite successful for flood forecasting (for e.g. Zhang et al., 2012), but conceptual differ from a fuzzy number based approach. As described above, a fuzzy number based method is when the uncertainty in a parameter is represented through a specific type of fuzzy set based on possibility theory.

While the benefits of a fuzzy number based approach for flow rate prediction have been investigated, the use of fuzzy linear regression (FLR) to predict peak flow rate has not been explored. FLR is a method used to extend simple linear regression for applications involving fuzzy numbers, i.e. for uncertain or imprecise systems (Khan & Valeo, 2015a). This property makes them ideally suited for data-driven techniques that model environmental systems. It provides an alternative method when simple linear
regression may not be possible, e.g. when assumptions of simple linear regression are not met (including when the assumption of linearity is invalid), or if there is obvious fuzziness or uncertainty in the underlying data or process coefficients (Savic and Pedrycz, 1991; Chang & Ayyub, 2001). 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). In simple linear regression models, the independent variable \(x\) predicts dependent variable \(y\) and in ordinary least squares regression, it is assumed that \(x\) are observed without error (i.e. they are fixed). This assumption is typically only true when the independent variables are controlled and the effect on the dependent variable is measured. Often in engineering applications, the dependent data are observed with an error, or bias, or are random rather than fixed (Maddala, 1988). This introduces a bias in the model coefficients such that the least squares estimates of the regression coefficients will be underestimated (Fuller, 1987; Maddala 1988). Unlike a traditional simple linear regression model, FLR allows uncertainty in the input parameters to be included in the analysis. The benefit of using a fuzzy number based approach with a data-driven method (FLR) to predict flow rate is that real-time data, which is often routinely collected, can be used to accurately predict peak flow rate. The uncertainty in the system and in the data can be collectively represented in a fuzzy number. A risk analysis can then be directly conducted using fuzzy numbers to assess the risk of flow magnitudes beyond a given threshold.

6.1.1 Chapter objectives

The objectives of this research is to improve peak flow rate prediction, and flood risk assessment, in southern Alberta, using a data-driven method in order to provide water resource managers sufficient lead time (e.g. between one and seven days) to implement flood defence systems. Data from three sites on two rivers (the Bow River and Elbow River) which flow through Calgary, Alberta are used. Currently, the Government of Alberta provides information to municipalities, including Calgary, which are at risk of flooding. This includes real-time and daily flow forecasting (Walford, 2014), which is required, for example, by the City of Calgary to empty the Glenmore Reservoir (on the Elbow River) in the event of a risk of a large flood (City of Calgary, 2014). A number of
factors impact the accuracy of these predictions during extreme events, including the use of data from hundreds of different sites, lack of knowledge of the precise magnitude and location of upstream precipitation, the destruction of upstream monitoring stations, and uncertainty with meteorological forecasts (Ardell, 2013a; Ardell, 2013b; Walford, 2014).

Thus, there is a need for improved real-time flow rate predictions, particularly with respect to the timing and magnitude of the peak flow rate, to help mitigate against the impacts of the floods, as experienced in 2013. An FLR method (developed in in Khan and Valeo, 2015) is proposed that uses real-time data to predict daily peak flow rate using lagged daily flow rate. This approach retains the simplicity and benefits of simple linear regression, but with the ability to incorporate uncertainty that improves risk assessment and the prediction of extreme floods. This FLR method is unique because fuzzy number inputs, outputs and regression coefficients are used, whereas other fuzzy regression techniques only use fuzzy representations for some (not all) of these parameters (Khan and Valeo, 2015). In addition, this method uses probability-possibility transformation to define fuzzy numbers with uniquely non-linear membership functions; this is much more suitable for analysis of flow rate.

The results from this analysis are compared to observations, as well as the results from an error-in-variables (EIV) regression model. Both these methods attempt to describe an uncertain system with an error in the observations (Bárdossy et al., 1990) using a relatively simple data-driven model, with the key difference that one method is based on probability theory (EIV) while the other is based on possibility theory (FLR).

6.2 Methods\textsuperscript{10}

6.2.1 Site description and data collection

The Bow River basin, located in southern Alberta, has an area measuring approximately 25,123 km$^2$ and provides significant economic activity in the form of supplying drinking water for a major urban centre (the City of Calgary), and irrigation water for several irrigation districts in the Prairie region in Alberta (BRBC, 2010a). For this research, three monitoring sites in the Bow River basin were selected for peak flow rate prediction and flood risk assessment. These sites are: “Bow River at Banff”

\textsuperscript{10}The MATLAB code associated with this section is included in Appendix J.
(Environment Canada WSC Station ID: 05BB001), “Bow River at Calgary” (ID: 05BH004), and “Elbow River at Bragg Creek” (ID: 05BJ004). The first two sites are located on the Bow River, whereas the last is located on the Elbow River (a tributary that merges with the Bow River in downtown Calgary); see Figure 6-1 below for the relative location of each site in Alberta.

![Figure 6-1 The locations of the three sites used in this chapter: (from L to R) Bow River at Banff, Bragg Creek at Elbow River, and Bow River at Calgary](image)

The Bow River originates in Bow Lake, located in the Rocky Mountains and flows south-easterly through Banff (drainage area of 2210 km$^2$) and Calgary (drainage area of 7870 km$^2$), meeting the Oldman River, and ultimately draining into Hudson Bay (Robinson et al., 2009; Environment Canada, 2014). The Bow River averages a 0.4% slope over its 645 km length, measuring as high as 7 m/km in the Rocky Mountains (BRBC, 2010a; BRBC, 2010b). The Bow River is supplied by precipitation accumulated in the snowpack in the Rocky Mountains, precipitation and discharge from shallow groundwater. Runoff peaks in the spring (typically in June) while low flows are seen in the winter (January) (BRBC, 2010b). While the River is unregulated at the Banff monitoring station, Bearspaw Dam regulates flow immediately upstream of Calgary’s city limits. The headwaters of the Elbow River are located at Elbow Lake in the Rocky Mountains. The river flows easterly, through the hamlet of Bragg Creek (drainage area of 791 km$^2$), and merges with the Bow River in downtown Calgary. The Elbow River has an average grade of 1% over its 124 km length; the total drainage area of the Elbow River
sub-basin is 1235 km² (Environment Canada, 2014; BRBC, 2010b; BRBC 2010c). The Glenmore Dam in Calgary regulates the downstream flow of the Elbow River through Calgary. These rivers flow through many residential communities and the commercial centre of Calgary. Thus, the importance of peak flow prediction and flood risk assessment of these rivers is extremely important for the safety of the residents in Calgary, and for protection of significant assets in downtown Calgary.

Eleven years of hourly flow rate data for each station was obtained from Environment Canada for the period January 1, 2000 to December 31, 2010. The annual median flow rate at Banff, Calgary and Bragg Creek varied between 11 – 26 m³/s, 47 – 85 m³/s, and 4 – 8 m³/s, respectively for the selected period. The annual peak flow rate at Banff, Calgary and Bragg Creek varied between 142 – 306 m³/s, 172 – 787 m³/s, and 15 – 307 m³/s, respectively. The highest peak flow rate at Banff (306 m³/s) occurred in 2007, and in 2005 for Calgary (787 m³/s) and Bragg Creek (307 m³/s). The peak flow rates in 2005 are associated with a flood event in Calgary and southern Alberta, which estimated to have cost approximately $400 million in damage (Environment Canada, 2005; Valeo et al., 2007). This flood event was the last major flood in southern Alberta prior to the floods in June, 2013. It is important to note that though the drainage area at Bragg Creek is almost three times smaller than the drainage area at Banff, the maximum annual peak flow for these sites is similar, and the median annual flow at Bragg Creek is much lower than the peak flow rate, indicating that this site is prone to flash flooding.

The hourly flow rate collected for each site was filtered by removing dates where shift corrections (usually due to ice conditions) were applied by Environment Canada. Typically for the eleven year period, the data ranged between mid-March to early-November for Banff and Calgary, and from mid-February to late October for Bragg Creek. Any data points within the selected range that had corrections applied were removed from further analysis. In addition to this, a second filter was applied to remove the low flow rate periods, since the primary objective of this research was to predict peak flow rates. To do this, the first and last day where the daily peak flow rate was greater than the ice-free median flow rate were used as the lower and upper limits of the data range, respectively. This reduced the data from an original, unfiltered set of 4018 days of...
hourly flow rate data to 1583 days for Banff, 1860 for Calgary, and 1633 for Bragg Creek, i.e. representing about 40% of available data.

6.2.1.1 Selection of dependent variables

The filtered data was used to calculate the daily peak flow rate \( Q_P \). An initial correlation analysis was conducted using \( Q_P \) and several other variables, including daily mean flow rate \( Q_D \), daily precipitation, lagged \( Q_P \), (computed at different lags between one to seven days), and various combinations of these variables as well. This preliminary analysis showed that while an autoregressive approach (using lagged peak flow rate) to predict \( Q_P \) showed promise, a better overall performing model structure was to use \( Q_D \) at different lags. This was based on the magnitude of the correlation coefficient for each site at different lags. It should be noted that a similar analysis was conducted using upstream data (i.e. at Banff) to predict peak flow rate downstream (i.e. at Calgary), and lower correlations were found for the entire study period using upstream data rather than local data (correlation coefficient of 0.82 vs. 0.95 on average, respectively). Details of this analysis, including different combinations of variables tested as the independent variable, are not discussed here but may be found in Khan (2014) and Khan & Valeo (2014b). Thus, these two variables were used to construct the EIV and FLR regression models to predict \( Q_P \), in the form of:

\[
Q_P(t) = f(Q_D(t-d))
\]

Equation 6-1

where \( Q_P(t) \) is the daily peak flow on day \( t \), \( Q_D(t-d) \) is the mean daily flow on day \( t-d \), and \( d \) is the lag in days, selected as either 1, 2, 3 or 7 days for this research. These lags were specifically selected to give sufficient lead time to operators of the proposed method to enact flood defence systems in a timely manner. The 7 day lag represents the limit of the applicability of the proposed method. Model performance is expected to diminish in proportion to the number of lag days between a lag of 3 and 7 days, thus, analysis for lags of 4, 5 or 6 days is not included.
6.2.2 Error-in-variable linear regression

A simple linear regression model is proposed to predict \( Q_P \) using \( Q_D \) at different lags. However, ordinary least squares regression assumes that the independent variables (in this case \( Q_D \)) are observed without error (i.e. they are fixed). This assumption is only true for controlled experiments, and not when monitoring data is being used to find suitable relationships between variables. The bias introduced by this error results in an underestimation of regression coefficients, even for very large samples (Fuller, 1987; Maddala 1988). Thus, to account for the uncertainty in the calculated \( Q_D \) being used to predict \( Q_P \), an EIV regression model is specified as follows:

\[
Q_P = A + BQ_D
\]

Equation 6-2

where \( Q_P \) is the daily peak flow rate of interest, \( A \) and \( B \) are the regression coefficients to be calculated, and \( Q_D \) is the calculated mean daily flow rate (i.e., the average of the 24 hourly data points observed). The authors propose that this calculated \( Q_D \) is not the true mean daily flow rate, but a value with uncertainty. The true mean daily flow rate \( \Phi_D \) is given by:

\[
\Phi_D = Q_D + u
\]

Equation 6-3

where \( u \) is a random error term, which we assume to be the standard error of the mean, estimated as:

\[
u = \pm T \frac{\sigma}{\sqrt{n}}\]

Equation 6-4

where \( T \) is the \( t \)-statistic for 23 degrees of freedom, \( \sigma \) the standard deviation of the observations of hourly flow rate and \( n \) (which equals 24) is the number of data points. Therefore, for each calculated \( Q_D \) (mean of observations), we also calculate the daily standard deviation (\( \sigma \)) and then calculate the daily error (\( u \)). By taking this uncertainty in
the calculated mean daily flow rate, an unbiased estimate of the slope coefficient $B$ can be calculated by:

$$B = \frac{COV(Q_D, Q_P)}{VAR(Q_D) - VAR(u)}$$  \hspace{1cm} \text{Equation 6-5}

The constant $A$ can be estimated from $(Q_P - BQ_D)$. Once the regression coefficients have been calculated, the unbiased estimate of the variance of the predictions, and the confidence interval of the predictions can be calculated using the standard procedure (Wittink, 1988). The derivation of this EIV regression method is widely available in many regression or econometric textbooks with good descriptions in Fuller (1987) and Maddala (1988). It is worth noting here that the similarity of EIV and FLR is that they both attempt to consider error in the observations. The differences between the two is that the basic assumptions of each method are completely distinct: the error in EIV is assumed to be the same for all values, whereas in the fuzzy case each value will have its own membership function), and the data requirements for the EIV are higher (Bárdossy et al., 1990).

Typically, after a regression analysis has been conducted, an analysis of residuals must be conducted to ensure that the initial assumptions of the model have not been violated. The four principal assumptions of linear regression are: (i) the mean of residuals $r$ is zero, $E(r) = 0$; (ii) the residuals have a constant variance (with respect to time, and the data), $\text{var}(r) = \sigma^2$; (iii) the residuals are independent or uncorrelated, $cov(r_1, r_2) = 0$; and (iv) the residuals are normally distributed, $r \sim N(0, \sigma^2)$. Of these assumptions, the third assumption (autocorrelation of the residuals) is most important for time series data (such as the dataset considered in this chapter). A violation of this assumption, however, is not critical if the first assumption holds. Then, the presence of significant autocorrelation suggests that though the regression coefficient estimates are still unbiased, they are no longer the most efficient estimates, in other words they are not minimum variance estimates (Wittink, 1988; Montgomery et al., 2006). This means that the predicted variance of the coefficients is lower than the actual variance, the standard error is artificially low, and that the coefficient of determination is artificially higher. Thus, the
presence of significant autocorrelation will show a stronger relationship between the dependent and independent variables than what truly exists, and the use of confidence intervals and hypotheses testing may not be appropriate (Montgomery et al., 2006).

### 6.2.3 Fuzzy linear regression

The objective of the FLR method is similar to the ordinary least squares approach to linear regression, however, instead of minimising the residual between an observed and regressed value, the distance between two fuzzy numbers is minimised 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 \( i = 1, 2, \ldots, n \) a regression 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 optimization problem:

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

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

Using the FLR method means that the output of the method is also a fuzzy number, in this case a set of values corresponding to the upper and lower limits of \( \alpha \)-cuts at 0, 0.25, 0.5 0.75 and 1. These five levels were selected to give a full spectrum of possible values of the fuzzy number. The predicted membership value can be used to calculate the exceedance probability of a value within the fuzzy set using a possibility to probability transformation. This is useful since probabilities are more readily understood by water...
resource managers and the general public. A description of a probability to possibility
transformation and its inverse are described below.

6.2.3.1 Probability to possibility transformations
A number of different methods exist to create fuzzy numbers from observed data;
these methods are known as probability-to-possibility transformations. A recent summary
of different conceptual approaches to these transformations is provided in Mauris (2013).
For this research a method by Dubois et al. (2004) is adapted and implemented to convert
the sub-daily flow rate observations to $Q_P$ and $Q_D$ into fuzzy numbers. This method was
designed to convert a uni-modal discrete probability distribution to a triangular fuzzy set.
It was adapted by Khan & Valeo (2014b) to create non-linear discrete fuzzy numbers.
The basic premise of this transformation is to convert a non-specific probability
distribution to a membership function where the modal value has a $\mu = 1$, and the support
(the limits of the $\alpha$-cut interval) is calculated using an uncertainty value $e$ (details of this
value are discussed below). The values of the fuzzy number at other membership levels
are calculated using a relationship between the probability and possibility that matches
the area under the probability density function (pdf) to the highest membership level.

To convert the observed $Q_D$ and $Q_P$ to fuzzy numbers, two different approaches were
taken. For $Q_D$, a non-specific probability distribution was converted to a fuzzy number
using the method described above. Hourly observations (24 points), the modal value
(where the membership level was equal to one) and a support calculated using an
uncertainty value $e$ (details of this are given below) were used. For $Q_P$, a specific pdf
(assumed to be a Normal distribution, with the mean value as the observed peak flow rate
for a given data, and a variance calculated using the error value $e$) was used to construct a
fuzzy number using the method outlined above. This represents a hybrid modelling
approach, where a probabilistic parameter ($Q_P$) is used with a possibilistic parameter ($Q_D$)
in one regression model.

6.2.3.2 Significance of the error value $e$
The value of $e$ used for these transformations represents the uncertainty associated with
the measured flow rate. Typically, analysts assume that the in situ measured data from
river flow rate measurements (such as that provided by Environment Canada) are within
±5% of the true value at the 95% confidence interval (Hamilton & Moore, 2012; Papa et al., 2012). Others consider that this random uncertainty associated with the measurement of the flow rate to be negligible (Di Baldassarre & Montanari, 2009) or as low as 1% of the true value at the 95% confidence interval (Shrestha & Simonovic, 2010). However, if other components of uncertainty associated with determining flow rate are included (e.g. stage-discharge relationship, shape of river bed, etc.), this accuracy declines significantly. Many in the research community advocate for uncertainty levels between 15 and 20% of the true value (Papa et al., 2012), with 15% considered as “optimistic” (Di Baldassare, 2012). McMillan et al. (2012) provide a list of typical quantitative results of combined flow rate uncertainty. In one studies this uncertainty is listed as high as 100% for low flows, 10% for medium flows, and 20% for high flows (McMillan et al., 2012; Krueger et al., 2010). Daily discharge uncertainty is listed with a range of ±100 – 200% for low flows and ±100% for high flows by Harmel and Smith (2007) and up to 50% by Hamilton and Moore (2012) for all magnitudes. Pappenberger et al. (2006) reported uncertainty with peak flow rates to range between 8 and 25%, Di Baldassare and Montanari (2009) reported a range from 6.2% to 42.8% at the 95% confidence interval, and Westerberg et al. (2011) give a range between -43 to 73%.

The literature shows a very wide range of error values, with little consensus on the optimal value. Thus, three different error values, $e$, were selected for this research: 5%, 10% and 20% and the analysis was conducted using each value. These values represent a conservative estimate of flow rate uncertainty, which have been selected to demonstrate the application of a fuzzy number based method and for comparison with a probabilistic method; the authors recognise that the value of $e$ is not limited to only these values (i.e. higher $e$ values can be used). The 5% represents the case where only random uncertainty is considered, whereas the 20% represents the case for high flow rates, which are of primary interest for this research. In another application using fuzzy number analysis to quantify and characterise peak flow rate, Shrestha and Simonovic (2010) used an error value of 50% to represent the support of the fuzzy number: thus, the selected values of $e$ in this chapter are within a conservative range. Figure 6-2 below shows examples of the $Q_D$ and $Q_P$ transformations for an $e$ value of 20%.
Figure 6-2 A sample of results of transforming observed hourly flow rate to fuzzy mean daily flow rate, and fuzzy mean peak flow rate, for July 2, 2005. An $e$ value of 20% was used for these conversions.
6.2.3.3 Possibility to probability transformations

Defuzzification, or transforming the possibility distribution of a fuzzy number to a probability distribution, is important when an estimate is needed on whether something is “probable” rather than “possible”. For example, while fuzzy regression might give a prediction of future peak flow rate as a fuzzy number, an estimate of the probability associated with the magnitude of peak flow rate is needed, often to communicate the risk of a flood. For example, a fuzzy number might predict that the peak flow rate to be 500 m$^3$/s with a membership level equal to 0.75, but a decision maker might need to know how likely a flow rate of that magnitude is, before making a decision. For this case, an inverse transformation can be used to calculate the point estimate of probability for any given value (based on Oussalah, 2000 and Dubois et al., 2004). However, in most hydrological applications, we are interested in exceedance probability $P(x > X)$ rather than probability of a single event $x$. Thus, an inverse transformation to calculate the exceedance probability was developed specifically for non-symmetric fuzzy numbers for discrete systems.

For any $x$ in $X$ in the support of a fuzzy number $[a, b]$, we have the corresponding membership level $\mu(x)$ and the paired value $\mu(x')$ which also shares the membership level. The value $\mu(x)$ is the sum of the cumulative probability distribution between $[a, x']$ and $[x, b]$, labelled $A_L$ and $A_R$, respectively:

$$\mu(x) = A_L + A_R \quad \text{Equation 6-8}$$

where $A_R$ represents the exceedance probability that the predicted value is greater than $x$. Given the fact that the fuzzy number is not symmetrical, we use the lengths of the two intervals $[a, x']$ and $[x, b]$ to establish a relationship between $A_L$ and $A_R$. Using this ratio, we can estimate $A_R$ as:

$$A_R = P(x > X) = \frac{\mu(X)}{\left(1 + \frac{x' - a}{b - x}\right)} \quad \text{Equation 6-9}$$
Thus, this predicted value $P(x > X)$ can be used by water resource managers to determine if there is a serious risk of high peak flow rate in the predicted time period, and if it warrants the implementation of flood defence strategies. For this research a flow rate magnitude corresponding to the 2% flow rate ($Q_{P2\%}$) for the eleven year period (from 2000 to 2010) was calculated for each site. This value, $Q_{P2\%}$, is exceeded 98% of the time in the eleven year period at each site. These values were 173 m$^3$/s at Banff, 305 m$^3$/s at Calgary, and 59 m$^3$/s at Bragg Creek. These values were then used to establish the probability of predicted peak flow rate for a given day to be higher than $Q_{P2\%}$. This procedure can be used by decision makers to create a set of rules relating to flood risk management. For example, if there is more than a 5% chance of the predicted flow to be greater than $Q_{P2\%}$, a flood defence warning is issued.

### 6.2.4 Quantifying model performance

Three metrics are used to quantify and compare the performance of the EIV and FLR model. These are the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970), percent bias (PBIAS), and the ratio of the root mean square error to the observed standard deviation (RSR). If a data set of observations $y_i$ has a mean $\bar{y}$ and corresponding predictions at each point $i$ denoted by $y_{i,\text{pred}}$ then:

\[
\text{NSE} = 1 - \frac{\sum_i (y_i - y_{i,\text{pred}})}{\sum_i (y_i - \bar{y})} \quad \text{Equation 6-10}
\]

\[
\text{PBIAS} = 100 \frac{\sum_i (y_i - y_{i,\text{pred}})}{\sum_i y_i} \quad \text{Equation 6-11}
\]

\[
\text{RSR} = \frac{\sqrt{\sum_i (y_i - y_{i,\text{pred}})^2}}{\sqrt{\sum_i (y_i - \bar{y})^2}} \quad \text{Equation 6-12}
\]

These metrics were calculated for each crisp and fuzzy simulation. For the fuzzy data, the metrics were calculated at each membership level ($\mu = 0, 0.25, 0.50, 0.75$ and 1). However, this meant that for each metric, there were five intervals of metric values for each fuzzy number (corresponding to each membership level), making it difficult to compare results between the crisp and fuzzy results. Thus, a rating system developed by
Moriaisi et al. (2007) was extended for use with fuzzy numbers. We assign a numerical value to each of the qualitative ranks provided by Moriasi et al. (2007), as listed in Table 6-1. For each fuzzy number, each metric was calculated at each membership level, and assigned a rating value according to Table 6-1. The overall metric for each fuzzy number was then the average of the five (corresponding to each μ). Furthermore, this rating was assigned to each of the calculated metrics for the crisp results to enable a quantitative comparison between models.

Table 6-1 Rating system used to rank and compare performance of the models
(adapted from Moriasi et al. (2007).

<table>
<thead>
<tr>
<th>Description</th>
<th>NSE</th>
<th>RSR</th>
<th>PBIAS</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very good</td>
<td>0.75 &lt; NSE ≤ 1.00</td>
<td>0.00 &lt; RSR ≤ 0.50</td>
<td>PBIAS ≤</td>
<td>10%</td>
</tr>
<tr>
<td>Good</td>
<td>0.65 &lt; NSE ≤ 0.75</td>
<td>0.50 &lt; RSR ≤ 0.60</td>
<td></td>
<td>10% ≤ PBIAS &lt;</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>0.50 &lt; NSE ≤ 0.65</td>
<td>0.60 &lt; RSR ≤ 0.70</td>
<td></td>
<td>15% ≤ PBIAS &lt;</td>
</tr>
<tr>
<td>Unsatisfactory</td>
<td>NSE ≤ 0.5</td>
<td>RSR ≥ 0.50</td>
<td>PBIAS ≥</td>
<td>25%</td>
</tr>
</tbody>
</table>

To directly compare the crisp and fuzzy results, these metric ratings were then combined to give an “average rating” (AR) of each model: this was the average of the rating (i.e. 0, 0.33, 0.66 and 1) of each metric (NSE, RSR or PBIAS). The AR were then assigned a qualitative rating similar to Moriasi et al. (2007) for comparison and are shown in Table 6-2.

Table 6-2: Rating system used to compare the average rank of each simulation.

<table>
<thead>
<tr>
<th>Description</th>
<th>Average Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very good</td>
<td>AR &gt; 0.9</td>
</tr>
<tr>
<td>Good</td>
<td>0.66 &lt; AR ≤ 0.9</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>0.33 &lt; AR ≤ 0.66</td>
</tr>
<tr>
<td>Unsatisfactory</td>
<td>0.0 &lt; AR ≤ 0.33</td>
</tr>
</tbody>
</table>

Once this step was completed for all metrics (NSE, RSR and PBIAS) an overall rating was calculated by using the same rating table listed in the table above. In doing so, one model could be compared to another by the use of only one overall metric. This system was used to make more direct comparisons between different models (i.e. models at different lags, or with different error values).
In addition to these three metrics, a performance metric, $D$, was calculated for the risk analyses stage of this research. $D$ measured the relative difference between the observed peak flow rate and upper limit of either the fuzzy support, or the 95% confidence interval, for events where the observed peak flow rate was greater than the $Q_{P2\%}$. The objective of calculating and then comparing this metric was to determine which model performed better (i.e. was closest in magnitude to the observation) in cases where both models under-predicted the observed peak flow rate. The distance was calculated as:

$$D = |y_{upper} - y_i|$$

Equation 6-13

where $y_{upper}$ is either the upper limit of the fuzzy support $[\ldots]_{\mu} = 0$, or the 95% confidence interval.

6.2.5 Model implementation

Two distinct modelling approaches were taken. In the first phase, two proof-of-concept models (one using EIV method, the second using the fuzzy method) were constructed using the entire dataset, at different lags. In the second phase, a recursive algorithm was employed for both regression methods, where an additional year of data was sequentially added to the models, and updated the model parameters. The purpose of the first phase is to show the global appropriateness of the data-driven technique to predict peak flow rate and quantify the performance of both regression methods. The purpose of the second method is to demonstrate the utility of the methods in a quasi-real-time state. Lastly, the final version of the recursive model was then used for a test data set comprised of the 2013 flood year in Alberta.

6.2.5.1 Proof-of-concept models

A proof-of-concept model was constructed for both the EIV regression method and the FLR method at each of the three sites. In each case, six years of data was used for model construction, specifically data from 2000, 2002, 2004, 2006, 2008 and 2010. Model validation was done using another five years of data, specifically, 2001, 2003, 2005, 2007 and 2009. The flood year of 2005 was not included in the calibration. This modelling
exercise was conducted at four different lags, (i.e. lagged $Q_D(t - d)$) where $d = 1$ day, 2 days, 3 days or 7 days. In addition to this, three different error values, $e$, were used for fuzzy number construction, namely $e$ of 5%, 10% and 20%. Thus, a total of 12 models were constructed and validated at each of the 3 sites (with 4 lags and 3 error values). Error analyses were conducted for each model.

6.2.5.2 Recursive models
For each type of regression model, the recursive modelling set started with using data from the year 2000 for calibration, and data from the year 2001 for validation. An error analysis was then conducted for both model types. Following this, the calibration dataset was updated to include the years 2000 and 2001, while the validation dataset used data from the year 2002; the regression parameters were updated and another error analysis was conducted. This process was continued until ten years of data were used for calibration, and the final year (2010) was used for validation. This meant a total of ten models were constructed and validated for each of the three sites, at four different lags (1, 2, 3 and 7 days), at three error values (5%, 10% and 20%). The rating system described earlier was used to condense the results from the error analyses on each model to ease comparison between different simulations.

A risk analysis was then conducted on days when the observed flow was higher than $Q_{P2%}$. It should be noted that for each case where this was true, the predicted flow from the validation dataset was used for the analysis. For brevity’s sake this analysis was only conducted on models that used a 1 day lag for $Q_D$, and an error value $e$ of 20%, for each site.

6.2.5.3 Test case
To test the developed models, an independent dataset was used. This was daily peak flow rate and daily mean flow rate at Calgary for the year 2013, which was an extreme flood year in southern Alberta. Regression coefficients from the last recursive model, i.e. where ten years of data from 2000 to 2009 was used for model construction, was used to predict peak flow rate using the 2013 mean daily flow rate (only with a lag of one day, and an error value of 20%). An error analysis between the observed and predicted peak flow rate was conducted. Following this a risk analysis was conducted on the days in
2013 where the observed peak flow rate was higher than the historical (for the 2000 – 2010 period) $Q_{p2\%}$.

6.3 Chapter results and discussion

6.3.1 Model performance comparison

6.3.1.1 Proof-of-concept
The proof-of-concept models were constructed using 6 years of data, and validated using 5 years of data. These models were constructed at each of the 4 lags, and 3 error values, for each site. Figure 6-3 to Figure 6-5 shows sample results for one particularly model: Bow River at Calgary, with a lag of 1 day and $e$ of 20%. Figure 6-3 shows the trend plots of two years, 2010 (which was used for calibration, and is one of the lowest flow years in the dataset) and 2005 (which was used for validation and was a flood year). The figure shows that in general the observed peak flow rate falls within the 95% confidence interval (CI) and within the interval defined by the $\alpha$-cut at $\mu = 0$. For the Bow River at Banff site, the amount of observations captured within the predicted fuzzy interval (at $\mu = 0$) for lags of 1, 2, 3 and 7 days were 92.6%, 82.2%, 79.9% and 57.1%, respectively. At the Bow River at Calgary site, the amount of observations captured within the predicted fuzzy interval (at $\mu = 0$) for lags of 1, 2, 3 and 7 days were 97.2% 92.1%, 85.2%, and 72.1%, respectively. Lastly, at the Elbow River at Bragg Creek site, the amount of observations captured within the predicted fuzzy interval (at $\mu = 0$) for lags of 1, 2, 3 and 7 days were 91.0%, 84.3%, 82.1%, and 71.3%, respectively. These results show that there was minimal impact of the lead time or peak flow prediction timing since for the majority of events (especially at low lags), the observations fell within the predicted fuzzy interval for a given day. Of note, in both years is that the fuzzy interval increases and decreases at different points throughout the years reflecting the certainty associated with the prediction, whereas, the EIV regression interval is constant and independent of the magnitude of the predicted peak flow. Thus, during periods of high flow rate, the fuzzy interval expands to reflect the associated uncertainty with high flow predictions; and at lower flows, it reflects the opposite. In some lower flow cases show in Figure 6-3, the fuzzy interval is actually smaller than the crisp interval. While the
majority of observations are captured within the intervals, the fuzzy interval is generally closer to the observations when the observations do not fall within the interval.

A plot of the observed versus predicted peak flow rate for the entire dataset are included in Figure 6-4. The figure shows that the fuzzy numbers (represented here as a black square corresponding to the values at $\mu = 0$) and crisp numbers from both regression generally follow the 1:1 line. The fuzzy predictions tend to expand for higher peak flow rates (representing the higher uncertainty associated with high peak flow). The flood event of 2005 can be seen clearly in the validation plot: the fuzzy intervals are closer to the 1:1 line compared to the crisp results for the flood level flows.

Figure 6-5 plots the observed mean daily flow and daily peak flow for both the calibration and validation datasets. This figure clearly illustrates that the fuzzy intervals increase with the magnitude of peak flow rate, and thus, capture more of the uncertainty in the observations compared to the EIV regression method.

Figure 6-6 summarises three performance metrics in order to compare the performance of the EIV and FLR for the $e = 20\%$ proof-of-concept models (results for the other $e$ values are included in the Appendix I). In general, performance decreases, for both methods, as the lags are increased. The model rankings (discussed below) show that in general the model performances decline from “very good” at a 1 day lag to “satisfactory” at a 7 day lag. Thus, this indicates that the optimum approach for this data-driven method is limited to short term predictions, between 1 and 3 days. In general, better performance is seen with the higher $e$ value.
Figure 6-3 Results from the proof-of-concept models for a lag of 1 day, and e of 20%: trend plots of predicted daily peak flow rate from the error-in-variables and fuzzy linear regression methods for (top) the calibration (shown for 2010 only), and (bottom) the validation dataset (shown for 2005 only) for the Bow River at Calgary.
Figure 6-4 Results from the proof-of-concept models for a lag of 1 day, and e of 20%: observed versus predicted peak flow rate plots for the error-in-variables (black circle with line) and fuzzy linear regression (black boxes) methods for (left) the calibration, and (right) the validation dataset for the Bow River at Calgary.
Figure 6-5 Results from the proof-of-concept models for a lag of 1 day, and e of 20%: observed peak flow rate, the error-in-variables regression line, 95% confidence intervals, and fuzzy interval at μ = 0 for the (left) calibration and (right) validation dataset for Bow River at Calgary
For the Bow River at Banff, the average ranking for the validation datasets for the EIV and FLR models is “very good” for lags of 1 and 2 days for all $e$ values. The ranking decreases to “good” for a 3 day lag, and “satisfactory” for a 7 day lag for all $e$ values. Results for the calibration datasets (not shown) are typically equal to or higher than the validation dataset. For the Bow River at Calgary, the average ranking for the validation dataset for the FLR model is “very good” for lags at 1 day, “good” for lags at 2 and 3 days, and “satisfactory” for a lag of 7 days. The EIV model has ranking of “very good” for lags of 1 and 2 days, and “good” and “satisfactory” for lags of 3 and 7 days, respectively, for the validation dataset. Again, calibration results (not shown) were either equal to or higher than the validation datasets. For the Elbow River at Bragg Creek, the average ranking for the FLR model is “unsatisfactory” for all validation cases, except for a lag of 1 day and an $e$ of 20% where it is classified as “satisfactory”. The calibration dataset are ranked “satisfactory” or “good” for all cases. The validation dataset for the EIV model is ranked “satisfactory” (with AR = 0.33) for all four lags.

Between the sites, the Elbow River at Bragg Creek has noticeably lower performance than the other two sites. The Bragg Creek site has a much smaller drainage area than the other two sites, and a much larger range of annual peak flow rate ($15 – 307 \text{ m}^3/\text{s}$). This suggests that this site experiences more flash flooding events than the other two sites, and perhaps this model structure is not the optimum for this site. An independent parameter other than $Q_D$ could improve model performance and be more suitable for this site.
Figure 6-6 RSR, NSE and PBIAS values for the validation dataset for e = 20% proof-of-concept models: the markers (circle, square, rhombus, and triangle) represent results for different lags (1, 2, 3 and 7 days, respectively): Banff (top), Calgary (middle) and Bragg Creek (bottom)
6.3.1.2 Recursive methods

Figure 6-7 to Figure 6-9 show results for recursive Model 5 for the Bow River at Calgary (Model 5 uses data from 2000 to 2004 for model construction and 2005 for validation). The trend plot in Figure 6-7 shows results from 2005, for a lag of 1 day, and $e$ of 20%. These results again highlight the flexibility of the FLR approach: the predicted intervals increase and decrease as a function of the observed mean flow rate, and capture the majority of the observations. This is also highlighted in the results show in Figure 6-8 and Figure 6-9. Note that in Figure 6-9, for results at a lag of 1 day, for two high flow instances, the EIV results do not come close to the 1:1 line, but the FLR method (black squares) are closer to, or intersect, the 1:1 line. This means that for the same case, the FLR method can come closer to predicting the peak flow rate for a major flood event (2005) one day in advance, as compared to the EIV method.

Figure 6-10 show plots of the three performance metrics used, for each model, for each site, for each lag, and for $e = 20\%$. In general, the RSR plots show that the performance of all models, at all sites, does not change as the amount of data used is increased. This suggests that the issue of “over-learning” have been avoided in this case. The RSR performance decreases with increasing lags, and generally, a better performance is seen for the fuzzy system as the $e$ value increases. Better performances are generally seen at Banff and at Calgary than at Bragg Creek. The NSE values for Bragg Creek are much lower than those of the other two sites at lag 7; indicating the lack of a linear relationship. The improvement with increasing $e$ is more noticeable for Calgary and Bragg Creek than for Banff, where performance is high for all three $e$ cases. The results confirm that performance decreases as lag increases, and an improvement can be seen (especially for Calgary and Bragg Creek) as $e$ increases.
Figure 6-7 Results from the recursive model for a lag of 1 day, and $e$ of 20%: trend plots of predicted daily peak flow rate from the error-in-variables and fuzzy linear regression methods for the validation dataset (for 2005) for the Bow River at Calgary
Figure 6-8 Results from the recursive models for all lags, and e of 20%: observed versus predicted peak flow rate plots for the error-in-variables (grey lines) and fuzzy linear regression (black boxes) methods for the validation dataset (for the year 2005 only) for the Bow River at Calgary
Figure 6-9 Results from the recursive models for all lags, and e of 20%: observed peak flow rate, the error-in-variables regression line, 95% confidence intervals, and fuzzy interval at $\mu = 0$ for the validation dataset (for the year 2005 only) for Bow River at Calgary
For the FLR validation case, for an $e$ of 20%, the average ranking for Banff ranges from “very good” for all 10 models at 1 day lag, to a range between “good” and “very good” for a 2 day lag, “satisfactory” to “good” for a 3 day lag, and “satisfactory” for the 7 day lag (with one unsatisfactory case for Model 10). For Calgary, the average ranking for the FLR model was “very good” for both 1 and 2 day lags (with the exception of Model 5 for the two day lag which had a rating of “good”). The ranking was “satisfactory” to “very good” for the 3 day lag, and “satisfactory” to “good” for the 7 day lag. Lastly, for Bragg Creek, the average ranking for the FLR models was between “unsatisfactory” and “very good” for a 1 day lag, “unsatisfactory” and “good” for the 2 day lag, and “unsatisfactory” to “satisfactory” for the 3 and 7 day lags. Similar results were seen for the EIV case.

The last component of the model performance evaluation was calculating $D$. First, all days where the observed peak flow rate was higher than the calculated $Q_{p2%}$ flow rate were isolated. There were 30 occurrences of these high flows at Banff, 36 at Calgary, and 32 at Bragg Creek between 2000 and 2010. Then, for each model (only recursive models with $e$ of 20% are considered for this analysis) the days when the observed flow did not fall within the predicted intervals for both EIV and FLR was isolated. The distance between the upper limit of the interval and the observed peak flow rate was then calculated, and the mean of these distances was summarised for each model for each site; these results are shown in Table 6-3.
Figure 6-10 RSR, NSE and PBIAS values for the validation dataset for all ten recursive models with $e = 20\%$: the markers (circle, square, rhombus, and triangle) represent result for different lags (1, 2, 3 and 7 days, respectively), for each of the three sites: Banff (top), Calgary (middle) and Bragg Creek (bottom)
This table shows that for days when the models entirely under predict the observed peak flow rates, the FLR has a smaller distance between the highest point of the interval and the observation for a majority of cases. For example, in Banff the FLR has a significantly smaller distance (calculated using the two-sample Kolmogorov-Smirnov test at the 5% level) for the 1 and 2 day lag cases than for EIV. For Calgary, the FLR performs substantially better than EIV for lags 1 to 3; however, these differences were not significant and were impacted by the low sample size. Lastly, for the Bragg Creek models, the EIV has a smaller distance than the FLR for all cases, but these were not found to be significantly different.

<table>
<thead>
<tr>
<th>Lag</th>
<th>Banff</th>
<th>Calgary</th>
<th>Bragg Creek</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLR</td>
<td>EIV</td>
<td>FLR</td>
</tr>
<tr>
<td>1 day</td>
<td>0</td>
<td>24</td>
<td>15</td>
</tr>
<tr>
<td>2 days</td>
<td>19</td>
<td>39</td>
<td>8</td>
</tr>
<tr>
<td>3 days</td>
<td>46</td>
<td>52</td>
<td>11</td>
</tr>
<tr>
<td>7 days</td>
<td>102</td>
<td>74</td>
<td>22</td>
</tr>
</tbody>
</table>

6.3.2 Flood risk assessment

Due to the importance of flood risk assessment for the City of Calgary in the wake of the June 2013 floods, the risk assessment reported here is only for the Calgary site using the recursive model with a 1 day lag and $e$ of 20% (only validation data was used for this analysis). A subset of the full dataset was taken to highlight the difference between the models. For each site, days when the observed flow was greater than the $Q_{P2%}$ was selected, reflecting high peak flow observations. The data is further filtered to include only the days when observed values are not captured within the EIV model interval, and when the EIV model under predicts the observations. A plot of the 95% CI of the predictions, the corresponding observations, the fuzzy number, and the $Q_{P2%}$ (the “Warning” flow”) are plotted in Figure 6-11.

On each subplot, two probabilities calculated using the possibility to probability transformation are listed. The first value represents the probability that the predicted
fuzzy peak flow will be higher than the $Q_{P2\%}$ flow, and the second represents the probability that the predicted fuzzy peak flow will be higher than the observed peak flow. These plots show that even when the EIV model’s predicted peak flow rate does not capture the observed flow within its 95% CI, the fuzzy predictions not only capture this high flow but can estimate the probability of that event. For example, in Calgary on 8 June 2005 with a one day lag, the EIV interval is between 241 and 282 m$^3$/s, whereas the observed peak flow for the day was 312 m$^3$/s. The fuzzy number meanwhile ranges between 170 and 355 m$^3$/s, and predicts the probability that the peak flow would be greater than 312 m$^3$/s (what was observed) as 12%. Based on this, a water resource manager in the municipality could issue a flood advisory a day before this high peak flow was seen.

The utility of the fuzzy number approach lies in its ability to calculate the probability of exceedance from its predictions. A water resource manager can provide a risk tolerance, e.g. if the probability that a predicted peak flow will be higher than $Q_{P2\%}$ is greater than 5%, a “flood advisory” can be issued. Another tolerance for a more certain outcome, e.g. if the probability that a predicted peak flow will be higher than $Q_{P2\%}$ is greater than 50%, a “flood warning” can be issued. In general, the fuzzy predictions are equal to, if not better than the EIV method, but in addition, fuzzy numbers can provide more meaningful information for flood mitigation.
Figure 6-11 Possibility to probability transformations for high peak flow rate days (>Q_{P2%}), for Bow River at Calgary, with a lag of 1 day, and e of 20%
6.3.3 Model application: 2013 flood data

An analysis using the developed model was conducted on an independent dataset: Model 10 (both EIV and FLR) for the 1 day lag and e of 20% were used to predict the peak flow rate at Calgary in 2013. A similar analysis as for the previous years was conducted and is summarised in Figure 6-12 to Figure 6-14 below.

As Figure 6-12 shows, the fuzzy interval encompasses the peak flow rate for the entire year except for the one day leading up to the flood. This is highlighted in Figure 6-13, where the black boxes (representing fuzzy numbers) all cross the 1:1 line at all times, except for one occasion. However (discussed in detail below) during this particular day the fuzzy number is still able to predict the risk of high peak flow (i.e. \( P(x > Q_{P2\%}) \)) whereas the EIV method does not warn of this risk. The RSR value for the FLR ranged between 0.50 and 0.72, whereas the EIV had a value of 0.60. The NSE values for FLR ranged between 0.48 and 0.76, whereas the EIV had a value of 0.62. Lastly, the PBIAS for FLR ranged between 1.7 and 9.2\%, whereas the EIV had a value of 5\%. For both models, the average ranking was “satisfactory”, which is somewhat lower compared to the other 1 day, 20% e models for Calgary discussed earlier.

To focus on the flood event, a critical 12 day period from the 18\(^{th}\) to the 29\(^{th}\) of June, 2013 is plotted in Figure 6-12 as an insert. The figure clearly shows that the FLR method is more flexible and reactive to the flood event, by coming closer to the observed peak flow on the 20\(^{th}\) of June 2013 (unlike the EIV method) and capturing the peak flow rate on the 21\(^{st}\) of June 2013, which lies within the fuzzy interval. In fact, during this period, the EIV did not capture the observed peak flow within its 95\% CI interval 9 out of the 12 days, whereas the FLR method only missed the observations on 3 out of 12 occasions.

Figure 6-14 shows the possibility to probability transformation for this 12 day period in 2013. It shows that on June 20\(^{th}\) 2013, the fuzzy method still predicts a ~4\% chance that the observed peak flow will be greater than the warning level (\( Q_{P2\%} \)) whereas the EIV interval is still under the threshold. Thus, even with an event that rises extremely rapidly (from 235 m\(^3\)/s to 1600 m\(^3\)/s in 18 hours on June 20, 2013), the FLR method can predict a small risk of high peak flow a day in advance.
Figure 6-12 Results from the test case of 2013 for a lag of 1 day, and e of 20%: trend plot of predicted daily peak flow rate from the error-in-variables and fuzzy linear regression methods for the Bow River at Calgary. The insert shows the days of interest in June 2013 when the highest flows were measured.
Figure 6-13 Results from the test case of 2013 for a lag of 1 day, and $\epsilon$ of 20%: (left): observed versus predicted peak flow rate plots for the error-in-variables (grey lines) and fuzzy linear regression (black boxes) methods; and (right) observed peak flow rate, the error-in-variables regression line, 95% confidence intervals, and fuzzy interval at $\mu = 0$ for the Bow River at Calgary.
Figure 6-14 Possibility to probability transformations for high peak flow rate days (> Q_P), for Bow River at Calgary in 2013 during the flood, with a lag of 1 day, and a of 20%
6.4 Chapter summary

A fuzzy linear regression model was developed in this chapter and tested against an error-in-variables implementation of ordinary least squares regression for flood prediction in a large river basin. A unique way of transforming a possibility to a probability for this particular type of data-driven modelling technique was also developed. In addition, the authors proposed a method to compare and evaluate fuzzy linear regression modelling to ordinary least squares regression modelling by modifying Moriasi et al.’s (2007) categories of model performance. In applying the models to the Bow River system in southern Alberta that experienced devastating floods in 2013, the research showed that the flood in June 2013 could not have been predicted with any confidence using current, conventional methods of probabilistic reasoning but could have been predicted a full day in advance, and potentially three days in advance for the City of Calgary using only observed real-time flow data gauged in the Bow River inside Calgary (WSC ID: 05BH004). This would provide enough advance warning to initiate flood response measures. The implementation of this fuzzy linear regression would be amenable to reproduction as a web-based application for use by forecasters or the public.

The proposed method can be further fine-tuned to improve model performance. For example, given the availability of sub-hourly flow rate data at some Environment Canada stations, it may be possible to use hourly, or perhaps 6-hour averaged data, rather than daily (24-hour averaged) data that was used in forming the models in this chapter. Thus, the effect of a 1 time-unit lag would be less significant than in the current method, even though in the FLR method, the impact of the lag is compensated by the size of the fuzzy interval in most cases. However, it is important to note that by reducing the lead time to sub-daily lags, that there will not be enough time to for water resource managers to implement flood defence systems, which is the primary objective of this risk-based peak flow prediction method.

Another method of improving performance could be by increasing the error value \( e \) to up to 40% or 50% (from the maximum of 20% assumed for this analysis). Increasing the window of the fuzzy number would improve model performance and capture more events within the fuzzy interval. It should be noted that even with such a high uncertainty band the proposed model has shown that it still reduces to a small interval when necessary (i.e.}
during low flow). Also, it was assumed that the daily peak flow rate was a probabilistic parameter and converted to a fuzzy number for analysis; this is the most straightforward approach. However, a number of propositions on converting a single flow measurement to a fuzzy number have been proposed and can be considered.

Lastly, of the three sites considered, the models did not perform as well at the Elbow River at Bragg Creek site, as they did at the other two sites. Given this site’s smaller drainage area and history of flash flooding, a different independent parameter might improve model performance, such as using lagged peak flow rate rather than lagged daily mean flow rate.
7. Dissolved oxygen prediction and risk analysis using a fuzzy neural network\textsuperscript{11}

7.1 Chapter introduction
The dissolved oxygen (DO) concentration of a waterbody is the most fundamental indicator of overall aquatic ecosystem health (Dorfman & Jacoby, 1972; Hall, 1984; CCME, 1999; Kannel et al., 2007; Khan & Valeo, 2015a). A number of factors, biotic and abiotic (non-living, physical and chemical attributes), regulate or influence DO in riverine environments, e.g. re-aeration from the atmosphere, increased streamflow, or respiration of aquatic plants (Pogue & Anderson 1995; Hauer & Hill 2007; He et al., 2011; Wen et al., 2013; Khan et al., 2013; Khan & Valeo, 2015a). The DO concentration in rivers is influenced by a combination of direct impacts, such as contaminants that consume oxygen, and indirect impacts, such as nutrients that lead to an increase in vegetation which in turn leads to increase in photosynthesis which consumes oxygen (Antanasijević et al., 2014). Low DO concentration in particular, can increase the risk of adverse effects to the aquatic environment. While the impact of long-term effects are largely unknown, low oxygen events can have immediate devastating effects on ecosystems (Adams et al., 2013). Thus, DO is widely measured and modelled, and the identification and quantification of DO trends in rivers is of primary interest to water resource managers (Wen et al., 2013).

Changes in watershed land-use due to urbanisation add additional complexity to DO variability, making modelling and prediction more challenging. The interaction of numerous factors, over a relatively small area and across different temporal scales means that DO trends in urban areas are more difficult to predict (Hall, 1984; Niemczynowicz, 1999). Secondly, rapid changes in the urban environment – e.g. land-use changes which can affect runoff characteristics, major flood-events which can transform the existing ecosystem, the impact of various water quality improvement strategies employed within the watershed – means that the factors and regimes influencing water quality and DO in the riverine environment might also change rapidly. This means that managing and evaluating water quality in urban riverine environments is a difficult task (Hall, 1984).

\textsuperscript{11} Material from this chapter was submitted for publication in Environmental Modelling & Software.
In Calgary, Alberta, Canada, the Bow River has occasionally experienced low DO conditions in recent years. This River is extremely important for the region because it provides potable water for the residents of the expanding City; and is used for industrial and irrigation purposes in Southern Alberta, commercial and recreational fishing, and other recreational activities (Robinson et al., 2009; BRBC, 2010a). Thus, maintaining a high water quality standard is extremely important to meet the needs of the community and to create a sustainable aquatic ecosystem.

The stress on the Bow River is due, in part, to the high rate of development in the watershed. The river is one of the most regulated rivers in Alberta (BRBC, 2010a) resulting in the river approaching its assimilative capacity. High sediment and nutrient loads, effluent from three wastewater treatment plants, and stormwater runoff have contributed to reducing the health of the Bow River. The City of Calgary is mandated to meet the requirements of the provincial surface water quality guidelines, which means maintaining DO concentrations above 5 mg/L (1-day minimum), and above 6.5 mg/L (7-day average) (Alberta Environment, 1997). Thus, the City has implemented numerous strategies to limit loadings into the river in an effort to improve water quality and maintain high DO concentration.

As part of this strategy, the City of Calgary uses numerical modelling to predict the impact of different water quality improvement strategies within the watershed, as well as to forecast the future state of the Bow River under different scenarios. The physical processes that govern the behaviour of DO in the aquatic environment are quite complex, poorly understood and thus, physically-based models (that are typically used for managing DO in water bodies) require the parameterisation of a several different variables, which can be unavailable, expensive and time consuming (Antanasijević et al., 2014; Wen et al., 2013; Khan et al., 2013). In addition to this, many physically-based models cannot account for the types of rapid changes seen in urban areas (as described above), e.g. two major floods in 2005 and 2013 in Calgary completed altered the ecosystem; new wastewater treatment plants have come online and discharge their effluent into the River; and the real-time water quality monitoring stations have consistently been relocated further downstream. These factors highlight the fact that the uncertainty associated with a system increases proportionally to its complexity – the
uncertainty can arise as a result of vaguely known relationships among all these factors, in addition to the randomness in the system (Deng et al., 2011). Thus, DO predictions are extremely difficult and beset with uncertainty; hindering water resource managers from making objective decisions.

Thus, there is a need to create a numerical modelling method that can accurately predict DO concentration in a riverine environment, like that in Calgary. This method should be able to account for the uncertainty in DO prediction caused by the lack of understanding of governing processes, or issues related to parameterisation and data availability. This type of uncertainty, stemming from an incomplete understanding of the physical system is known an epistemic uncertainty. It is in contrast to aleatory uncertainty, which is typically dealt with using probability theory. An alternative but complimentary basis to deal with epistemic uncertainty in known as possibility theory (Jacquin, 2010; Deng et al., 2011).

### 7.1.1 Possibility theory & fuzzy numbers for uncertainty analysis

Possibility theory, an extension of fuzzy set theory, is an information theory used for representing uncertain, vague or imprecise information (Zadeh, 1978). Fuzzy sets were proposed by Zadeh (1965) in order to express imprecision in complex systems, and can be described as a generalisation of classical set theory (Khan & Valeo, 2015a). In classical set theory, an element \( x \) either belongs or does not belong to a set \( A \). In contrast, using fuzzy set theory, the elements \( x \) have a degree of membership within the fuzzy set \( A \). This degree of membership (typically denoted using the symbol \( \mu \)) is defined in the interval between 0 and 1. If \( \mu \) equals 0, then \( x \) does not belong in \( A \), and \( \mu = 1 \) means that it completely belongs in \( A \), while a value \( \mu = 0.5 \) means that it is only a partial member of \( A \).

Fuzzy numbers are an extension of fuzzy set theory, and express an uncertain or imprecise quantity. A fuzzy number does not refer to a single value, but represents the set of all possible values that define the quantity. A fuzzy number is defined as a fuzzy set that is *normal* and *convex*, and typically membership functions (denoted \( \mu(x) \) or \( \pi(x) \)) have a triangular shape, though this might not always be appropriate for hydrological applications (Khan et al., 2013; Khan & Valeo, 2015a). Fuzzy numbers are particularly useful for dealing with uncertainties when data are limited or imprecise (Bárdossy et al.,...
1990; Guyonnet et al., 2003; Huang et al., 2010; Zhang & Achari 2010a), and when a probabilistic representation of parameters may not be possible, since the exact values of parameters may be unknown, or only partial information is available (Zhang, 2009).

While fuzzy set theory based applications, particularly applications using fuzzy logic in neural networks, have been widely used in many fields including hydrology (Bárdossy et al., 2006; Abrahart et al., 2012), the use of fuzzy numbers and possibility theory based applications are rare (Bárdossy et al., 2006; Jacquin, 2010). Some examples include creating maps of soil hydrological properties (Martin-Clouaire et al., 2000), remotely-sensed soil moisture data (Verhoest et al., 2007), climate modelling (Mujumdar & Ghosh, 2008), subsurface contaminant transport (Zhang et al., 2009), and streamflow forecasting (Alvisi & Franchini, 2011). In the last example, Alvisi & Franchini (2011) introduce a new method to train an artificial neural network (ANN), a type of data-driven model, which uses fuzzy numbers to quantify the total uncertainty in the weights and biases of the network. These types of models are different than the more common physically-based or conceptual models, and are particularly well suited to use in conjunction with fuzzy numbers.

7.1.2 Data-driven models for DO prediction

Data-driven models are a class of numerical models based on generalised relationships, links or connections between input and output datasets (Solomatine & Ostfeld, 2008). These models can characterise a system with limited assumptions and are useful in solving practical problems, especially when there is lack of understanding of the underlying physical processes, time-series are of insufficient length, or when existing models are inadequate (Solomatine et al., 2008; Napolitano et al., 2011). Data-driven models typically have a simpler model structure, which means that propagating uncertainty from different sources is easier. With increased use of real-time water quality monitoring stations, there exists a great opportunity to utilise high resolution data that is continuously being collected at numerous locations in many jurisdictions. Thus, the availability of data, which is often cited as a limiting factor in the success of data-driven models (e.g. Solomatine et al., 2008) is not considered to be a major issue where high resolution datasets are readily available (e.g. in urban areas). In addition to this, the very nature of data-driven models means that many site-specific surveys that are needed to
calibrate physically-based models are no longer necessary, and arguably, data-driven models have lower data requirements than conceptual models (Antanasijević et al., 2014).

The use of data-driven models, such as neural networks, statistical methods, fuzzy rule-based systems, model trees, and genetic programming, has been widespread in hydrology (Shrestha & Solomatine, 2008; Solomatine et al., 2008; Elshorbagy et al., 2010) including DO prediction in rivers. Wen et al. (2013) used an ANN to predict DO in a river in China using ion concentration as the predictors. Antanasijević et al. (2014) used ANNs to predict DO in a river in Serbia using a Monte Carlo approach to quantify the uncertainty in model predictions and temperature as a predictor. Chang et al. (2015) also used ANNs coupled with hydrological factors (such as precipitation and discharge) to predict DO in a river in Taiwan. Singh et al. (2009) used water quality parameters to predict DO and biochemical oxygen demand in a river in India. Other studies (e.g. Heddam, 2014; Ay & Kisi, 2012; He et al., 2011) have used regression methods to predict DO in rivers using water temperature, or electrical conductivity, amongst others, as inputs. In general, these studies have demonstrated that there is a need for less complex DO models, and this has led to an increase in popularity of data-driven models (Antanasijević et al., 2014). These studies have demonstrated that data-driven models have suitable level of performance for DO prediction.

7.1.3 Chapter objectives

Given the importance of DO concentration as an indicator of overall aquatic ecosystem health, there is a need to accurately model and predict DO in urban riverine environments. In this chapter a new data-driven method is proposed that attempts to address the issues that plague numerical modelling of DO concentration in the Bow River in Calgary, Canada. A data-driven method that uses abiotic parameters as inputs, and fuzzy numbers to represent uncertainty is proposed. This method: (i) accounts for the complexity of the physical-system by using a data-driven approach (with generalised input and output relationships); (ii) uses abiotic inputs (such as water temperature and flow rate) since they are routinely collected and thus, a large dataset is available; and (iii) captures and propagates both the aleatory and epistemic uncertainty in the system through the use of fuzzy numbers.
Specifically, a fuzzy neural network (FNN; first proposed by Alvisi & Franchini, 2011) is adapted and advanced through the development of possibility theory-based intervals to predict daily minimum DO concentration in the Bow River; with fuzzy number-based network weights and biases to represent the total uncertainty. A method to find the optimum network architecture is also presented. Additionally, the application of the developed model will be demonstrated by analysing the risk of low DO concentration in the Bow River through a possibility-to-probability transformation. This method will then be used to create a tool for water resource managers to assess the risk of conditions that lead to low DO in the Bow River.

7.2 Methods\textsuperscript{12}

7.2.1 DO monitoring & modelling in the Bow River

The City of Calgary is located in the Bow River Basin (approximately 25123 km\textsuperscript{2} in area) in southern Alberta, Canada. The Bow River is 645 km long and averages a 0.4\% slope over its length (BRBC, 2010\textsuperscript{a}). The headwaters of the river are located at Bow Lake, in the Rocky Mountains, from where it flows south-easterly to Calgary (drainage area of 7870 km\textsuperscript{2}), meeting the Oldman River and eventually draining into Hudson Bay (Robinson et al., 2009; Environment Canada, 2014). The river is supplied by snowmelt from the Rocky Mountains, rainfall and runoff, and discharge from groundwater. The River has an average annual discharge of 90 m\textsuperscript{3}/s, and an average width and depth of 100 m and 1.5 m, respectively (Khan & Valeo, 2015b).

The City of Calgary routinely samples a variety of water quality parameters along the Bow River to measure the impacts of urbanisation, particularly from three wastewater treatment plants and numerous stormwater runoff outfalls that discharge into it. Real-time water quality monitoring systems are stationed at the upstream (at the Bearspaw reservoir) and downstream (currently at Highwood) ends of the City (as shown in Figure 7-1). Comparing water quality data from the stations shows the direct impact of the urban area on the Bow River (Khan & Valeo, 2014\textsuperscript{a}). DO concentration measured at the upstream site is generally high throughout the year, with little diurnal variation (He et al., 2011; Khan et al., 2013; Khan & Valeo, 2015\textsuperscript{a}). In contrast, the concentration

\textsuperscript{12}The MATLAB code associated with this section is included in Appendix K.
downstream of the City limits is typically lower, and experiences much higher fluctuations. The three wastewater treatment plants (shown in Figure 7-1) are located upstream of this monitoring site, and are thought to be responsible, along with other impacts of urbanisation, for the degradation of water quality at the site.

Figure 7-1 An aerial view of Calgary, Canada showing the locations of (a) Water Survey of Canada flow monitoring site “Bow River at Calgary”, (b) Bonnybrook, (c) Fish Creek, and (d) Pine Creek wastewater treatment plants, and two water quality sampling sites (e) Stier’s Ranch and (f) Highwood
For this research, nine years of DO concentration data was collected from the downstream station for the period from 2004 to 2012. From 2004 to 2007 the monitoring station was located at Pine Creek and sampled every 30 minutes (for 2004 and 2005), and every 15 minutes (for 2006 and 2007). In 2008 the station was moved to Stier's Ranch where it remained until 2011, and sampled data every hour (in 2008) and every 15 minutes from there on. The site was moved further downstream to its current location (at Highwood) in 2012 where it sampled every 15 minutes. During this period a number of low DO events were observed. The number of days where the observed DO was measured to be below 5 mg/L (the acute guideline) was 25 days in 2004, 1 day in 2005 and 25 in 2006. The number of days with DO concentration below 6.5 mg/L (the chronic guideline) was more frequent with a total of 184 occurrences: 41, 26, 70, 27, 5 and 15 days in 2004, 2005, 2006, 2007, 2008 and 2009, respectively.

A YSI sonde is used to monitor DO and the sonde is not accurate in freezing water, thus only data from the ice-free period was considered, which is approximately from April to October for most years (YSI Inc., 2013). Since low DO events usually occur in the summer (corresponding to high water temperature and lower discharge), the ice-free period dataset contains the dates that are of interest for low DO modelling.

The City has implemented numerous strategies to limit loadings into the river – such as the Total Loadings Management Plan and the Bow River Phosphorus Management Plan (Neupane et al., 2014) – and use numerical modelling as part of its strategy to predict not only the impacts of strategies on water quality, but also to forecast the future state of the river under different scenarios. Currently, the Bow River Water Quality Model (see Tetra Tech, 2013 and Golder, 2004 for details), a physically–based model, is used to predict DO and other water quality parameters. This model suffers from many of the same issues related to complexity and uncertainty that are discussed above. In response to this, recent research by He et al., (2011), Khan et al. (2013), Khan & Valeo (2014a) and Khan & Valeo (2015a) has shown that data–driven models, particularly those that use abiotic factors as inputs, have promising results to predict DO concentration in the River. Examples of the abiotic factors used in these studies are water temperature, nutrient concentration, flow rate and solar radiation; these factors are commonly monitored in many jurisdictions including in Calgary. The advantage of using readily available data in
these studies was that if a suitable relationship between these factors and DO can be found, then changing the factors (e.g. increasing the flow rate downstream of a treatment plant) could improve DO at times when the risk of low DO is high.

For this research mean daily water temperature (T) and mean daily flow rate (Q) were selected as the abiotic input parameters for the FNN data-driven model. The reason for selecting these two variables was their use in previous studies in this river basin (He et al., 2011; Khan et al., 2013, Khan & Valeo, 2015a). Results from those studies have shown that these two variables are good predictors of daily minimum and mean DO concentration.

The same YSI sonde used to sample DO concentration was used to collect T. The Water Survey of Canada site “Bow River at Calgary” (ID: 05BH004 shown in Figure 7-1) was used to collect Q. This data is collected hourly throughout the year; data where considerable shift corrections are applied (usually due to ice conditions) were removed from the analysis. The mean annual water temperature ranged between 9.23 and 13.2°C, and the annual mean flow rate was between 75 and 146 m³/s, for the selected period.

7.2.2 Fuzzy neural networks

7.2.2.1 Background on ANN

ANNs are a type of data-driven model that are defined as a massively parallel distributed information processing system (Elshorbagy et al., 2010; Wen et al., 2013). ANNs were designed to mimic biological neural networks (BNN) in the human brain and attempt to model at least two features of BNNs: (i) the ability to perform functions in parallel rather than linearly, and (ii) produce an output (or firing) only when a certain input threshold is reached. The benefit of these properties is that when ANNs are used as a predictive model they can capture complex, nonlinear relationships that exist between variables without an explicit understanding of the physical phenomenon (Alvisi & Franchini, 2011; Antanasijević et al., 2014), making it an ideal candidate for DO prediction in riverine environments. Also, some types of ANNs are considered to be universal approximators (Hornik et al., 1989), which means that theoretically the models have the capacity of approximating any linear or nonlinear mapping (from input to output), whether it be for time-series forecasting or pattern recognition (ASCE 2000; Elshorbagy et al., 2010; Napolitano et al., 2011; Kasiviswanathan et al., 2013). This has
resulted in significant use of ANN models in hydrology where the complexity of the physical systems is high owing partially to an incomplete understanding of the underlying processes, and lack of availability of necessary data (He et al., 2011; Kasiviswanathan et al., 2013). Further, ANNs arguably require less data and do not require an explicit mathematical description of the underlying physical process (Antanasijević et al., 2014).

Over the last two decades, countless research has been published on the use of ANNs in hydrology. Recent surveys on the state of the use of ANNs in the field were conducted by Abrahart et al., (2012) and Maier et al., (2010). The majority of the published literature has focused on streamflow prediction, and less than 10% have focused on water quality (Maier et al., 2010). Maier et al., (2010) also identified that since ANN models are data-driven, they can best-use the available water quality data (which is typically less than water quantity data) as compared to process-driven water quality models that require difficult to obtain and more extensive data.

There are three major types of ANNs used in hydrology: feedforward networks where data propagation is only in one direction (from input to output), recursive networks where feedback loops may be used (often to model dynamic systems with time delays), and lastly a hybrid network where principles of ANNs are combined with other modelling techniques (Maier et al., 2010; Betrie et al., 2014). Multilayer Perceptron (MLP) is a type of feedforward ANN and is one of the most commonly used in hydrology (Maier et al., 2010). One of the reasons for the popularity of MLPs is that a trained network can be used as a universal approximator with a relatively simple architecture: a single hidden-layer (Hornik et al., 1989). Further, the popularity of MLP has meant that subsequent research has continued to use MLP (He & Valeo 2009; Napolitano et al., 2011) and thus, form a reference for the basis of comparing ANN performance (Alvisi & Franchini, 2011).
In the simplest case, an MLP consists of three layers: an input-layer, a hidden-layer, and an output-layer (as shown in Figure 7-2). Each layer consists of a number of neurons (or nodes) that each receive a signal, and on the basis of the strength of that signal, emit an output (similar to the function of a BNN). Thus, the final output-layer is the synthesis and transformation of all the input signals from the input and hidden-layers (He & Valeo, 2009). The number of neurons in the input ($n_I$) and output ($n_O$) layers corresponds to the number of variables used as the input and the output, of the MLP, respectively. The number of neurons in the hidden-layer ($n_H$) are reflective of the complexity of the system being modelled; more neurons represent a more complex system (Elshorbagy et al., 2010). Each neuron in the hidden-layer, receives a signal from the input-layer and is combined with an individual weight and a bias. The weight determines the strength of the signal. This linear combination of input signal, weights and biases is then fed through a transfer function that maps this modified signal into a separable space (typically between
This means that the strength of the input signal is governed by the magnitude of the output from the transfer function, e.g. firing when the value is 1, and not firing when the value is -1, similar to the BNN function in the human brain. The combined output signals from the hidden-layer are then combined with another set of weights and biases, and finally fed through a second transfer function, which is typically a simple linear function (Elshorbagy et al., 2010), giving the final value of the output-layer. This type of MLP can be expressed mathematically as follows:

\[ H = f_{\text{HID}}(W_{\text{IH}}*I + B_{\text{H}}) \]  \hspace{1cm} \text{Equation 7-1}

\[ O = f_{\text{OUT}}(W_{\text{HO}}*H + B_{\text{O}}) \]  \hspace{1cm} \text{Equation 7-2}

where \( I \) is the \( i \)th observation (an \( n_{\text{I}} \times 1 \) vector) from of a total of \( n \) observations, \( W_{\text{IH}} \) is a \( n_{\text{H}} \times n_{\text{I}} \) matrix of weights between the input and hidden-layer, \( B_{\text{H}} \) is a vector (\( n_{\text{H}} \times 1 \)) of biases in the hidden-layer, and \( H \) is the \( i \)th output (an \( n_{\text{H}} \times 1 \) vector) of the input signal through the hidden-layer transfer function, \( f_{\text{HID}} \). Similarly, \( W_{\text{HO}} \) is an \( n_{\text{O}} \times n_{\text{H}} \) matrix of weights between the hidden and output-layers, \( B_{\text{O}} \) is an \( n_{\text{O}} \times 1 \) vector of biases in the output-layer, and \( f_{\text{OUT}} \) the final transfer function to generate the \( i \)th modelled output \( O_i \) (an \( n_{\text{O}} \times 1 \) vector).

The values of all the weights and biases in the MLP are calculated by training (i.e. calibrating) the network by minimising the error (typically mean squared error, MSE, is used) between the modelled output and the observed data (also known as the target) (He & Valeo, 2009). A number of training algorithms exist: one of the most commonly used variant is the Levenberg–Marquardt algorithm (LMA) (Alvisi et al., 2006). In LMA the error between the output and target is back-propagated through the model using a gradient method where the weights and biases are adjusted in the direction of maximum error reduction. The LMA is particularly well-suited for ANN problems that have a relatively small number of neurons.

There are two major consequences of using this LMA-based training approach: firstly, the optimum result from LMA may result in overfitting, which means the MLP has a limited capacity to predict the output beyond the data used for training (Abrahart et al., 2012). To counteract this problem, an early-stopping procedure is used (Alvisi et al.,
of regularisation where the data is split into three subsets (for training, validation and testing) and the training is terminated when the error on the validation subset increases from the previous iteration. In this case, during each minimisation iteration, the performance (MSE) is calculated for both the training and validation subsets, and the metric is expected to decrease. However, at some point the MSE will continue to decrease for the training subset, but not for the validation subset, indicating the start of overfitting. Thus, at this point the minimisation is stopped, and the final MSE is calculated using the testing dataset. Secondly, the starting point of the LMA minimisation problem can impact the final value of the weights and biases (due to multiple, closely-spaced local minima), suggesting that a range of values for the weights and biases may be possible for the same dataset (Maier et al., 2010).

These issues highlight a broader issue with ANN application, namely that the choice of network architecture – the number of hidden-layers, the number of neurons in each hidden-layer, the selection of transfer function between each layer, the amount of data in each subset, the training algorithm used, and the performance metric selected for optimisation – will all impact the final values of the weights and biases, and the overall model performance. This suggests that each MLP model has uncertainty pertaining to the trained parameters, and that the impact of the choice of final parameters selected (e.g. one that gives the lowest MSE while subjecting some constraints to the other variables) on model performance is hard to pinpoint. In fact, recent reviews of ANN applications in hydrology have indicated that the lack of uncertainty quantification is a major reason for the limited appeal of ANN by water resource managers (Abrahart et al., 2012; Maier et al., 2010). Most ANN applications have a deterministic structure that generate point predictions without a quantification of the intervals corresponding to these predictions (Alvisi & Franchini, 2012; Kasiviswanathan & Sudheer, 2013). This means that end-users of these models may have excessive confidence in the forecasted values, and overestimate the applicability of the results (Alvisi & Franchini, 2011). However, as with all numerical models, the appropriate characterisation of uncertainty in a model is essential for both research and operational purposes (Alvisi et al., 2013). Without this characterisation, the results produced by these models have limited value (Kasiviswanathan & Sudheer, 2013).
In this chapter, two separate methods are proposed to quantify the uncertainty in MLP modelling. First, the *total* uncertainty in the weights and biases of an MLP are quantified using a possibility theory-based FNN. Here total uncertainty (as defined by Alvisi & Franchini, 2011) represents the overall uncertainty in the modelling process, and not in the individual components (e.g. randomness in observed data). Second, the uncertainty due to model structure is reduced by developing a transparent algorithm to select the number of neurons, and the amount of data for each subset (train, validate and test), to maximise model performance. These two methods are developed and presented in the following section.

7.2.2.2 Possibility theory based intervals for FNN

Alvisi & Franchini (2011) proposed a method to create a FNN, where the weights and biases, and therefore the output, of the neural network are fuzzy numbers rather than crisp (deterministic or non-fuzzy) numbers. These fuzzy numbers quantify the total uncertainty of the calibrated parameters. This application is unique compared to the more typical use of fuzzy set theory based applications that involve ANN, e.g. the widely used fuzzy logic based Adaptive Neuro-Fuzzy Inference System, where automated IF-THEN rules are used to create crisp outputs (Abrahart et al., 2010; Alvisi & Franchini, 2011). These types of methods do not quantify the uncertainty of the predictions.

In the FNN method developed by Alvisi & Franchini (2011), the MLP model presented in Equation 7-1 and Equation 7-2 is modified to predict an interval rather than a single value for the weights, biases and outputs, corresponding to an $\alpha$-cut interval (an interval corresponding to the lower and upper limits of a fuzzy set at a defined membership level $\alpha$). This is repeated for several $\alpha$-cut levels thus, building a discretised fuzzy number at a number of membership levels. This is done by using a stepwise, constrained optimisation approach:

\[
[H^L \ H^U] = f_{hid}([W_{hi}^L \ W_{hi}^U] * I + [B_{hi}^L \ B_{hi}^U])
\]

\[
[O^L \ O^U] = f_{out}([W_{ho}^L \ W_{ho}^U] *[H^L \ H^U] + [B_{o}^L \ B_{o}^U])
\]

Equation 7-3

Equation 7-4

where all the variables are as described as before, and the superscripts U and L represent the upper and lower limits of the interval, respectively. The constraints are such
that to find the upper and lower limits of each weight and bias (in both layers) to
minimise the width of the predicted interval, and the following equations are used:

\[
\begin{align*}
\text{minimise: } & (\Sigma |O^L - O^U|) \\
\text{subject to: } & 1/n \Sigma (\delta) \geq PI \\
\text{where } & \delta = 1 \text{ if } O^L \leq t \leq O^U \\
& = 0 \text{ otherwise}
\end{align*}
\]

where \(t_i\) is the \(i^{th}\) target (observed data) and \(PI\) is a predefined percentage of data. Alvisi
& Franchini (2011) defined the \(PI\) to be 100% at \(\mu = 0\), 99% at \(\mu = 0.25\), 95% at \(\mu = 0.5\) and 90% at \(\mu = 0.75\). This algorithm was built starting at \(\mu = 0\) and moving to higher
membership levels to maintain convex membership functions of the generated fuzzy
numbers, by using the results of the previous optimisation as the upper and lower limit
constraints for the proceeding optimisation. Lastly, at \(\mu = 1\), the interval collapses to a
singleton, representing the crisp results from a non-fuzzy ANN. Thus, these \(\alpha\)-cut
intervals of the output represent a type of confidence interval (in the possibility sense
rather than probability sense) of the observed data to fall within the predicted interval.
Thus, the advantage of fuzzy outputs is that it provides the uncertainty of the predictions
while the fuzzy parameters reflect the uncertainty in the model structure.

However, one major drawback of the approach proposed by Alvisi & Franchini (2011)
is that the selection of the \(PI\) in the predicted interval is arbitrary and limiting. End-users
may be interested in the risk of events that are not captured within the selected \(PI\) and
may require a full spectrum of possible values for a given prediction. Thus, the original
approach is further refined by utilising the relationship between possibility theory and
probability theory (known as probability-possibility transformations), giving a more
objective means of designing FNNs with fuzzy weights, biases and outputs.

Probability-to-possibility transformations are methods to create fuzzy numbers from
observed data. Recent discussions on possibility theory and summaries of different
conceptual approaches to these transformations are provided in Klir & Parvais (1992),
Oussalah (2000), Jacquin (2010), Mauris (2013), and Dubois & Prade (2015). In this
chapter, we use the principles of the transformation proposed by Dubois et al., (2004),
where the possibility is viewed as the upper envelope of the family of probability measures (Jacquin, 2010; Ferrero et al., 2013; Betrie et al., 2014), and implemented by Zhang et al., (2009) and Khan & Valeo (2014a; 2015a; 2015b). In this transformation, the interval \([a \ b]_\alpha\) created by the \(\alpha\)-cut at \(\mu = \alpha\) implies that:

\[p(x < a) + p(x > b)] = \alpha\]  \hspace{1cm} \text{Equation 7-8}

\[p(a < x < b) = (1 - \alpha)\]  \hspace{1cm} \text{Equation 7-9}

This means that there is a probability of \((1 - \alpha)\) that the random variable \(x\) falls with the interval \([a \ b]_\alpha\). Thus, in other words the \(\alpha\)-cuts of a possibility distribution (at any \(\mu\)) correspond to the \((1 - \alpha)\) confidence interval of the probability distribution of the same variable (Serrurier & Prade, 2013).

This principle is used to select the different \(PI\) for the optimisation constraints rather than the predetermined \(PI\) selected by Alvisi & Franchini (2011). Thus, the constraints for the FNN then become: 100% at \(\mu = 0\), 80% at \(\mu = 0.20\), 60% at \(\mu = 0.4\), 40% at \(\mu = 0.6\), 20% at \(\mu = 0.8\), and 0% at \(\mu = 1.0\). For practical purposes, instead of calculating a \(PI\) at 0% for \(\mu = 1\), the crisp ANN results are used following the procedure of Alvisi & Franchini (2011). Similarly, to prevent over-fitting, the interval at \(\mu = 0\) was adjusted to 99.5%. The reason for selecting these specific six specific \(PI\) is to provide results with sufficient detail rather than the excessive smoothness that occur at a lower number of levels – e.g. at two levels the fuzzy number reduces to a triangular shaped fuzzy number.

Using this process, the least credible, and hence the widest interval with most uncertainty, will capture all the observations within the predicted interval. As the membership level increases, the amount of data captured reduces in relation to the \(PI\). Note that the \(PI\) values used by Alvisi & Franchini (2011) can still be implemented in the proposed method: for 99%, the \(\mu\) necessary would be 0.01 as opposed to the 0.25 they used (following Equation 7-9), and would be 0.05 for the 95% interval, 0.10 for the 90% interval. Using the proposed \(PI\), these intervals are captured between the \(\mu = 0\) and 0.2 membership levels.

This FNN optimisation algorithm was implemented in MATLAB (version 2015a). The crisp results of the ANN (at \(\mu = 1\)) were conducted using the built-in MATLAB Neural
Network Toolbox with an early-stopping procedure to prevent over-fitting. Subsequent optimisation for the other intervals was conducted using a two-step approach: first the Shuffled Complex Evolution algorithm (commonly known as SCE-UA, Duan et al., 1992) was used to find an initial solution to the minimisation problem. Then, further refinement of the solution was conducted using the built-in MATLAB minimisation function \textit{fmincon}. The optimisation is run such that the intervals at higher membership levels govern the upper and lower bounds of the predicted interval in order to preserve the convexity of fuzzy numbers.

7.2.2.3 Selecting the FNN architecture and structure

For this research a three layer, feedforward MLP architecture was selected to model minimum daily DO (the output) using Q and T as the inputs. The three layers consist of an input-layer, an output-layer, and a hidden-layer. Previous studies modelling minimum DO in the Bow River have also used a three-layer MLP feedforward network (see He et al., 2011), one of the most common structures for MLPs.

Two transfer functions were required – one between the input and hidden-layer, \( f_{\text{HID}} \), which was selected as the hyperbolic tangent sigmoid function, and one between the hidden-layer and output-layer, \( f_{\text{OUT}} \), which was a pure linear function (both selections following Alvisi & Franchini, 2011; Wen et al., 2012; Elshorbagy et al., 2010):

\[
\begin{align*}
  f_{\text{HID}} &= \frac{e^x - e^{-x}}{e^x + e^{-x}} \\
  f_{\text{OUT}} &= x
\end{align*}
\]

Equation 7-10
Equation 7-11

To calibrate the crisp ANN (i.e. at \( \mu = 1 \)), a backpropagation algorithm, the LMA method was used to train the network, minimising MSE. The input and output data were pre-processed before training, validation and testing. The data was normalised so that all input and output data fell within the interval [-1 1].

Two factors relating to ANN structure and data-handling, the number of neurons in the hidden-layer (\( n_H \)), and the amount of data used for training, validation and testing (known as data-division), still need to be identified. There is no consistent method used in the literature for each of these factors (Abrahart et al., 2010; Dubois & Prade, 1993; Civanlar & Trussel, 1986). Typically, an ad hoc, or trial-and-error method is used to select the
number of neurons (Maier et al., 2010; Elshorbagy et al., 2010; Alvisi & Franchini, 2011; He et al., 2011). The number of neurons selected must balance the complexity and generalisation of the final model; too many neurons increase the complexity and hence the processing speed, while reducing the transparency of the model. Not enough neurons risk reducing model performance and forgoing the ability of modelling non-linear systems. Similarly, the issue of data-division, which can have significant impacts on final model structure, is also predominantly conducted in an ad hoc and trial-and-error basis (Maier et al., 2010). Generally speaking two broad methods are available: in the first, each subset should have data that is statistically similar, including similar patterns or trends. Conversely, a method can be selected where each subset is based on some physical property, such as grouping the subsets chronologically.

In this chapter, we propose a coupled method to select the optimum $n_H$ and data-division for the ANN model described. This method uses the independent test dataset for all inferences regarding the error statistic, MSE, to be minimised. The smallest number of neurons and the least amount of data for training is targeted. The first is to reduce computational effort. The second is to prevent the risk of overfitting to the training data, and to have a larger dataset for testing for more robust statistical inference of that dataset.

First the dataset is randomly split into a 50%-25%-25% ratio for training, validation and testing, respectively. For each subset, data is randomly sampled to group into each subset. Then, the network is trained using 1 to 20 neurons. This process is repeated 100 times to account for the different selection of randomly sampled data in each subset. This is because the random initialisation of the ANN can cause variability in overall model performance (Napolitano et al., 2011). Thus, each iteration is different than the previous. The MSE for each test dataset is then calculated and compared, as well as the number of epochs for each model is measured. Epochs are the number of times each weight and bias is modified in the optimisation algorithm. The configuration that leads to the lowest MSE for the testing dataset and the lowest number of epochs (for the training dataset only, by definition) was collected. This process is then repeated by sequentially increasing the amount of data used for training (and thus, reducing the amount of data equally partitioned for validation and testing) by increments of 0.5% from 50% to 75%, conducting each iteration of this change 100 times.
Using this process, a number of different circumstances where the highest performance was seen can be listed, and the best combination (i.e. number of neurons and amount of training data) can be objectively selected. In doing so, both the processing time (i.e. the number of epochs, the amount of training data and the number of neurons) and the complexity (number of neurons) of the system is accounted for in the final model architecture.

### 7.2.3 Risk analysis for low DO

A major component of this research is to be able to use a trained fuzzy ANN model to predict the risk of low DO using abiotic inputs. Risk analyses for complex systems, generally speaking, is challenging for a number of reasons, including an insufficient understanding of the failure mechanisms (Deng et al., 2011). An advantage of the proposed approach is that using a system that utilises imprecise information (e.g. through the use of fuzzy numbers) is an effective method of conducting risk analysis (Deng et al., 2011). Another advantage is that by identifying the abiotic factors that seem to influence DO concentration, then the possibility of low DO can be directly inferred from the FNN outputs for any combination of input parameters (in this case Q and T).

However, communicating uncertainty, which can be done in a number of different ways, is an important, yet difficult task (Van Steenbergen et al., 2012). Given the overall preference of the general public and water resource managers for using probabilistic measures (rather than possibilistic measures), it is important to convert the possibility of low DO to an equivalent probability for communicating risk and uncertainty. It is worth noting here that even linguistic parameters such as “most likely” that are often used to convey uncertainty (Van Steenbergen et al., 2012) have a probability based underlying meaning (in this case likelihood). Thus, there are several advantages to use probability theory for the communication of uncertainty.

In this chapter, *defuzzification* is used to convert the possibility of low DO (i.e. predicted minimum DO to be below a given threshold) to a probability measure. Mathematically, the *defuzzification* method used is the inverse of the transformation used to guide the selection of the \( \alpha \)-cut intervals for the FNN optimisation in Section 7.2.2.2. The transformation is given as follows (following Khan & Valeo, 2014a, 2015b): for any \( x \) in the support (the \( \alpha \)-cut interval at \( \mu = 0 \)) of a fuzzy number \([a, b] \), we have the
corresponding $\mu(x)$ and the paired value $x'$ which also shares the same membership level. The value $\mu(x)$ is the sum of the cumulative probability distribution between $[a, x]$ and $[x', b]$, labelled $A^L$ and $A^R$, respectively:

$$\mu(x) = A^L + A^R$$

Equation 7-12

where $A^L$ represents the cumulative probability between $a$ and $x$ which is assumed to equal the probability $P(X < x)$, since the fuzzy number defines any values to less than $a$ to be impossible (i.e. $\mu = 0$). Given the fact that the fuzzy number is not symmetrical, the lengths of the two intervals $[a, x]$ and $[x', b]$ can be used to establish a relationship between $A^L$ and $A^R$. Then, $A^L$ can be estimated as:

$$A^L = P(X < x) = \mu(x)/[1 + [(b - x')/(x - a)]]$$

Equation 7-13

Thus, Equation 7-13, gives the probability that the predicted minimum DO for a given day is below the threshold value $x$. For example, if the lowest acceptable DO concentration of 5 mg/L (Alberta Environment, 1997) is selected as this threshold, then this transformation can be used to calculate the probability that daily minimum DO is below 5 mg/L. Similarly, the same principle can be used to map out all combinations of input values that lead to a risk of low DO (i.e. below 5 mg/L or any other threshold value). Once this is mapped out, combinations that lead to a major risk of low DO can be identified by water resource managers. The overall objective is to provide a simple tool to quantify the risk of low DO under selected T and Q conditions. For this research, the calibrated FNN model was used to predict minimum DO for various combinations of Q and T: where Q ranged between 40 m$^3$/s and 220 m$^3$/s (at 2 m$^3$/s intervals), and T ranged between 0 and 25°C (at 0.2°C intervals). For each prediction, the risk of low DO to be below either 5, 6.5 or 9.5 mg/L threshold was calculated using the previously described defuzzification technique. These intervals of Q and T were selected to reflect typical conditions in the Bow River. The thresholds correspond to the minimum acceptable DO concentration for the protection of aquatic life for 1-day (at 5 mg/L) (AENV, 1997), and
for the protection of aquatic life in cold, freshwater for early-life stages (at 9.5 mg/L) and other-life stages (6.5 mg/L) (CCME, 1999).

7.3 Chapter results and discussion

7.3.1 FNN structure

For this research, a three-layer, feedforward MLP network architecture was selected to model minimum daily DO (the output-layer), with two variables, Q and T in the input-layer. A coupled method was used to determine the optimum number of neurons in the hidden-layer and the percentage of data for each of the training, validation and testing subsets. Sample results of the proposed method are shown in Figure 7-3. Figure 7-3a shows the mean MSE (solid black line) of the test dataset for the initial 50%-25%-25% data-division scenario, with \( n_H \) varying between 1 to 20 neurons. This simulation was repeated 100 times to account for the random selection of data and the upper and lower limits of MSE for each of these simulations are shown in grey. This figure demonstrates that the number of neurons did not have a noticeable impact on the MSE for this configuration. The most significant outcome of this process is that the variability (the difference between the upper and lower limits) of the performance seems to decrease after \( n_H = 6 \) and increases again after \( n_H = 12 \), with the lowest MSE at \( n_H = 10 \). This result has two important implications: first, increasing the model complexity results in limited improvement of model performance, suggesting that a simpler model structure may be more suitable to describe the system. Second, the variability in performance indicates that the initial selection of data in each of the training subset can highly influence the performance of the test dataset, especially at the lower (i.e., \( n_H < 6 \)) and higher (i.e., \( n_H > 12 \)) ends of the spectrum of the proposed number of neurons. This suggests that an optimum selection of hidden neurons lies within this range (6 < \( n_H < 12 \)).

Figure 7-3b shows the change in the mean (solid black line) and the variability (in grey) of the number of epochs needed to train the network for the initial data-division scenario, as \( n_H \) increases from 1 to 20. While the mean value does not show a notable change, the variability of the time needed for training (i.e. the number of epochs) drastically decreases as \( n_H \) increases from 1 to 5. This means that a simpler model structure may require more time to train, and the performance of these simpler architectures (\( n_H = 1 \) to 5) is more variable. This is likely due to the fact that the initial
dataset selection has a higher impact on the final model performance for less complex models. The lowest number of mean epochs for this analysis occurred at $n_H = 19$, with 26 epochs. However, note that the variability of MSE at $n_H = 19$ (in Figure 7-3a) is high, and that $n_H = 19$ fall outside the range $6 < n_H < 12$, identified above.

The impact of changing the amount of data used for training, validation and testing on the model performance (MSE) was generally inconclusive as the amount of data used for training was increased from 50% to 75% at 0.5% intervals. Figure 7-3c shows sample results for the $n_H = 10$ scenario, which was the best performing scenario, i.e. had the lowest mean MSE for each data-division scenarios when compared to other $n_H$ values. However, the subplot illustrates that the MSE for the test dataset does not show a major trend as the amount of data for training is increased from the initial 50%. This means that for this scenario ($n_H = 10$) increasing the amount of data used for training has minimal impact on model performance, indicating that using the least amount of data for training (and thus having a higher fraction available for validating and testing) would be ideal. Note that the mean MSE values were generally higher for all data-division scenarios when the selected $n_H$ was between 1 and 5 (following the example shown in Figure 7-3a).

The number of epochs needed for training the network at different data-division scenarios was inconclusive. Figure 7-3d shows sample results for the $n_H = 10$ case, which demonstrates that the mean and the variability of the number of epochs does not demonstrate a clear trend, as the amount of data used for training is increased. The significance of this analysis is that the amount of computational effort (or time) does not necessarily decrease as a larger fraction of data is used for training. Given this result, the least amount of training data (50%) is the preferred choice for the number of neurons that result in the lowest MSE, which is $n_H = 10$ as described above. For the $n_H = 10$ case, the overall mean number of epochs for each data-division scenario is low ranging between 24 and 40 epochs.
Figure 7-3 Sample results of the coupled method to determine the optimum number of neurons in the hidden-layer and percentage of data for training, validation and testing subsets; the mean (solid black line) and upper and lower limits (in grey) of (a) the MSE for the test dataset for each number of hidden neurons; (b) the number of epochs for training; (c) the MSE for a range of training data size; and (d) the number of epochs for 10 hidden neurons
Based on these results, \( n_H = 10 \) with a 50%-25%-25% data-division was selected as the optimum architecture for this research. The fact that the mean and the variability of MSE was the lowest at \( n_H = 10 \) makes it a preferred option over the \( n_H = 19 \) case, which as a lower number of mean epochs but had higher variability in MSE. In other words, higher model performance was selected over model training speed (mean epochs at \( n_H = 10 \) ranged between 9 and 122 for the 50%-25%-25% data-division scenario). Secondly no significant trend was seen as the amount of data used for training, validation and testing was altered, however lower MSE values were seen at \( n_H = 10 \) compared to other at \( n_H \) values. Thus, the option that guarantees the largest amount of independent data for validation and training is preferred. Given the fact that the mean MSE for the testing dataset does not show a significant change as the amount of training data is increased from 50% to 75%, the initial 50%-25%-25% division is maintained as the final selection.

The overall outcome of this component of this research was that that instead of using the typical trial-and-error based approach to selecting neural network architecture parameters, the proposed method can provide objective results. Specifically, systematically exploring different numbers of hidden layers and fraction of training data, can help select a model with the highest performance, whilst accounting for the randomness in data-selection. Once these neural network architecture parameters were identified, subsequent training of the FNN was completed by using Equation 7-3 to Equation 7-7. The results of the training and optimisation are presented in the next section.

### 7.3.2 FNN training, validation and testing results

First, the network was trained at \( \mu = 1 \) using the network structure outlined in the previous section. The crisp, abiotic inputs (\( Q \) and \( T \)) were used to estimate the values of each weight and bias in the network. This amounted to 20 \( W_{IH} \) (10 for each input), 10 \( B_{IH} \), 10 \( W_{HO} \), and 1 \( B_o \). The MSE and the Nash-Sutcliffe model efficiency coefficient (NSE; Nash & Sutcliffe, 1970) for the training, validation and testing scenarios for the \( \mu = 1 \) case are shown in Table 7-1 below.
Table 7-1 MSE and NSE for the neural network at a $\mu = 1$

<table>
<thead>
<tr>
<th></th>
<th>MSE (mg/L)$^2$</th>
<th>NSE (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>1.04</td>
<td>0.59</td>
</tr>
<tr>
<td>Validation</td>
<td>1.22</td>
<td>0.61</td>
</tr>
<tr>
<td>Test</td>
<td>1.42</td>
<td>0.51</td>
</tr>
</tbody>
</table>

The MSE for each dataset are low, approximately between 10 and 18% of the mean annual minimum DO seen in the Bow River for the study period. The NSE values are greater than 0.5 for each subset, which is higher than NSE values for most water quality parameters (when modelled daily) reported in the literature (Moriasi et al., 2007) and is considered “satisfactory” using the ranking system proposed by Moriasi et al. (2007). These two model performance metrics highlight that predicting minimum DO using abiotic inputs and a data-driven approach is an effective technique. Additional advantages of the proposed approach are seen when the fuzzy component of the results are analysed.

Once the crisp network was trained, a top-down approach was taken to train the remaining intervals, starting at $\mu = 0.8$ where 20% of the observations should be captured within the corresponding predicted output interval, and continuing to $\mu = 0$ where 99.5% of the observations should be captured within the predicted output intervals. Each set of optimisation (both the SCE-UA and fmincon algorithms) for each of the five remaining membership levels ($\mu = 0, 0.2, 0.4, 0.6$ and 0.8) took approximately 2 hours using a 2.40 GHz Intel® Xeon microprocessor (with 4 GB RAM). The results of this optimisation are summarised in Table 7-2, which shows the amount of data captured within the resulting $\alpha$-cut intervals after each optimisation.

Table 7-2 Percentage of data (%) (%) captured within each fuzzy interval for the FNN model

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.80</td>
<td>20.02</td>
<td>16.10</td>
<td>15.85</td>
</tr>
<tr>
<td>0.60</td>
<td>40.05</td>
<td>34.63</td>
<td>33.41</td>
</tr>
<tr>
<td>0.40</td>
<td>60.07</td>
<td>52.44</td>
<td>52.44</td>
</tr>
<tr>
<td>0.20</td>
<td>80.10</td>
<td>80.00</td>
<td>80.24</td>
</tr>
<tr>
<td>0.00</td>
<td>99.51</td>
<td>98.78</td>
<td>99.02</td>
</tr>
</tbody>
</table>
For the training dataset, the coupled algorithm was able to capture the exact amount of data it was required to. Similar results can be seen for the validation and testing datasets, where the amount of data captured are close to the constraints posed. For the two independent datasets, the amount of coverage decreases (i.e. lower performance) as the membership level increases, which is unavoidable when the width of the uncertainty bands decrease. Similar results were seen for the testing dataset in Alvisi & Franchini (2011).

The result of this optimisation was the calibrated values of the fuzzy weights and biases; sample membership functions for four weighs and biases are shown in Figure 7-4. The figure illustrates that the shapes of all membership functions were convex, a consequence of the top-down calibration approach, where the interval at lower membership levels is constrained to include the entire interval at higher membership levels. In other words, since lower membership levels include a higher amount of data, the corresponding interval to that level is wider than at higher membership levels. Furthermore, since the crisp network was used at $\mu = 1$, there is at least one element in each fuzzy number with $\mu = 1$, meaning that each weight and bias is a normal fuzzy set.

The membership functions of the weights and biases are assumed to be piecewise linear following the assumption made in Alvisi & Franchini (2011), Khan et al. (2013) and Khan & Valeo (2015a). This means that enough $\alpha$-cut levels need to be selected to completely define the shape of the membership functions. In this chapter six levels were selected to equally span, and to give a full spectrum of possibilities, between 0 and 1. Overall, the results of the weights and biases in the figure demonstrate that indeed enough levels were selected to define the shape of the membership function. If a smaller number of levels were selected, e.g. two levels, one at $\mu = 0$ and one at $\mu = 1$, the fuzzy number collapses to a triangular fuzzy number. This type of fuzzy number does not provide a full description of the uncertainty and how it changes in relation to the membership level. Thus, intermediate intervals are necessary and the results demonstrate that the functions are in fact not triangular shaped functions and not necessarily symmetric about the modal value (at $\mu = 1$). A consequence of this is that the decrease in size of the intervals does not follow a linear relationship with the membership level.
Figure 7-4 Sample results of the FNN optimisation algorithm to estimate the fuzzy number values of each weight and bias in the model

Similarly, a higher number of intervals than the six selected for this research could be used, e.g. 100 intervals, equally spaced between 0 and 1. The risk in selecting many intervals is that as the membership level increases (closer to 1) the intervals become narrower as a consequence of convexity. This will result in numerous closely spaced intervals, with essentially equal upper and lower bounds, making the extra information redundant. This is demonstrated in the sample membership functions for $W_{ih} = 10$ in Figure 7-4, where increasing the number of membership levels between 0.4 and 1.0
would not improve the shape or description of the membership function. This is because the existing intervals are already quite narrow. Defining more uncertainty bands between the existing levels would not add more detail but would merely replicate the information already calculated.

Table 7-2 and Figure 7-4 demonstrate the overall success of the proposed approach to calibrate an FNN model. Whereas in many fuzzy set applications the membership function is not defined or selected based on a consistent, transparent or objective method (Abrahart et al., 2010; Dubois & Prade, 1993; Civanlar & Trussel, 1986); or is selected arbitrarily (as noted by Alvisi & Franchini, 2011); the method proposed in this chapter is based on possibility theory and provides an objective method to create membership functions. The results of the process show that the method is capable in creating fuzzy number weights and biases that are convex and normal, and capture the required percentage of data within each interval. However, it is worth emphasising that the uncertainty represented by these fuzzy weights and biases is not the uncertainty of the particular parameter, but is the total forecasting uncertainty defined by the quantifying bands around the crisp predicted value.

Figure 7-5 shows the results of observed versus crisp predictions (black dots) and fuzzy predictions (black line) of minimum DO at the membership level $\mu = 0$ for the three different data subsets. The figure illustrates that nearly-all (i.e. ~99.5%) of the observations fall within the $\alpha$-cut interval defined at $\mu = 0$ since the black dots are enveloped by the grey high-low lines of the fuzzy interval. The figure also demonstrates the advantage of the FNN over a simple ANN: while some of the crisp predictions veer away from the 1:1 line (especially at low DO, defined to be less than 6.5 mg/L here), nearly-all (i.e. ~99.5%) of the fuzzy predictions intersect the 1:1 line. This also illustrates that the NSE values reported in Table 7-1 (which were calculated for the crisp case at $\mu = 1$ only) are not representative of the NSE values of the fuzzy number predictions, and that there is a need to develop an equivalent performance metric when comparing crisp observations to fuzzy number predictions.

Lastly, the figure demonstrates the superiority of the FNN to be able to predict more of the low DO events compared to the crisp method. The low DO range (at DO = 6.5 mg/L) is highlighted in Figure 7-5, and it is apparent that within this window both models tend
to over predict minimum DO as they fall above the 1:1 line. However, the fuzzy intervals (grey lines) predicted by the FNN intersect the 1:1 line for the majority of low DO events, and hence predict some possibility (even if it is a low probability) at any $\mu$ of the low DO events, whereas the crisp do not predict any possibility at all. Thus, generally speaking the ability of the FNN to capture 99.5\% of the data within its predicted intervals guarantees that most of the low DO events are successfully predicted. This is a major improvement over conventional methods used to predict low DO.

Figure 7-6 and Figure 7-7 show trend plots of observed minimum DO and predicted fuzzy minimum DO for the years 2004, 2006, 2007 and 2010. These particular years were selected due to the high number of low DO occurrences in each year, and thus, are intended to highlight the utility of the proposed method to predict minimum DO using abiotic factors in the absence of complete understanding of the physical mechanisms. Note that for each year shown approximately 50\% of the data constitute training data, while the other 50\% constitute a combination of validation and testing data. However, for clarity this difference is not explicitly shown in these figures.

The most number of days where minimum daily DO was below the 5 mg/L guideline was observed in 2004 and 2006, with 25 days in each year. The year 2007 had the third most (after 2006 and 2004, respectively) number of days below the 6.5 mg/L guideline, with 27 days. Lastly, 2010 had the second most number of days below the 9.5 mg/L guideline with 180 days (after 2007 which had 182) below the guideline. Thus, these four years collectively represent the years with the lowest minimum daily DO during the study period. It is noteworthy that even though minimum DO was observed to be below 5 mg/L on several occasions in 2004 and 2006, the DO was below on 9.5 mg/L only 107 and 164 days, respectively, for these years. In contrast, in 2007 and 2010, no observations below 5 mg/L were seen, however 182 and 180 days, respectively, below the 9.5 mg/L were seen for those years. This indicates that even if no observations of DO < 5 mg/L are seen, it is not a good indicator of a healthy ecosystem, since the overall mean DO of the year might be low (e.g. a majority of days below the 9.5 mg/L). The implication of this is that only using one guideline is not a good indicator of overall aquatic ecosystem health.
Figure 7-5 A comparison of the observed and predicted crisp (black dots) and fuzzy intervals at $\mu = 0$ (grey lines) for minimum DO in the Bow River for the training, validation and testing datasets.

In Figure 7-6 and Figure 7-7, the predicted minimum DO at equivalent membership levels (e.g. $0^L$ or $0^R$) at different times steps are joined together creating bands representative of the predicted fuzzy numbers calculated at each time step. In doing so, it
is apparent that all the observed values fall within the $\mu = 0$ interval for the years 2006, 2007 and 2010, and nearly-all observations in 2004. This difference in 2004 is due to the fact that the PI for $\mu = 0$ was selected to be 99.5% rather than 100% to prevent overfitting; the optimisation algorithm was designed to eliminate the outliers first to minimise the predicted interval. The low DO values in 2004 are the lowest of the study period, and thus are not captured by the FNN.

The width of each band correspond the amount of uncertainty associated with each membership level, for example, the bands are the widest at $\mu = 0$, meaning the results have the most vagueness associated with it. Narrower band are seen as the membership level increases until $\mu = 1$, which gives crisp results. This reflects the decrease in vagueness, increase in credibility, or less uncertainty of the predicted value. In each of the years shown note that the observations fall closer to the higher credibility bands, except for some of the low DO events. This means that the low DO events are mostly captured with less certainty or credibility (typically between the $0^L$ and $0.2^L$ levels) compared to the higher DO events. However, it should be noted that compared to a crisp ANN, the proposed method provides some possibility of low DO, whereas the former only predicts a crisp result without a possibility of low DO. Thus, the ability to capture the full array of minimum DO within different intervals is an advantaged of the proposed method over existing methods.

Results from 2004 show that minimum DO decreases rapidly starting in late June and continuing until late July, followed by a few days of missing data and near-zero measurements, before increasing to higher DO concentration. Details of this trend are shown in Figure 7-8. The reason for this rapid decrease is unclear and may be an issue with the monitoring device. However, it demonstrates that the efficacy of data-driven methods is dependent on the quality of the data. One of the advantages of the proposed method is that while it is able to capture nearly-all of the observations (including outliers) within the least certainty band (at $\mu = 0$), other observations are mostly captured within higher certainty bands ($\mu > 0$). As the data length increases (i.e. the addition of more data and the FNN is updated), the number of outliers included with the $\mu = 0$ band will decrease because the optimisation algorithm (Equation 7-5 to Equation 7-7) searches for
the smallest width of the interval whilst including 99.5% of the data. Thus, with more data, the 0.5% not included in the interval will be the type of outliers seen in 2004.

The time series plot for 2006 shows that all the observations fall within the predicted intervals. The majority of the 25 low DO (i.e. less than 5 mg/L) occur starting in mid-July and continuing occasionally until mid-September. Unlike in 2004, all of these low DO events are captured within both the $\mu = 0.2$ and 0 intervals. This suggests that the model predicts these low DO events with more credibility than in the 2004 case. Details of some of the low DO events are also plotted in Figure 7-8 (for September). This plot shows that the low DO events are captured between $\mu = 0$ and 0.2 intervals. Even though the membership level is low, the general trend profile of the observed minimum daily DO is captured by the modelled results.

Figure 7-7 illustrates the time series predictions for 2007, and it clearly demonstrates that most of the observations are captured at higher membership levels, unlike the 2004 and 2006 examples, i.e. only a limited number of observations are seen between the $\mu = 0$ and 0.2 intervals for the entire year. In addition to this, 26 out of the 27 low DO observations (in this case below 6.5 mg/L) for this year are captured within the predicted membership levels (between $\mu = 0$ and 0.2). Figure 7-8 shows details of a low DO period in 2007 in July and August, which shows that the observations are evenly scattered around the $\mu = 1$ line.
Figure 7-6 Time-series comparison of the observations and FNN minimum DO for 2004 and 2006
The trend plot for 2010 is shown Figure 7-7 and it is clear that nearly-all observations for the year are below than 10 mg/L, and about 87% of all observations are below the 9.5 mg/L guideline. As with the 2007 case, the bulk of the observations are captured within high credibility intervals, owing to the lack of extremely low DO (i.e. below 6.5 or 5 mg/L). The trend plot illustrates that the FNN generally reproduces the overall trend of observed minimum DO. This can be seen in a period in early May where DO falls from a high of 10 mg/L to a low of 7 mg/L, and all the predicted intervals replicate the trend. This is an indication that the two abiotic input parameters are suitable parameters for predicting minimum DO in this urbanised watershed. Figure 7-8 shows details of a low DO (below 9.5 mg/L) event in 2010 in late July through late August. The bulk of low DO events are captured between the $\mu = 0.6$ and $0.2$ intervals – demonstrating that these values are predicted with higher credibility than the low DO cases ($< 5$ mg/L) in 2004 and 2006. Some of the low DO ($< 9.5$ mg/L) events highlighted in this plot are underestimated by the crisp ANN method. This means that in general the crisp method tends to over-predict extremely low DO events (i.e. those less than 5 mg/L) while under-predicting the less than 9.5 mg/L events. In both cases, the fuzzy method is able to capture the observations within its predicted intervals.
Figure 7-7 Time-series comparison of the observations and FNN minimum DO for 2007 and 2010
Figure 7-8 Detailed view of time series for minimum observed DO and predicted fuzzy DO for 2004, 2006, 2007 and 2010, corresponding to days with low DO events
The analysis of the trend plots for these four sample years show that the proposed FNN method is extremely versatile in capturing the observed daily minimum DO in the Bow River using abiotic (Q and T) inputs. The crisp case (at $\mu = 1$) cannot capture the low DO events (as shown in Figure 7-5 and Figure 7-6), while the FNN is able to capture these events within some membership level. Generally speaking the top-down training method selected for the FNN has been successful in creating nested-intervals to represent the predicted fuzzy numbers. The width of the predicted intervals corresponds to the certainty of the predictions (i.e. larger intervals for more uncertainty). The utility of this method is further demonstrated in the proceeding section, where a risk analysis tool for low DO events in the Bow River is presented.

7.3.3 Risk analyses and low DO tool

Daily minimum DO was observed to be below 5 mg/L on 51 occasions in the Bow River for the study period between 2004 and 2012. The FNN method predicted DO be less than 5 mg/L on each of the 51 occasions (at some possibility level), giving a 100% success rate, whereas the crisp ANN only correctly predicted one of these low DO events, a 2% success rate. Similarly, for the 6.5 mg/L guideline, the FNN was able to correctly identify all 184 occasions where minimum daily DO was less than the threshold. The crisp ANN only predicted 52% (96) of these low DO events. Lastly, of the 1151 occasions where the daily minimum DO was observed to be less than 9.5 mg/L – the FNN identified low DO events for each of these days, whereas the crisp method predicted 97% of these events. To summarise, the FNN method is able to predict all the low DO events, and performs significantly better compared to the crisp case for the most extreme low DO guideline (i.e. at the 5 mg/L).

Figure 7-9 shows sample plots of observed and predicted fuzzy minimum DO, corresponding to the lowest DO concentration observed for each year for the study period. Note that the membership functions of the predicted fuzzy numbers show that each one is convex and normal (matching the results of the weights and biases), highlighting the fact that the proposed optimisation algorithm successfully produces nested intervals at each membership levels. As with the weights and biases shown in Figure 7-4, the intervals are the largest at $\mu = 0$, which decreases in size as the membership level increases. The shape of the membership functions are not triangular
shaped as assumed in many fuzzy set based applications. This is of significance because it shows that the amount of uncertainty (or credibility) does not change linearly with the magnitude of DO, which has important implications regarding the risk of low DO, discussed in detail below.

Comparing the crisp results at $\mu = 1$ to fuzzy number predictions, it is apparent that on a number of occasions (e.g. 2006, 2005 and 2004) when the observed DO is below 5 mg/L, the crisp prediction does not predict low DO while the fuzzy number predicts a possibility of DO to be below 5 mg/L, (between $\mu = 0$ and 0.2 for the illustrated examples). Even in the 2004 case, where both the crisp and fuzzy predictions over predict the observed DO, the fuzzy prediction still predicts some possibility of low DO, whereas the crisp results do not. These examples illustrate that low DO prediction using the FNN is superior to the crisp case.

The calibrated FNN model was used to create a low DO risk tool where the risk (i.e. probability) of low DO was calculated from the fuzzy DO using the inverse transformation described in Equation 7-13. First, values of Q and T were selected to represent average conditions in the Bow River. The flow rate selected was between 40 and 220 m$^3$/s at 2 m$^3$/s intervals, and water temperature was between 0 and 25°C at 0.2°C intervals. For each combination of Q and T, the fuzzy DO was calculated using the FNN. The inverse transformation was used to calculate the probability of predicted DO to be below 5 mg/L for each combination of inputs.

The results of this analysis are illustrated in Figure 7-10 which shows the change in probability of low DO for different inputs. Generally the figure correctly recreates the conditions that lead to low DO events in the Bow River: low flow rates and high water temperature. The highest risk of low DO is when T ranges from 21 to 24°C and Q ranges from 40 to 100 m$^3$/s: the probability of low DO is more than 90%. The risk of low DO decreases with higher flow rate and lower temperature. The utility of this method is that a water-resource manager can use forecasted water temperature data and expected flow rates to quantify the risk of low DO events in the Bow River, and can plan accordingly. For example, if the risk of low DO reaches a specified numerical threshold or trigger, different actions or strategies (e.g. increasing flow rate in the river by controlled release from the upstream dams) can be implemented. The quantification of the risk to specific
probabilities means that the severity of the response can be tuned to the severity of the calculated risk.

Figure 7-9: Membership functions of the predicted minimum DO and observed minimum DO for corresponding to the lowest DO observation for each year between 2004 and 2012

It is worth highlighting here (as mentioned above) that this demarcation of probabilities for different inputs would not have been possible if only two membership levels (at $\mu = 0$ and 1) were used to construct the fuzzy number weights, biases and output. This is
because using only two membership levels would result in triangular membership functions which would show a linear change in probabilities with the change in magnitude of inputs. However, as the results in the previous section have shown, the change in the width of the intervals versus the change in membership levels is not linear. This is also highlighted in Figure 7-10, where there is no linear change in the risk of low DO with the inputs.

This process was repeated for two more cases to calculate the probability of low DO to be below 6.5 mg/L and 9.5 mg/L; the resulting risk of low DO are shown in Figure 7-11 and Figure 7-12. Similar results can be seen for these two cases, where the risk of low DO increases with increasing temperature and decreasing flow rate, as expected. These figures demonstrate that the probability of low DO is generally high for the type of conditions seen in the Bow River. The mean annual water temperature and flow rate for the study period was between 9.23 and 13.2°C, and 75 and 146 m³/s, respectively. For these conditions, the probability of DO to be less than 9.5 mg/L ranges between ~50% to more than 90% (based on results presented in Figure 7-12). This risk increases in the summer months where the average daily water temperature in the Bow River is usually above 10°C; under this condition there is a high risk of low DO even at high flow rates, as seen in Figure 7-12. In contrast to this, Figure 7-11 shows that there is a relatively higher risk of low DO (below 6.5 mg/L) in the spring (April and May) and late summer (September), when flow rate can be as low as 50 m³/s, and the water temperature varies between 10 to 15°C, resulting in a risk of low DO of about ~60%. These examples are meant to illustrate the potential utility of the data-driven and abiotic input parameter DO model, that can be used to assess the risk of low DO. Given that it is a data-driven approach, the model can be continually updated as more data is available, further refining the predictions.
Figure 7-10 Results of the low DO identification and risk analyses tool for DO less than 5 mg/L
Figure 7-11 Results of the low DO identification and risk analyses tool for DO less than 6.5 mg/L
Figure 7-12 Results of the low DO identification and risk analyses tool for DO less than 9.5 mg/L
7.4 Chapter summary

In this chapter a new method to predict daily minimum DO concentration in the Bow River in a highly urbanised watershed (Calgary, Canada) is presented. Due to the difficulties in calibrating and quantifying the uncertainty associated with physically-based models, a data-driven approach is proposed that used abiotic factors (non-living, physical and chemical attributes) namely water temperature and flow rate as inputs to the model. A possibility theory based refinement to an existing fuzzy neural network (FNN) model that was first proposed by Alivisi & Franchini (2011) is developed and implemented. In addition to this, a method of determining the optimum network architecture is presented, that is used to select the number of hidden neurons and the amount of data used for training, validation and testing.

The proposed FNN was successfully calibrated using a combination of SCE-UA and \textit{fmincon} optimisation. The results of the FNN method clearly demonstrate its advantage over a non-fuzzy approach: specifically more of the low DO events (identified when DO concentration is less than either 5, 6.5 or 9.5 mg/L) were correctly identified using the fuzzy method. The results of the method demonstrate the applicability of an abiotic input and data-driven based approach to predict DO concentration, in highly complex and uncertain environment.

The calibrated model was then used to create a low DO identification and risk analysis tool. Various combinations of inputs were used to predict fuzzy minimum DO, and a defuzzification process was used to predict the probability of low DO for a given pair of input values. This tool can be used by water resource managers and decision-makers to identify and quantify conditions that lead to low DO, and thus can implement suitable strategies to prevent the occurrence of low DO.
8. Dissolved oxygen prediction using possibility theory based neural network

8.1 Chapter introduction

The City of Calgary is a major economic hub in western Canada. With a rapidly growing population, currently estimated in excess of 1 million, the City is undergoing expansion and urbanisation to accommodate the changes. The Bow River is a relatively small river (with an average annual flow of 90 m$^3$/s, with an average width of 100 m, and depth of 1.5 m) that flows through the City and provides approximately 60% of the residents with potable water (Khan & Valeo, 2015a; 2015b). In addition to this, water is diverted from within the City for irrigation, is used as a source for commercial and recreational fisheries, and is the source of drinking water for communities downstream of Calgary (Robinson et al., 2009; BRBC, 2010a). This highlights the importance of the Bow River, not just as a source of potable water, but also as a major economic resource.

However, urbanisation has the potential to reduce the health of the Bow River, which is fast approaching its assimilative capacity and is one of the most regulated rivers in Alberta (BRBC, 2010a). Three wastewater treatment plants (shown in Figure 8-1) and numerous stormwater outfalls discharge their effluent into the River and are considered to be a major cause of water quality degradation in the River (He et al., 2015). This highlights some of the major impacts on the Bow River from the surrounding urban area. A number of municipal and provincial programs are in place to reduce the loading of nutrients and sediments into the river such as the Total Loadings Management Plan and the Bow River Phosphorus Management Plan (Neupane et al., 2014) as well as modelling efforts – namely the Bow River Water Quality Model (Tetra Tech, 2013; Golder, 2004) – to predict the impact of different water management programs on the water quality.

One of the major concerns is that low dissolved oxygen (DO) concentration has occurred on a number of occasions over the last decade in the Bow River within the City limits. DO is an indicator of overall health of the aquatic ecosystem (Dorfman & Jacoby, 1972; Hall, 1984; Canadian Council of Ministers of the Environment, 1999; Kannel et al., 2007; Khan and Valeo, 2014a; 2015a), and low DO – which can be caused by a number

---

13 Material from this chapter was submitted for publication in *Hydrology & Earth System Sciences*. 
of different factors (Pogue and Anderson 1995; Hauer and Hill 2007; He et al., 2011; Wen et al., 2013) – can impact various organisms in the waterbody. While the impact of long-term effects of low DO are largely unknown, acute events can have devastating effects on aquatic ecosystems (Adams et al., 2013). Thus, maintaining a suitably high DO concentration, and water quality in general, is of utmost importance to the City of Calgary and downstream stakeholders, particularly as the City is being challenged to meet its water quality targets (Robinson et al., 2009).

A number of recent studies have examined the DO in the Bow River, and the factors that impact its concentration. Iwanyshyn et al., (2008) found the diurnal variation in DO and nutrient (nitrate and phosphate) concentration was highly correlated, suggesting that biogeochemical processes (photosynthesis and respiration of aquatic vegetation) had a dominant impact on nutrient concentration rather than wastewater treatment effluent. Further, Robinson et al., (2009) found that the DO fluctuations in the River were primarily due to periphyton rather macrophyte biogeochemical processes. In both studies, the seasonality of DO, nutrients, and biological concentration, and external factors (e.g. flood events) were demonstrative of the complexity in understanding river processes in an urban area, and that consideration of various inputs, outputs and their interaction is important to fully understand the system. He et al. (2011) found that seasonal variations in DO in the Bow River could be explained by a combination of abiotic factors (such as climatic and hydrometric conditions), as well as biotic factors. The study found that while photosynthesis and respiration of biota are the main drivers of DO fluctuation, the role of nutrients (from both point and non-point sources) was ambiguous. Neupane et al. (2014) found that organic materials and nutrients from point and non-point sources influence DO concentration in the River. The likelihood of low DO was highest downstream of wastewater treatment plants, and that non-point sources have a significant impact in the open-water season. Using a physically-based model, Neupane et al. (2014) predicted low DO concentration more frequently in the future in the Bow River owing to higher phosphorus concentration in the water, as well as impacts of climate change.

A major issue of modelling DO in the Bow River is that rapid urbanisation within the watershed has resulted in substantial changes to land-use characteristics, sediment and nutrient loads, and to other factors that govern DO. Major flood events (like those in
2005 and 2013) completely alter the aquatic ecosystem, while new wastewater treatment plants (e.g. the Pine Creek wastewater treatment plant) added in response to the growing population further increases the stress downstream. These types of changes in a watershed increase the complexity of the system, making DO trends and variability more challenging to model. The interaction of numerous factors, over a relatively small area and across different temporal scales means that DO trends in urban areas are more difficult to predict and evaluating water quality in urban riverine environments is a difficult task (Hall, 1984; Niemczynowicz, 1999).

The implication of this is that the simplistic representation described in conceptual, physically-based models is not suitable for complex systems, i.e. where the underlying physical mechanisms behind the factors that govern DO are still not clearly understood, in a rapidly changing urban environment. Physically-based models require the parameterisation of a several different variables which may be unavailable, expensive and time consuming (Antanasijević et al., 2014; Wen et al., 2013; Khan et al., 2013). In addition to this, the increase in complexity in an urban system proportionally increases the uncertainty in the system. This uncertainty can arise as a result of vaguely known relationships among all the factors that influence DO, in addition to the inherent randomness in the system (Deng et al., 2011). The rapid changes in an urban area render the system dynamic as opposed to stationary, which is what is typically assumed for many probability-based uncertainty quantification methods. Thus, not only is DO prediction difficult, it is beset with uncertainty, hindering water resource managers from making objective decisions.

In this chapter, we propose a new method to predict DO concentration in the Bow River using a data-driven approach (as opposed to a physically-based method) that uses possibility theory and fuzzy numbers to represent the uncertainty rather than the more commonly used probability theory. Data-driven models are a class of numerical models based on generalised relationships, links or connections between input and output datasets (Solomantine & Ostfeld, 2008). These models can characterise a system with limited assumptions and are useful in solving practical problems, especially when there is lack of understanding of the underlying physical process, the time series are of
insufficient length, or when existing models are inadequate (Solomatine et al., 2008b; Napolitano et al., 2011).

Figure 8-1 An aerial view of the City of Calgary, Canada showing the locations of (a) the flow monitoring site Bow River at Calgary (Water Survey of Canada ID: 05BH004), three wastewater treatment plants at (b) Bonnybrook, (c) Fish Creek, and (d) Pine Creek, and two water quality sampling sites (e) Stier’s Ranch and (f) Highwood.
8.1.1 Fuzzy numbers and data-driven modelling

Possibility theory is an information theory that is an extension of fuzzy sets theory for representing uncertain, vague or imprecise information (Zadeh, 1978). Fuzzy numbers are an extension of fuzzy set theory, and express an uncertain or imprecise quantity. These types of numbers are particularly useful for dealing with uncertainties when data are limited or imprecise (Bárdossy et al., 1990; Guyonnet et al., 2003; Huang et al., 2010; Zhang & Achari 2010a) – in other words when epistemic uncertainty exists. This type of uncertainty is in contrast to aleatory uncertainty that is typically handled using probability theory. Possibility theory and fuzzy numbers are thus useful when a probabilistic representation of parameters may not be possible, since the exact values of parameters may be unknown, or only partial information is available (Zhang, 2009). Thus, the choice of using a data-driven approach in combination with possibility theory lends itself well to the constraints posed by the problem in the Bow River: the difficulty in correctly defining a physically-based model for a complex urban system and the use of possibility theory to model the uncertainty in the system when probability theory based methods may be inadequate.

Data-driven models, such as neural networks, regression-based techniques, fuzzy rule-based systems, and genetic programming, have seen widespread use in hydrology, including DO prediction in rivers (Shrestha & Solomatine, 2008; Solomatine et al., 2008b; Elshorbagy et al., 2010). Wen et al. (2011) used artificial neural networks (ANN) to predict DO in a river in China using ion concentration as the predictors. Antanasijević et al., (2014) used ANNs to predict DO in a river in Serbia using a Monte Carlo approach to quantify the uncertainty in model predictions and temperature as a predictor. Chang et al., (2015) also used ANNs coupled with hydrological factors (such as precipitation and discharge) to predict DO in a river in Taiwan. Singh et al., (2009) used water quality parameters to predict DO and BOD in a river in India. Other studies (e.g. Heddam, 2014 and Ay & Kisi, 2012,) have used regression to predict DO in rivers using water temperature, or electrical conductivity, amongst others, as inputs. In general, these studies have demonstrated that there is a need and demand for less complex DO models, has led to an increase in the popularity of data-driven models (Antanasijević et al., 2014), and that the performance of these models is suitable. Recent research into predicting DO
concentration in the Bow River in Calgary using abiotic factors (these are non-living, physical and chemical attributes) as inputs have shown promising results (He et al., 2011; Khan et al., 2013; Khan & Valeo, 2015a). The advantage of using readily available data (i.e. the abiotic inputs) in these studies is that if a suitable relationship between these factors and DO can be found, changing the factors (e.g. increasing the flow rate downstream of a treatment plant) can potentially reduce the risk of low DO.

While fuzzy set theory based applications, particularly applications using fuzzy logic in neural networks, have been widely used in many fields including hydrology (Bárdossy et al., 2006; Abrahart et al., 2010), the use of fuzzy numbers and possibility theory based applications has been limited in comparison (Bárdossy et al., 2006; Jacquin, 2010). Some examples include maps of soil hydrological properties (Martin-Clouaire et al., 2000), remotely sensed soil moisture data (Verhoest et al., 2007), climate modelling (Mujumdar and Ghosh, 2008), subsurface contaminant transport (Zhang et al., 2009), and streamflow forecasting (Alvisi & Franchini, 2011). Khan et al. (2013) and Khan & Valeo (2015a) have introduced a fuzzy number based regression technique to model daily DO in the Bow River using abiotic factors with promising results. Similarly, Khan & Valeo (2014a) used an autoregressive time series based approach combined with fuzzy numbers to predict DO in the Bow River. In these studies the use of fuzzy numbers meant that the uncertainty in the system could be quantified and propagated through the model. However, due to the highly non-linear nature of DO modelling, the use of an ANN based method is of interest since these types of models are effective for modelling complex, nonlinear relationships without the explicit understanding of the physical phenomenon governing the system. A fuzzy neural network method proposed by Alvisi & Franchini (2011) for streamflow prediction that uses fuzzy weights and biases in the network, is further refined in this chapter for predicting DO concentration.

8.1.2 Chapter objectives

Given the importance of DO concentration as an indicator of overall aquatic ecosystem health, there is a need to accurately model and predict DO in urban riverine environments, like that in Calgary, Canada. In this chapter a new data-driven method is proposed that attempts to address the issues that plague numerical modelling of DO concentration in the Bow River. The FNN method proposed by Alvisi & Franchini (2011)
is adapted and extended in two critical ways. The existing method uses crisp (i.e. non-fuzzy) inputs and outputs to train the network, producing a set of fuzzy number weights and biases, and fuzzy outputs. The method is adapted to be able to handle fuzzy number inputs to produce fuzzy weights and biases, and fuzzy outputs. The advantage is that the uncertainties in the input observations are also captured within the model structure. To do this, a new method of creating fuzzy numbers from observations is presented based on a probability-possibility transformation. Second, the existing training algorithm is based on capturing a predetermined set of observations (e.g. 100%, 95% or 90%) within the fuzzy outputs. The selection of the predetermined set of observations in the original study was an arbitrary selection. A new method that exploits the relationship between possibility theory and probability theory is defined to create a more objective method of training the FNN. A consequence of this is that the resulting fuzzy number outputs from the model can then be directly used for risk analysis, specifically to quantify the risk of low DO concentration. This information is extremely valuable for managing water resources in the face of uncertainty.

Following previous research for this river, two abiotic inputs (daily mean water temperature, $T$ and daily mean flow rate, $Q$) will be used to predict daily minimum DO. An advantage of using these factors is that they are routinely collected by the City of Calgary, and thus, a large dataset is available. Also, their use in previous studies has shown that they are good predictors of daily DO concentration in this river basin (He et al., 2011, Khan et al., 2013, Khan and Valeo, 2015a). The following sections outline the background of fuzzy numbers and existing probability-possibility transformations. This is followed by the development of the new method to create fuzzy numbers from observations. Then, the new FNN method using fuzzy inputs is developed mathematically using new criteria for training, also based on possibility theory. Lastly, a method to measure the risk of low DO is described.
8.2 Methods\textsuperscript{14}

8.2.1 Data collection
The Bow River is 645 km long and averages a 0.4% slope over its length (BRBC, 2010a) from its headwaters at Bow Lake in the Rocky Mountains to its confluence with the Oldman River in Southern Alberta, Canada (Robinson et al., 2009; Environment Canada, 2014). The river is supplied by snowmelt from the Rocky Mountains, rainfall and discharge from groundwater. The City of Calgary is located within the Bow River Basin and the river has an average annual discharge of 90 m\textsuperscript{3}/s, an average width and depth of 100 m and 1.5 m, respectively (Khan & Valeo, 2014b; 2015b).

The City of Calgary routinely samples a variety of water quality parameters along the Bow River to measure the impacts of urbanisation, particularly from three wastewater treatment plants and numerous stormwater runoff outfalls that discharge into the River. DO concentration measured upstream of the City is generally high throughout the year, with little diurnal variation (He et al., 2011; Khan et al., 2013; Khan & Valeo, 2015a). The DO concentration downstream of the City is lower and experiences much higher diurnal fluctuation. The three wastewater treatment plants are located upstream of this monitoring site, and are thought to be responsible, along with other impacts of urbanisation, for the degradation of water quality (He et al., 2015).

For this research, nine years of DO concentration data was collected from one of the downstream stations from 2004 to 2012. The monitoring station was located at Pine Creek and sampled water quality data every 30 minutes (from 2004 to 2005), and every 15 minutes (from 2006 to 2007). The station was then moved to Stier's Ranch and sampled data every hour (in 2008) and every 15 minutes (2009 to 2011). The monitoring site was moved further downstream to its current location (at Highwood) in 2012 where it samples every 15 minutes. During this period a number of low DO events have been observed in the River and are summarised below in Table 8-1 corresponding to different water quality guidelines.

Note that even though daily minimum DO was observed to be below 5 mg/L on several occasions in 2004 and 2006 (in Table 8-1), the minimum DO was below 9.5 mg/L only

\textsuperscript{14} The MATLAB code associated with this section is included in Appendix L.
107 and 164 days, respectively, for those two years. In contrast, in 2007 and 2010, no observations below 5 mg/L are seen yet 182 and 180 days, respectively, below the 9.5 mg/L guideline were seen for those years. The total amount of days below 9.5 mg/L constitute approximately 90% of all observations for those years. This highlights that despite no DO events below 5 mg/L, generally speaking DO concentration was quite low in these two years. The implication of this is that only using one guideline for DO might not be a good indicator of overall aquatic ecosystem health.

Table 8-1 A summary of low DO events in the Bow River between 2004 and 2012 and the corresponding minimum acceptable DO concentration guidelines

<table>
<thead>
<tr>
<th>Year</th>
<th>DO &lt; 5 mg/L&lt;sup&gt;a&lt;/sup&gt;</th>
<th>DO &lt; 6.5 mg/L&lt;sup&gt;b&lt;/sup&gt;</th>
<th>DO &lt; 9.5 mg/L&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>25</td>
<td>41</td>
<td>107</td>
</tr>
<tr>
<td>2005</td>
<td>1</td>
<td>26</td>
<td>133</td>
</tr>
<tr>
<td>2006</td>
<td>25</td>
<td>70</td>
<td>164</td>
</tr>
<tr>
<td>2007</td>
<td>0</td>
<td>27</td>
<td>182</td>
</tr>
<tr>
<td>2008</td>
<td>0</td>
<td>5</td>
<td>130</td>
</tr>
<tr>
<td>2009</td>
<td>0</td>
<td>15</td>
<td>85</td>
</tr>
<tr>
<td>2010</td>
<td>0</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>2011</td>
<td>0</td>
<td>0</td>
<td>122</td>
</tr>
<tr>
<td>2012</td>
<td>0</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>184</td>
<td>1179</td>
</tr>
</tbody>
</table>

<sup>a</sup> for the protection of aquatic life for 1-day (AENV, 1997)
<sup>b</sup> for the protection of aquatic life in cold, freshwater for other-life (i.e. not early) stages (CCME, 1999)
<sup>c</sup> for the protection of aquatic life in cold, freshwater for early-life stages (CCME, 1999)

A YSI sonde is used to monitor DO and T, and the sonde is not accurate in freezing water, thus only data from the ice free period was considered, which is approximately from April to October for most years (YSI Inc., 2014). Since low DO events usually occur in the summer (corresponding to high water temperature and lower discharge), the ice-free period dataset contains the dates that are of interest for low DO modelling.

Daily mean flow rate, Q, was collected from the Water Survey of Canada site “Bow River at Calgary” (ID: 05BH004) for the same period. This data is collected hourly throughout the year, thus, data where considerable shift corrections were applied (usually
due to ice conditions) were removed from the analysis. The mean annual water temperature ranged between 9.23 and 13.2 °C, and the annual mean flow rate was between 75 and 146 m$^3$/s, for this period.

8.2.2 Probability-possibility transformations

Fuzzy sets were proposed by Zadeh (1965) in order to express imprecision in complex systems, and can be described as a generalisation of classical set theory (Khan & Valeo, 2015a). In classical set theory, an element $x$ either belongs or does not belong to a set $A$. In contrast, using fuzzy set theory, the elements $x$ have a degree of membership, $\mu$, between 0 and 1 in the fuzzy set $A$. If $\mu$ equals 0, then $x$ does not belong in $A$, and $\mu = 1$ means that it completely belongs in $A$, while a value $\mu = 0.5$ means that it is only a partial member of $A$.

Fuzzy numbers express uncertain or imprecise quantities, and represent the set of all possible values that define a quantity rather than a single value. A fuzzy number is defined as a specific type of fuzzy set: a normal and convex fuzzy set. Normal implies that there is at least one element in the fuzzy set with a membership level equal to 1, while convex means that the membership function increases monotonically from the lower support (i.e., $\mu = 0^L$) to the modal element (i.e. the element(s) with $\mu = 1$) and then monotonically decreases to the upper support (i.e., $\mu = 0^R$) (Kaufmann & Gupta, 1985).

Traditional representation of a fuzzy numbers has been using symmetrical, linear membership functions, typically denoted as “triangular fuzzy” numbers. The reason for selecting this type of membership function has to do with its simplicity: given that a fuzzy number must, by definition, be convex and normal, a minimum of three elements are needed to define a fuzzy number (two elements at $\mu = 0$ and one element at $\mu = 1$). For example, if the most credible value for DO concentration is 10 mg/L ($\mu = 1$), with a support about the modal value between ($\mu = 0^L$) and 12 mg/L ($\mu = 0^R$). This implies that the simplest membership function is triangular, though not necessarily symmetrical. Also, as we demonstrate below, in some probability-possibility frameworks, a triangular membership function corresponds to a uniform probability distribution – the least specific distribution in that any value is equally probable and hence, represents the most uncertainty (Dubois & Prade, 2015; Dubois et al., 2004).
However, recent research (Khan et al., 2013; Khan & Valeo, 2014a; 2014b; 2015a; 2015b) has shown that such a simplistic representation may not be appropriate for hydrological data, which is often skewed, and non-linear. This issue is further highlighted if the probability-possibility framework mentioned above is used: it implies that for a triangular membership function, the fuzzy number bounded by the support [8 12] mg/L, has a uniform probability distribution bounded between 8 and 12 mg/L with a mean value of 10 mg/L, suggesting that values between the support are equally likely to occur. It not difficult to see that this an over-simplification of hydrological data, which often have skewed, non-symmetrical distributions. In many cases enough information (i.e. from observations) is available to define the membership function with more specificity, and this information should be used to define the membership function.

Multiple frameworks exist to transform a probability distribution to a possibility distribution, and vice versa; a comparison of different conceptual approaches are provided in Klir & Parvais (1992), Oussalah (2000), Jaquin (2010) Mauris (2013) and Dubois & Prade (2015). However, a major issue of implementing fuzzy number based methods in hydrology is that there is no consistent, transparent and objective method to convert observations (e.g. time series data) into fuzzy numbers, or generally speaking to construct the membership function associated with fuzzy values (Abrahart et al., 2010; Dubois & Prade, 1993; Civanlar & Trussel, 1986).

A popular method (Dubois et al., 1993; 2004) converts a probability distribution to a possibility distribution by relating the area under a probability density function to the membership level (Zhang, 2009). From this point of view, the possibility is viewed as the upper envelope of the family of probability measures (Jacquin, 2010; Ferrero et al., 2013; Betrie et al., 2014). There are two important considerations for this transformation, first it guarantees that something must be possible before it is probable; hence, the degree of possibility cannot be less than the degree or probability – this is known as the consistency principle (Zadeh, 1965). Second is order preservation, which means if the possibility of $x_i$ is greater than the possibility of $x_j$ then the probability of $x_i$ must be greater than the probability of $x_j$ (Dubois et al., 2004). For a discrete system, this can be represented as:

if $p(x_1) > p(x_2) > \ldots > p(x_n)$,

then the possibility distribution of $x (\pi(x))$, follows the same order, that is:
The transformation is given by:

For \( p(x_1) > p(x_2) > \ldots > p(x_n) \):

\[
\pi(x_1) = 1
\]

\[
f(x) = \begin{cases} 
\sum_{j=i}^{n} p_j, & \text{if } p_{i-1} > p_i \\
\pi(x_{i-1}), & \text{else}
\end{cases}
\]

Equation 8-1

where the \( x_i \) are elements of a fuzzy number \( A \), \( \pi(x_i) \) is the possibility of element \( x_i \), and \( p(x_i) \) is the probability of element \( x_i \). The concept of this transformation may be more illustrative when viewed in the continuous case: for any interval \([a, b]\), the membership level \( \mu \) (where \( \pi(a) = \pi(b) = \mu \)) is equal to the sum of the areas under the probability density function curve between \((-\infty, a)\) and \((b, \infty)\) (Zhang et al., 2009). It is important to highlight that this particular transformation has an inverse transformation associated with, where a probability distribution can be estimated from the possibility distribution.

However, a major drawback of this transformation is that it theoretically requires a full description of the probability density function, or in the finite case, the probability associated with each element of the fuzzy number, the probability mass function. For many hydrology and time series based applications this might not be possible because the hourly data (that is typically collected) may not adequately fit the mould of a known class of probability density functions, or one distribution amongst many alternatives may have to be selected based on best-fit. This best-fit function may not be universal, e.g. data from one 24 hour period may be best described by one class or family of probability density function, while the next day by a completely different class of density function. This means working with multiple classes of distribution functions for one application, which can be cumbersome. Also, given that each day may only have 24 data points (or fewer on days with missed samples) it is difficult to select one particular function.

In previous research by Khan & Valeo (2015a), a new approach to create a fuzzy number based on observations was developed. This process used a histogram-based approach to estimate the probability mass function of the observations, and then Equation 8-1 was used to estimate the membership function of the fuzzy number. To create the
histogram, the bin-size was selected based on the extrema observations for a given day and the number of the observations. A linear interpolation scheme was then used to calculate the values of the fuzzy number at five predefined membership levels. This method has a few shortcomings, namely: first, the bin-size selection was arbitrarily selected based on the magnitude and number of observations which does not necessarily result in the optimum bin-size. This lack of optimality means that the resulting histogram may either be too smooth so as not to capture the variability between membership levels, or too rough and uneven so that the underlying shape of the membership function is difficult to discern. This is a common issue with histogram selection in many applications (Shimazaki & Shinomoto, 2007). Secondly, the aforementioned transformation used by Khan and Valeo (2015a) only allows one element to have $\mu = 1$ when $p(x)$ is maximum. However, there are a number of cases (e.g. bimodal distributions, or arrays when all elements are equal) where multiple elements have joint-equal maximum $p(x)$), and hence multiple elements with $\mu = 1$. This means that all elements within the $\alpha$-cut interval $[a \ b]_{\mu=1}$ (where $a$ and $b$ are the minimum and maximum elements with $\mu = 1$) must by definition also have a membership level equal to 1. Thus, a method is necessary to be flexible enough to accommodate these issues.

In this chapter, a two-step procedure is proposed to create fuzzy numbers on the inputs (i.e. $Q$ and $T$) using hourly (or sub-hourly) observations. First, a bin-size optimisation method is used (an extension of an algorithm proposed by Shimazaki & Shinomoto, 2007) to create histograms to represent the estimate of discretised probability density functions of the observations. This estimate of the probability distribution is then transformed to the membership function of the fuzzy number using a new numerical procedure and the transformation principles described in Equation 8-1. This updated method requires no assumptions regarding the distribution of the underlying data or selection of an arbitrary bin-size, has the flexibility to create different shapes of fuzzy numbers depending on the distribution of the underlying data, and allows multiple elements to have equal $\mu = 1$. The proposed algorithm is described in the proceeding section.
8.2.2.1 A new algorithm to create fuzzy numbers

Shimazaki & Shinomoto (2007) proposed a method to find the optimum bin-size of a histogram when the underlying distribution of the process is unknown. The basic premise of the method is that the optimum bin-size \( D_{\text{opt}} \) is one that minimises the error between the theoretical (but unknown) probability density and the histogram generated using the \( D_{\text{opt}} \). The error metric used by Shimazaki & Shinomoto (2007) is the mean integrated squared error (\( E_{\text{MISE}} \)) which is frequently used for density estimation problems. It is defined as:

\[
E_{\text{MISE}} = \frac{1}{P} \int_0^P \mathbb{E}[f_n(t) - f(t)]^2 dt \quad \text{Equation 8-2}
\]

where \( f(t) \) is the unknown density function, \( f_n(t) \) is the histogram estimate of the density function, \( t \) denotes time and \( P \) is the observation period, and \( \mathbb{E}[\cdot] \) is the expectation. In practice, \( E_{\text{MISE}} \) cannot be directly calculated since the underlying distribution is unknown and thus, an estimate of the \( E_{\text{MISE}} \) is used in its place (see \( C_D \) below). Thus, \( f_n(t) \) can be found without any assumptions of the type of distribution (e.g. class, unimodality, etc.); the only assumption is that the number of events (i.e., the counts \( k_i \)) in the \( i^{\text{th}} \) bin of the histogram follow a Poisson point process. This means that the events in two disjoint bins (e.g., the \( i^{\text{th}} \) and \( i+1^{\text{th}} \) bin) are independent, and that mean (\( k \)) and variance (\( v \)) of the \( k_i \) in each bin are equal due to the assumption of a Poisson process (Shimazaki & Shinomoto, 2007).

Using this property, the optimum bin-size can be found as follows. Let \( X \) be the input data vector for the observation period \( (P) \), e.g., a \([24\times1]\) vector corresponding to hourly samples for a given day. The elements in \( X \) are binned into \( N \) bins of equal bin-size \( D \). The number of events \( k_i \) in each \( i^{\text{th}} \) bin are then counted and the mean (\( k \)) and variance (\( v \)) of the \( k_i \) are calculated as follows:

\[
k = \frac{1}{N} \sum_{i=1}^{N} k_i \quad \text{Equation 8-3}
\]

\[
v = \frac{1}{N} \sum_{i=1}^{N} (k_i - k)^2 \quad \text{Equation 8-4}
\]
The \( k \) and \( v \) are then used to compute the \textit{cost-function} \( C_D \), which is defined as:

\[
C_D = \frac{2k - v}{D^2}
\]

Equation 8-5

This cost-function is a variant of the original \( E_{MISE} \) listed in Equation 8-2 and is derived by removing the terms from \( E_{MISE} \) that are independent of the bin-size \( D \), and by replacing the unobservable quantities (i.e. \( E[f(t)] \)) with their unbiased estimators (details of this derivation can be found in the original paper by Shimazaki & Shinomoto, 2007). The objective then is to search for \( D_{opt} \): the value of \( D \) that minimises \( C_D \). To do this two systematic modification are made: first, \( C_D \) is recalculated at different \textit{partitioning positions}, and secondly, the entire process is repeated for different values of \( N \) and \( D \), until a “reliable” estimate of minimum \( C_D \) and thus \( D_{opt} \) is found. Using different partitioning positions means that the variability in \( k_i \) resulting from the position of the bin (rather than the size of the bin) can be quantified. Repeating the analysis at different \( N \) and \( D \) accounts for the variability due to different bin-sizes. Both these techniques are ways of accounting for the uncertainty associated with estimating the histogram.

Partitioning positions are defined as the first and last point that define a bin. The most common way of defining a partitioning position is to centre it on some value \( a \), e.g. the bin defined at \([a-D/2, a+D/2]\) is centred on \( a \) and has a bin-size \( D \). Variations of this partitioning position can be found by using a moving-window technique, where the bin-size \( D \) is kept constant, but the first and last points are perturbed by a small value \( \delta \): \([a-D/2+\delta, a+D/2+\delta]\), where \( \delta \) ranges incrementally between 0 and \( D \). Using these different values of \( \delta \) whilst keeping \( D \) constant will result in different values of \( k_i \) and hence unique values of \( C_D \). Thus, for a single value of \( D \), multiple values of \( C_D \) are possible.

For this research this bin-size optimisation algorithm is implemented to determine the optimum histogram for the two input variables, \( Q \) and \( T \). The array of daily data, \( X \), (at hourly or higher frequency, see Sect. 2.1 for details regarding the sampling frequency of both inputs) for each variable was collected for the nine year period. The bin-size was calculated for each day as follows:
where the $x_{\text{max}}$ and $x_{\text{min}}$ are the maximum and minimum sampled values for $X$, respectively, and $N$ is the number of bins. As described above, a number of different $D$ were considered to find the optimum $C_D$. This was done by selecting a number of different values of $N$, ranging from $N_{\text{min}}$ to $N_{\text{max}}$. The minimum value $N_{\text{min}}$, was set at 3 for all days; this is the necessary number of bins to define a fuzzy number (two elements for $\mu = 0$, and one element for $\mu = 1$). The highest value, $N_{\text{max}}$ was calculated as:

$$N = \frac{x_{\text{max}} - x_{\text{min}}}{2r} \quad \text{Equation 8-7}$$

where $2r$ is the measurement resolution of the device used to measure either $Q$ or $T$, set at twice the accuracy ($r$) of the device. The rational for this decision is that as $N$ increases $D$ necessarily decreases (as per Equation 8-6). However, $D$ cannot be less than the measurement resolution; this constraint (i.e. $N \leq N_{\text{max}}$) ensures that the optimum bin-size is never less than what the measurement devices can physically measure. For this research, the accuracy for $T$ is listed as ±0.1 °C, and thus, the resolution ($2r$) is 0.2°C (YSI Inc., 2014). For $Q$ all measurements below 99 m$^3$/s have an accuracy of ±0.1 m$^3$/s and thus, a resolution of 0.2 m$^3$/s, while measurements above 99 m$^3$/s have an accuracy of ±1 m$^3$/s, and thus a resolution of 2 m$^3$/s. This is based on the fact that all data provided by the Water Survey of Canada is accurate to three significant figures. Note that for the case where $x_{\text{min}}$ equals $x_{\text{max}}$ (i.e. no variance in the daily observed data) then $D = 2r$, which means that the only uncertainty considered is due to the measurement.

Once the $N_{\text{max}}$ is determined, the bin-size $D$ was calculated for each $N$ between $N_{\text{min}}$ and $N_{\text{max}}$. Then, starting at the largest $D$ (i.e. $D = (x_{\text{max}} - x_{\text{min}})/N_{\text{min}}$), the cost-function $C_D$ is calculated at the first partitioning positing, where the first bin is centred at $x_{\min}$, $[(x_{\min} - D/2), (x_{\min} + D/2)]$, and the $N^{\text{th}}$ bin is centred on $x_{\text{max}}$, $[(x_{\text{max}} - D/2), (x_{\text{max}} + D/2)]$. Then, $C_D$ is calculated at the next partitioning positing, where the first bin is $[(x_{\min} - D/2 + \delta), (x_{\min} + D/2 + \delta)]$, and the $N^{\text{th}}$ bin is $[(x_{\text{max}} - D/2 + \delta), (x_{\text{max}} + D/2 + \delta)]$. The value of $\delta$ ranged between 0 and $D$ at $(D/100)$ intervals. Thus, for this value of $D$, 100 values of $C_D$ were
calculated since 100 different partitioning positions were used. The mean value of these $C_D$ was used to define the final cost-function value for the given $D$.

This process is then repeated for the next $N$ between $N_{\text{min}}$ and $N_{\text{max}}$, using the corresponding $D$ at 100 different partitioning positions, and so on until the smallest $D$ (at $N_{\text{max}}$). This results in $[N_{\text{max}}-N_{\text{min}}]$ values of mean $C_D$: the value of $D$ corresponding to the minimum value of $C_D$ is considered to be the optimum bin-size $D_{\text{opt}}$. This $D_{\text{opt}}$ is then used to construct the optimum histogram of each daily observation. This histogram can be used to calculate a discretised probability density function ($p(x)$), where for each $x$ (an element of $X$), the $p(x)$ is calculated by dividing the number of events in each bin by the total number of elements in $X$. The $x$ and $p(x)$ can then be used to calculate the possibility distribution using the transformation described in Equation 8-1.

First, the $p(x)$ are ranked from highest to lowest, and the $x$ corresponding to the highest $p(x)$ is has a membership level of 1. Then the $\pi(x)$ values for the remaining $x$ are calculated using Equation 8-1. For cases where multiple elements have equal $p(x)$, the highest $\pi(x)$ is assigned to each $x$. For example, if $p(x_i) = p(x_j)$, and $x_i > x_j$, then $\pi(x_i) = \pi(x_j)$. This means that in some cases, for each calculated membership level, $\pi(x)$, there exists an $\alpha$-cut interval $[a, b]_{\mu=\pi(x)}$ where all the elements between $a$ and $b$ have equal $p(x)$ and hence equal $\pi(x)$. By definition of $\alpha$-cut intervals, all values of $x$ within the interval $[a, b]$ have at least a possibility of $\pi(x)$. A special case of this occurs when multiple $x$ have joint-equal maximum $p(x)$, meaning that multiple elements have a membership level of $\mu = 1$. Thus, an $\alpha$-cut interval is created for the $\mu = 1$ case, creating a trapezoidal membership function, where the modal value of the fuzzy number is defined by an interval rather than a single element.

Once all the $\pi(x)$ are calculated for each element $x$ in $X$, a discretised empirical membership function of the fuzzy number $X$ can be constructed using the calculated $\alpha$-cut intervals. That is, the fuzzy number is defined by a number of intervals at different membership levels. The upper and lower limit of the intervals at higher membership levels define the extent of the limits of the intervals at lower membership levels. This way the constructed fuzzy numbers maintain convexity (similar to a procedure used by Alvisi & Franchini, 2011), where the widest intervals have the lowest membership level.
For example, the interval at $\mu = 0.2$ will contain the interval $\mu = 0.4$, and this interval will contain the interval at $\mu = 0.8$.

In creating this discretised empirical membership function this way (rather than assuming a shape of the function) means that this function best reflects the possibility distribution of the observed data. However, it also means that all fuzzy numbers created using this method are not guaranteed to be defined at the same $\pi(x)$, nor have an equal number of $\pi(x)$ intervals used to define the fuzzy number. Thus, direct fuzzy arithmetic between multiple fuzzy numbers using the extension principle is not possible since it requires each fuzzy number to be defined at the same $\alpha$-cut intervals (Kaufmann & Gupta, 1985). Thus, linear interpolation is used to define each fuzzy number at a pre-set $\alpha$-cut interval using the empirical $\pi(x)$ calculated using the transformation. To select the pre-set $\alpha$-cut intervals it is illustrative to see the impact of selecting two extreme cases: (i) if only two levels are selected (specifically $\mu = 0$ and 1) the constructed fuzzy number will reduce to a triangular fuzzy number. As discussed above there are important implications of using triangular membership functions that make it undesirable for hydrological data; (ii) if a large number of intervals (e.g. 100 intervals between $\mu = 0$ and 1) are selected, there is a risk that the number of pre-set intervals is much larger than the empirical $\pi(x)$, which means not enough data (empirical $\alpha$-cut levels intervals) to conduct interpolation, leading to equal interpolated values at multiple $\alpha$-cut levels. For this research, results (discussed in the following section) of the bin-size optimisation showed that most daily observations for $T$ and $Q$ resulted in 2 to 10 unique $p(x)$ values. Based on this, six pre-set $\alpha$-cut intervals were selected: 0, 0.2, 0.4, 0.6, 0.8 and 1. The empirical $\pi(x)$ can then be converted to a standardised function at pre-defined membership levels using linear interpolation.

8.2.3 Fuzzy neural networks

8.2.3.1 Background on artificial neural networks

Artificial neural networks (ANN) are a type of data-driven model that are defined as a massively parallel distributed information processing system (Elshorbagy et al., 2010; Wen et al., 2013). As a predictive model, ANNs can capture complex, nonlinear relationships that may exist between variables without the explicit understanding of the physical phenomenon (Alvisi & Franchini, 2011; Antanasijević et al., 2014). This has
resulted in significant use of ANN models in hydrology when the complexity of the physical systems is high owing partially to an incomplete understanding of the underlying process, and the lack of availability of necessary data (He et al., 2011; Kasiviswanathan et al., 2013). Further, ANNs arguably require less data and do not require an explicit mathematical description of the underlying physical process (Antanasijević et al., 2014), making it a simpler and practical alternative to traditional modelling techniques.

Multilayer Perceptron (MLP) is a type of feedforward ANN and is one of the most commonly used in hydrology (Maier et al., 2010). One of the reasons for the popularity of MLPs is that a trained network can be used as universal approximators with only a single hidden layer (Hornik et al., 1989). This means that models are relatively simple to develop, and theoretically have the capacity of approximating any linear or nonlinear mapping (ASCE 2000; Elshorbagy et al., 2010; Napolitano et al., 2011; Kasiviswanathan et al., 2013). Further, the popularity of MLP has meant that subsequent research has continued to use MLP (He & Valeo 2009; Napolitano et al., 2011) and thus, form a reference for the basis of comparing ANN performance (Alvisi & Franchini, 2011).

In the simplest case, an MLP consists of an input layer, a hidden layer, and an output layer as shown in Figure 8-2. Each layer consists of a number of neurons (or nodes) that each receive a signal, and on the basis of the strength of the signal, emit an output. Thus, the final output layer is the synthesis and transformation of all the input signals from both the input and the hidden layer (He & Valeo, 2009).

The number of neurons in the input ($n_I$) and output ($n_O$) layers corresponds to the number of variables used as the input and the output, respectively and the number of neurons in the hidden layer ($n_H$) are selected based on the relative complexity of the system (Elshorbagy et al., 2010). A typical MLP is expressed mathematically as follows:

$$y_i = f_{\text{HID}}(W_{IH}x_i + B_H)$$  \hspace{1cm} \text{Equation 8-8}

$$z_i = f_{\text{OUT}}(W_{HO}y_i + B_O)$$  \hspace{1cm} \text{Equation 8-9}

where $x_i$ is the $i^{th}$ observation (an $n_I$ x 1 vector) from of a total of $n$ observations, $W_{IH}$ is a $n_H$ x $n_I$ matrix of weights between the input and hidden-layer, $B_H$ is a vector ($n_H$ x 1) of biases in the hidden-layer, and $y_i$ is the $i^{th}$ output (an $n_H$ x 1 vector) of the input signal
through the hidden-layer transfer function, $f_{HID}$. Similarly, $W_{HO}$ is an $n_O \times n_H$ matrix of weights between the hidden and output-layers, $B_O$ is an $n_O \times 1$ vector of biases in the output-layer, and $f_{OUT}$ the final transfer function to generate the $i^{th}$ modelled output $z_i$ (an $n_O \times 1$ vector).

The values of all the weights and biases in the MLP are calculated by training the network by minimising the error – typically mean squared error ($E_{MSE}$) (He & Valeo, 2009) – between the modelled output and the target data (i.e. observations). A number of training algorithms can be used, and one of the most common methods is the Levenberg–Marquardt algorithm (LMA) (Alvisi et al., 2006). In this method, the error between the output and target is back-propagated through the model using a gradient method where the weights and biases are adjusted in the direction of maximum error reduction. The LMA is well-suited for ANN problems that have a relatively small number of neurons. To counteract potential over-fitting issues, an early-stopping procedure is used (Alvisi et al., 2006; Maier et al., 2010), which is a form of regularisation where the data is split into three subsets (for training, validation and testing) and the training is terminated when the error on the validation subset increases from the previous iteration. Most ANNs have a deterministic structure without a quantification of the uncertainty corresponding to the predictions (Alvisi & Franchini, 2012; Kasiviswanathan & Sudheer, 2013). This means that users of these models may have excessive confidence in the forecasted values, and misinterpret the applicability of the results (Alvisi & Franchini, 2011). This lack of uncertainty quantification is one reason for the limited appeal of ANN by water resource managers (Abrahart et al., 2012; Maier et al., 2010). Without this characterisation, the results produced by these models have limited value (Kasiviswanathan & Sudheer, 2013).

In this chapter, two methods are proposed to quantify the uncertainty in MLP modelling to predict DO in the Bow River. First, the uncertainty in the input data (daily mean water temperature and daily mean flow rate) is represented through the use of fuzzy numbers. These fuzzy numbers are created using the probability-possibility transformation discussed in the previous section. Second, the total uncertainty (as defined by Alvisi & Franchini, 2011) in the weights and biases of an MLP are quantified using a new possibility theory-based FNN. The total uncertainty represents the overall uncertainty in the modelling process, and not of the individual components (e.g.
randomness in observed data). The following section describes the proposed FNN method.

Figure 8-2 An example of a three-layer multilayer perceptron feed-forward ANN, with two input neurons, three hidden layer neurons, and one output neuron.

8.2.3.2 FNN with fuzzy inputs and possibility-based intervals
Alvisi & Franchini (2011) proposed a method to create a FNN, where the weights and biases, and by extension the output, of the neural network are fuzzy numbers rather than crisp (non-fuzzy) numbers. These fuzzy numbers quantify the total uncertainty of the calibrated parameters. While fuzzy set theory based applications of ANN have been limited in hydrology, most have used fuzzy logic, e.g. the widely used Adaptive Neuro-Fuzzy Inference System, where automated IF-THEN rules are used to create crisp outputs (Abrahart et al., 2010; Alvisi & Franchini, 2011). Thus, the advantage of fuzzy outputs is
that it provides the uncertainty of the predictions as well, while the fuzzy parameters reflect the uncertainty in the model structure. This uncertainty quantification can be used to by end users to assess the value of the model output.

In their FNN, the MLP model is presented in Equation 8-5 and Equation 8-6 is modified to predict an interval rather than a single value for the weights, biases and output, corresponding to an \( \alpha \)-cut interval (at a defined membership level \( \mu \)). This is repeated for several \( \alpha \)-cut levels, thus building a discretised fuzzy number at a number of membership levels. This is done by using a stepwise, constrained optimisation approach:

\[
\begin{align*}
[y^L_i, y^U_i] &= f_{\text{HID}}([W^L_{IH} W^U_{IH}] x_i + [B^L_{IH} B^U_{IH}]) & \text{Equation 8-10} \\
[z^L_i, z^U_i] &= f_{\text{OUT}}([W^L_{HO} W^U_{HO}] \times [y^L_i y^U_i] + [B^L_{HO} B^U_{HO}]) & \text{Equation 8-11}
\end{align*}
\]

where all the variables are as described as before, and the superscripts U and L represent the upper and lower limits of the \( \alpha \)-cut interval, respectively. The constraints are defined so that the upper and lower limits of each weight and bias (in both layers) minimise the width of the predicted interval:

\[
\min \left( \sum_{i=1}^{n} (z_i^L - z_i^U) \right)
\]

subject to:

\[
\frac{1}{n} \sum_{i=1}^{n} \delta_i > P_{CI}
\]

where \( \delta_i = \begin{cases} 1, & \text{if } z_i^L < t_i < z_i^U \\ 0, & \text{otherwise} \end{cases} \)

where \( t \) is that target (observed data) and \( P_{CI} \) is a predefined percentage of data. Alvisi & Franchini (2011) defined P to be 100% at \( \mu = 0 \), 99% at \( \mu = 0.25 \), 95% at \( \mu = 0.5 \) and 90% at \( \mu = 0.75 \). This algorithm was built starting at \( \mu = 0 \) and moving to higher membership levels to maintain convex membership functions of the generated fuzzy numbers by using the results of the previous optimisation as the upper and lower limit constraints for the proceeding optimisation. Lastly, at \( \mu = 1 \), the interval collapses to a singleton, represent the crisp results from non-fuzzy ANN. Therefore, these \( \alpha \)-cut
intervals of the FNN output quantify the uncertainty around the crisp prediction, within which is expected to contain $P_{CI}$ percentage of data.

In this chapter, this method is modified in two ways. First, the inputs $x$ are also fuzzy numbers, which means that Equation 8-10 and Equation 8-11 are revised as follows:

\[
[y^L_i, y^U_i] = f_{HID}([W_{IH}^L W_{IH}^U] \times [x^L_i, x^U_i] + [B^L_H B^U_H]) \quad \text{Equation 8-13}
\]

\[
[z^L_i, z^U_i] = f_{OUT}([W_{HO}^L W_{HO}^U] \times [y^L_i, y^U_i] + [B^L_O B^U_O]) \quad \text{Equation 8-14}
\]

Note that now the input vector is represented by its upper and lower limits. The major impact on this is that the training algorithm for the FNN needs to accommodate this fuzzy $\alpha$-cut interval, which requires the implementation of fuzzy arithmetic principles (Kaufmann & Gupta, 1985). The cost function for the optimisation remains unchanged.

The second modification of the original algorithm is related to the selection of the percent of data included in the predicted interval ($P_{CI}$). In the original, the selection is arbitrary and end-users of this method may be interested in the events that are not included in the selected $P_{CI}$. Thus, a full spectrum of possible values for a given prediction is required. Thus, the Alvisi & Franchini (2011) approach is further refined by utilising the same relationship between probability and possibility that was used to define the input fuzzy numbers, giving a more objective means of designing FNNs with fuzzy weights, biases and output.

In the adopted possibility-probability framework, the interval $[a, b]_\alpha$ created by the $\alpha$-cut at $\mu = \alpha$ implies that:

\[
[p(x < a) + p(x > b)] = \alpha \quad \text{Equation 8-15}
\]

This can be used to calculate the probability:

\[
[p(a < x < b)] = (1 - \alpha) \quad \text{Equation 8-16}
\]

This means that there is a probability of $(1 - \alpha)$ that the random variable $x$ falls within the interval $[a, b]_\alpha$. In other words the $\alpha$-cuts of a possibility distribution (at any $\mu$)
correspond to the \((1 - \alpha)\) confidence interval of the probability distribution of the same variable (Serrurier and Prade, 2013). This principle is used to select the different \(P\) for the optimisation constraints rather than the predetermined \(P_{CI}\) selected by Alvisi & Franchini (2011) and these are shown in Table 8-2.

<table>
<thead>
<tr>
<th>(\mu)</th>
<th>(P_{CI}) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.80</td>
<td>20.00</td>
</tr>
<tr>
<td>0.60</td>
<td>40.00</td>
</tr>
<tr>
<td>0.40</td>
<td>60.00</td>
</tr>
<tr>
<td>0.20</td>
<td>80.00</td>
</tr>
<tr>
<td>0.00</td>
<td>99.50</td>
</tr>
</tbody>
</table>

Table 8-2: Selected values for \(P_{CI}\) for the FNN optimisation

Note that for practical purposes, \(P_{CI}\) was selected as 99.50% at \(\mu = 0\) to prevent overfitting. The implication of this selection is that at \(\mu = 0\), nearly-all the observed data should fall within this predicted FNN interval, reflecting the highest uncertainty in the prediction. The uncertainty decreases as \(\mu\) increases. For the \(\mu = 1\) case the values of the weights and biases were determined to be the mid-point of the interval at \(\mu = 0.8\) to maintain convexity of the produced fuzzy numbers, and the difficulty in finding an interval containing 0% of the data.

8.2.3.3 Network architecture and implementation

For this research a three layer, feedforward MLP architecture was selected to model minimum daily DO (the output) using fuzzified daily flow rate \((Q)\) and fuzzified daily water temperature \((T)\) as the inputs. The three layers consist of an input layer, an output layer, and a hidden layer (with 5 neurons based on a trial-and-error search procedure). This architecture was selected for three reasons: it is one of the most commonly used in hydrology (Maier et al., 2010), it can be used as a universal approximator (Hornik et al., 1989), and as reference for comparing performance with previous research (He & Valeo 2009; Napolitano et al., 2011). In particular, a previous study modelling minimum DO in the Bow River used a three-layer MLP feedforward network (see He et al., 2011). Two transfer functions are required for FNN implementation: the hyperbolic tangent sigmoid
function was selected for $f_{\text{HID}}$, and a pure linear function for $f_{\text{OUT}}$. Both function selections follow Alvisi & Franchini (2011), Wen et al., (2012) and Elshorbagy et al., (2010), and are described as follows:

\begin{equation}
 f_{\text{HID}} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \\
 f_{\text{OUT}} = x
\end{equation}

The LMA method was used to train the network, minimising $E_{\text{MSE}}$. The input and output data was pre-processed before training, validating and testing: the data was normalised so that all input and output data fell within the interval [-1 1]. Further the data were randomly divided into training, validation and testing subsets, following a 50%-25%-25% split.

This FNN optimisation algorithm was implemented in MATLAB (version 2015a). First, the built-in MATLAB Neural Network Toolbox was used to estimate the value of weights and biases using the midpoint of the interval at $\mu = 1$. The results from this were used as the constraints to solve the FNN optimisation (Equation 8-12 to Equation 8-14) at subsequent lower membership levels. The Shuffled Complex Evolution algorithm (commonly known as SCE-UA, Duan et al., 1992) was used to find the optimisation solution. The optimisation is run such that the intervals at higher membership levels govern the upper and lower bounds of the predicted interval in order to preserve the convexity of fuzzy numbers.

8.2.4 Risk analysis using defuzzification

Risk analyses for complex systems is challenging for a number of reasons, including an insufficient understanding of the failure mechanisms (Deng et al., 2011). The use of imprecise information (e.g. fuzzy numbers) is an effective method of conducting a risk analysis (Deng et al., 2011). However, communicating uncertainty is an important, yet difficult task, and many different frameworks exist to do so; water quality indices (Sadiq et al., 2007; Van Steenbergen et al., 2012) are one example. Since water resource managers often prefer to use probabilistic measures (rather than possibilistic ones), it is important to convert the possibility of low DO to a comparable probability for effective
communication of risk analysis. Note that the linguistic parameters (e.g. “most likely”) that are often used to convey risk or uncertainty (Van Steenbergen et al., 2012) have a probability-based meaning – in this case “most likely” is a measure of likelihood.

In this chapter, a defuzzification procedure is used to convert the possibility of low DO to a probability measure, to represent the risk of observing low DO (below a given threshold) in the Bow River. This method uses the inverse of the transformation described in Equation 8-1; however, instead of calculating the probability of one element, \( p(x) \), which is of limited value in most applications, it is generalised to calculated \( P(\{ X < x \}) \), as follows (from Khan & Valeo, 2014a, 2015b): for any \( x \) in the support (defined as the \( \alpha \)-cut interval at \( \mu = 0 \)) of a fuzzy number \([a \ b]\) we have the corresponding \( \mu \) and the paired value \( x' \) which shares the same membership level. The value \( \mu \) is the sum of the cumulative probability between \([a \ x]\) and \([x' \ b]\), labelled \( P_L \) and \( P_R \), respectively:

\[
\mu(x) = P_L + P_R \tag{Equation 8-19}
\]

where \( P_L \) represents the cumulative probability between \( a \) and \( x \) which is assumed to equal the probability \( P(\{X < x \}) \), since the fuzzy number defines any values to less than \( a \) to be impossible (i.e. \( \mu = 0 \)). Given the fact that the fuzzy number is not symmetrical, the lengths of the two intervals \([a \ x]\) and \([x' \ b]\) can be used to establish a relationship between \( P_L \) and \( P_R \). Then, \( P_L \) can be estimated as:

\[
P \{ X < x \} = P_L = \frac{\mu}{1 + \frac{(b-x')}{(x-a)}} \tag{Equation 8-20}
\]

Thus, Equation 8-20 gives the probability that the predicted minimum DO for a given day is below the threshold value \( x \). For example, if the lowest acceptable DO concentration for the protection of aquatic life for cold water ecosystems (6.5 mg/L, Canadian Council of Ministers of the Environment, 1999) is selected, then this transformation can be used to calculate the probability that the predicted fuzzy DO will be below 6.5 mg/L.
8.3 Chapter results and discussion

8.3.1 Examples of the probability-possibility transformation

The bin-size optimisation and the probability-possibility transformation algorithms were applied to the collected $Q$ and $T$ data for the nine year period. The constructed fuzzy numbers were then used to calibrate the FNN model. This section compares the results of constructing a discretised probability distribution with and without the bin-size optimisation algorithm, and its impact on the resulting membership function of the fuzzy number. The comparison is illustrated through five examples each for $Q$ and $T$ as a means of illustrating the advantages of using the proposed approach.

Figure 8-3 shows sample results of converting hourly $Q$ observations to fuzzy numbers for five cases. The left most column in the figure shows the raw data, i.e. the observations sampled over the course of 24 hours. The resulting histogram-based probability functions are shown for both the optimised ($D_{\text{opt}}$, illustrated with circles) and original ($D_{\text{orig}}=2r$; see Sect. 2.2.1 for the definition, illustrated with squares) bin-sizes in the second column. The third and fourth columns in Figure 8-3 show the resulting discretised empirical membership function using each of the histograms. The five examples selected here represent a full spectrum of results for the bin-size optimisation. The first row shows an example of when the optimum result was equal to the measurement resolution ($D_{\text{orig}} = D_{\text{opt}} = 2$), followed by cases where the $D_{\text{opt}}$ was 4, 4.5, 10 and 20 times greater than the initial bin-size.

The example in the first row illustrates cases where the bin-size optimisation algorithm calculates an optimum bin-size, corresponding to the minimum cost-function $C_D$, which is equal to the instrument measurement resolution. Thus, the resulting probability distributions for both cases are equivalent, as are the membership functions. In most cases, this occurred when the calculated minimum $C_D$ would result in a $D_{\text{opt}}$ smaller than $D_{\text{orig}} = 2r$, and since this is not physically feasible (measurable) the algorithm did not consider any bin-sizes below $2r$. Of note in this example is that the transformation of the probability distribution results in five empirical membership levels. Only one element was found to have a membership level equal to 1 (at $Q = 161$ m$^3$/s). Thus, the $\alpha$-level cut at this level is a simple singleton: $[161]_{\mu=1}$. The next membership level was calculated as 0.58; again the resulting $\alpha$-cut level only has one element at $Q = 149$ m$^3$/s (which is less
than the modal value). However, at this level the upper and lower limits of \( \alpha \)-cuts at higher membership levels define the upper and lower limits of \( \alpha \)-cuts at lower levels. Thus, using the information from the \( \alpha \)-level cut at \( \mu = 1 \), the level at \( \mu = 0.58 \) was defined as \([149 \ 161]\)\(_{\mu=0.58}\). The next membership level calculated was 0.46, and four elements had membership levels equal to this value, ranging between 147 and 165. The \( \alpha \)-cut interval at this level was defined as: \([147 \ 165]\)\(_{\mu=0.46}\). Note that in this case, this interval captures both the intervals at higher membership levels within its limits, i.e. the lower limit is less than the lower limits of higher intervals, and the upper limit is greater than then upper limits at higher levels. The next membership level was calculated to be 0.125, and three elements between 157 and 171 were assigned this value. However, the lower limit at \( \mu = 0.46 \) (the next higher membership level) was 147, which is less than 157, and thus, for the \( \alpha \)-cut level at this membership, the interval is then revised to: \([147 \ 171]\)\(_{\mu=0.125}\) rather than \([157 \ 171]\)\(_{\mu=0.125}\) to maintain convexity. Again, the reason here is that if something is possible at \( \mu = 0.46 \) it must be possible (by definition) at \( \mu = 0.125 \).

The last membership level found for this particular example was \( \mu = 0 \), with six elements sharing this value, ranging from 145 to 173, resulting in an \( \alpha \)-cut level of \([145 \ 173]\)\(_{\mu=0}\). Together, these five membership levels define a discretised membership function of the fuzzy number for \( Q \) on 27 July 2008. Following this, linear interpolation was conducted to find the elements corresponding to the six predefined membership levels of \( \mu = 0, 0.2, 0.4, 0.6, 0.8 \) and 1. The results are not explicitly shown in the figure for clarity, but are essentially located on the dashed line in the last column on the corresponding membership levels.
The second row in Figure 8-3 shows the results for 20 August 2009, where the optimum bin-size was found to be four times higher than the original bin-size ($D_{\text{opt}} = 0.8$ vs. $D_{\text{orig}} = 0.2$). The impact of this change is clearly evident in the distribution functions in the second column. The original histogram is multi-modal, and with multiple candidates as the modal value (where $\mu = 1$), whereas the post-optimisation histogram is considerably smoother, with a definitive modal value at $Q = 91.4 \text{ m}^3/\text{s}$. The impact of this
increase in bin-size is that the resulting membership function is defined at four membership levels (0, 0.25, 0.54 and 1), whereas the original function was defined at six levels, including an interval (rather than singleton) at $\mu = 1$. This decrease in membership levels in this case has a consequence of smoothing out the membership function, as can be seen by comparing the shapes of the functions in columns three and four. The overall impact of this smoothing out of both the distribution and the membership functions is that the heightened specificity of the original function at $\mu = 0.54$ and above is reduced to a more generalised shape.

Since the objective of the bin-size algorithm was to reduce the error between the histogram created using the $D_{\text{opt}}$ and the unknown theoretical distribution, then the density function plotted in Figure 8-3 represents the closest distribution to the unknown distribution. Hence, the membership function generated using this optimum distribution better reflects the underlying phenomenon than the membership function generated using $D_{\text{orig}}$. Thus, in comparing columns 3 and 4 for the second row, the smoother membership function representing $D_{\text{opt}}$ is preferred. Linear interpolation is then performed on this membership function to get values of $Q$ at the six predefined membership levels.

Similar results can be seen in the third row in Figure 8-3, where the optimised bin-size is 4.5 times greater than the original bin-size, ($D_{\text{opt}} = 9$ vs. $D_{\text{orig}} = 2$). Again, the original histogram is extremely uneven, whereas the post-optimisation histogram is considerably smoother with a definitive modal value at $Q = 277$ m$^3$/s. The overall impact of this smoothing is that the specificity of the function at $\mu = 0.6$ and higher of the original function is reduced to a more general shape in the optimised function.

The fourth row shows a different phenomenon, where instead of smoothing out the original membership function, the combined bin-size optimisation and transformation algorithm, creates a membership function with more specificity. In this case $D_{\text{opt}}$ is ten times higher than $D_{\text{orig}}$, and the consequence of this increase is the smoother probability density function with one clear modal value (at $Q = 70$ m$^3$/s). In contrast, the original histogram had six elements with joint-equal $p(x)$, resulting in a membership function that is shaped similarly to a uniform distribution (column 3) and defined with only 3 membership levels. This means that all values are considered equally-possible and represents maximum vagueness. However, using the optimised value, this is no longer
the case and the modal value is assigned a membership level of 1, and the remaining elements defined at three other membership levels. This suggests that this modal value is more possible (since it has a higher possibility), and this is reflected in the observations. This example illustrates that the method can not only generalise the data to smoother functions (as shown in the first three examples) but can also be more specific when the underlying data demonstrates this but this is not captured by the non-optimised bin-width distribution function.

The last example for \( Q \) in Figure 8-3 is an example of a case where the number of membership levels for both the original and optimised membership function are equal (four in this case), however the bin-size is 20 times greater for the optimised case. In this case, an optimum bin-size was found that did not change the specificity of the membership function, i.e. it is still defined with the same number of intervals but at different membership levels. In this case, the probability for \( D_{\text{orig}} \) is extremely uneven, but smoothed out to a unimodal function with the \( D_{\text{opt}} \). The final membership function for \( D_{\text{opt}} \) is defined more generally (smoothly) especially at higher membership levels compared to the one defined by \( D_{\text{orig}} \). This example again demonstrates the utility of the new coupled optimisation-transformation method to create fuzzy numbers for data where the underlying distribution is unknown.

Figure 8-4 shows similar results for the five water temperature examples, where the \( D_{\text{opt}} \) was equal to the \( D_{\text{orig}} \) (the first example on the top row), or increased by a factor of 1.5, 2.5, 3 or 5. The first example shows a case with very little \( T \) variation over a given day and the water temperature falls between 5.2 and 6.2 °C for the entire day. This lack of variability is responsible for the minimal bin-size selection as \( D_{\text{opt}} \): a unimodal distribution is best constructed using smaller bin-sizes for these cases. The second example shows another case where \( D_{\text{opt}} \) is only slightly greater than the original, resulting in a somewhat smoother probability function, and a slightly smoother membership function.
Figure 8-4 Sample results of probability-possibility transformation for water temperature, $T$

A major difference between the $T$ and $Q$ data is that the former is strongly diurnal, increasing after sunrise in the morning, peaking in late afternoon, and then decreasing
through the night. This temporal trend is seen for all examples in Figure 8-4, but most significantly in the bottom three examples. A major implication for this in developing a probability density function for this data is that the resulting shape will have a tendency to be bimodal. This means that the resulting membership functions might be trapezoidal or near-trapezoidal (and hence most vague) in shape, which is clearly demonstrated in the functions created using $D_{\text{orig}}$ in the bottom three examples. However, in each case the optimised bin-size creates a smoother probability distribution with a clearer modal value, resulting in membership functions that are no longer trapezoidal.

Thus, without using the bin-size optimisation algorithm there is a risk that the resulting membership functions will be too vague and do not represent the information that can be gained from the observations. It is worth noting that for these three examples, if linear interpolation is used on the original membership function, the resulting interpolated fuzzy number will all have equal intervals (due to the trapezoidal shape), transferring no useful information to the final fuzzy number.

Overall, the above examples illustrate the advantages of using the couple method of bin-size optimisation and probability-possibility transformation to create fuzzy numbers for the FNN application. The applicability of this method is not necessarily restricted to this application and can be applied whenever there is a need to construct fuzzy numbers from observed data. The utility of the first component, bin-size optimisation to estimate the density function, is that in cases where either not enough information is available to define a probability distribution, or if the data do not follow the mould of a known density function, or if assumptions on the class of distribution cannot be made, the optimum bin-size can be calculated to define an empirical distribution for the probability-possibility transformation. The advantage of the second component, the algorithm to construct the possibility distribution (i.e. the membership function of the fuzzy number) is that it provides a consistent, transparent and objective method to convert observations (e.g. time series data) into fuzzy numbers, which has been cited as a major hurdle in implementing fuzzy number based applications in the literature (Abrahart et al., 2010; Dubois & Prade, 1993; Civanlar & Trussel, 1986). A noteworthy component of this algorithm is that the fuzzy numbers do not reduce to the simple, triangular shaped
functions that are widely used, but rather the functions better represent the information from the observations.

### 8.3.2 Training the fuzzy neural network

Once the observations of the abiotic input parameters ($Q$ and $T$) were converted to fuzzy numbers, the FNN training algorithm was run using five neurons in the hidden layer, to predict daily minimum DO in the Bow River. First, the values of the fuzzy numbers at $\mu = 1$ was used to train the crisp network. This was done to have initial estimates of the 10 $W_{IH}$ (5 for each input), 5 $B_{IH}$, 5 $W_{HO}$, and 1 $B_{O}$. These initial estimates were used to provide the upper and lower limits of the constraints for the proceeding optimisation algorithm. Once these estimates were calculated, the optimisation algorithm was used to calculate the fuzzy weights and biases using fuzzy inputs, and was started from $\mu = 0$ and moving sequentially to higher membership levels until $\mu = 0.8$. The final level (at $\mu = 1$) was calculated using the midpoint of the intervals estimated at $\mu = 0.8$.

The total optimisation time (using the SCE-UA algorithm) for the selected architecture took 13 hours using a 2.40 GHz Intel® Xeon microprocessor (with 4 GB RAM).

The $E_{MSE}$ and the Nash-Sutcliffe model efficiency coefficient ($E_{NSE}$; Nash & Sutcliffe, 1970) for the training, validation and testing scenarios for $\mu = 1$ are shown in Table 8-3. The $E_{MSE}$ for each dataset are low, between 11% and 16% of the mean annual minimum DO seen in the Bow River for the study period. The $E_{NSE}$ values are approximately equal to 0.5 for each subset, which is higher than $E_{NSE}$ values in the literature for water quality parameters when modelled daily (see Moriasi et al., 2007 for a survey of results) and is considered to be “satisfactory” by their standards. These two model performance metrics highlight that in general, predicting minimum DO using abiotic inputs and a data-driven approach is an effective technique.

<table>
<thead>
<tr>
<th>Subset</th>
<th>$E_{MSE}$ (mg/L)$^2$</th>
<th>$E_{NSE}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>1.52</td>
<td>0.52</td>
</tr>
<tr>
<td>Validation</td>
<td>1.19</td>
<td>0.49</td>
</tr>
<tr>
<td>Test</td>
<td>1.09</td>
<td>0.54</td>
</tr>
</tbody>
</table>
The results of the optimisation component of the algorithm are summarised in Table 8-4, which shows the percentage of data ($P_{CI}$) captured within the resulting $\alpha$-cut intervals for each of the three data subsets. The exact amount of data was captured within each of the training intervals, as was required by the constraints, except for the $\mu = 0.8$ interval. At this interval, the optimisation converged to a value of 29.91% rather than 20%. However, as required by Equation 8-9, the amount of data within the interval has to be greater-than or equal-to the limit defined by $P_{CI}$ (as per Alvisi & Franchini, 2011). This means that a solution to satisfy the constraints with a lower amount of data would either result in non-minimal intervals (though this is unlikely) or that the constraints on the values of the intervals could not be maintained (this issue will be discussed in detail with Figure 8-5 below). For the validation and testing datasets, similar performance is seen for both, with near perfect $P_{CI}$ captured at $\mu = 0.6$, 0.8 and 1, and more than $P_{CI}$ at $\mu = 0.2$ and 0.4. These results are similar to the testing dataset in Alvisi & Franchini (2011).

A sample of the fuzzy weights and biases produced through the optimisation are shown in Figure 8-5. Note that the membership functions are assumed to be piecewise linear (following similar assumptions made in Alvisi & Franchini, 2011; Khan et al. 2013; Khan & Valeo, 2015a), i.e. that the intervals at each membership levels can be joined to create a fuzzy number. This can be confirmed by the fact that each of the weights and biases are convex where intervals at lower levels are wider than intervals at higher levels, and are normal with at least one element with $\mu = 1$. Note that each weight and bias has a non-linear membership function, i.e. none of the functions produced follow the typical triangular functions and are not necessarily symmetric about the modal value.

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>Percent captured, $P_{CI}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
</tr>
<tr>
<td>1.00</td>
<td>-</td>
</tr>
<tr>
<td>0.80</td>
<td>29.91</td>
</tr>
<tr>
<td>0.60</td>
<td>39.93</td>
</tr>
<tr>
<td>0.40</td>
<td>59.95</td>
</tr>
<tr>
<td>0.20</td>
<td>79.98</td>
</tr>
<tr>
<td>0.00</td>
<td>99.39</td>
</tr>
</tbody>
</table>
The figure demonstrates that enough $\alpha$-cut levels (i.e. six levels equally spaced between 0 and 1) have been selected to completely define the shape of the membership functions. A smaller number of levels e.g. two levels, one at $\mu = 0$ and one at $\mu = 1$, the fuzzy number collapses to a triangular fuzzy number, which is not desirable for this research, as discussed in previous sections. When only two levels are selected, the figures demonstrate that significant differences exist between those simple functions and the ones generated using six membership levels: the decrease in the width of the intervals with an increase in membership level is not linear as is in triangular shaped function. Similarly, a higher number of intervals e.g. 100 intervals, equally spaced between 0 and 1, could be selected. The risk in selecting many intervals is that as the membership level increases (closer to 1) the intervals become narrower as a consequence of convexity. This will result in numerous closely spaced intervals, with essentially equal upper and lower bounds, making the extra information redundant. This is demonstrated in the sample membership functions in Figure 8-5 for $W_{IH}$ number 5 and $B_O$, where the intervals at the higher membership levels collapse to a singleton, or are extremely narrow. Thus, defining more uncertainty bands between the existing levels would not add more detail but would merely replicate the information already calculated.

Connecting this back to the results in Table 8-4, these two particular weights and biases show why the percentage of data calculated at $\mu = 0.2$ (for training) cannot be improved by further optimisation. At some point, if the intervals at $\mu = 0.2$ for the various weights and biases collapse to a single element, no further refinement in the model is possible (since all the constraints are met) and the minimum interval width of the predicted DO whilst capturing at least $P_{C1}$ amount of data has been reached. It is worth emphasising here that the uncertainty represented by these fuzzy number weights and biases is not the uncertainty of the particular weight or bias, but is the total forecasting uncertainty defined by the quantifying bands around the crisp predicted value.
Table 8-3, Table 8-4, and Figure 8-5 demonstrates the overall success of the proposed approached to calibrate an FNN model. The optimisation algorithm is defined based on the principles of possibility theory (i.e. defining the amount of data to include in each interval) and is a transparent, repeatable and objective (not arbitrary) method to create the fuzzy numbers for the FNN model.

The observed versus crisp predictions (black dots) and fuzzy predictions at $\mu = 0$ (grey lines) for daily minimum DO for the three different data subsets (training, validation and testing) are shown in Figure 8-6. The figure shows that nearly-all (specifically, 99.4%, 98.8% and 99.0% of the training, validation, and testing subsets, respectively) of the observations fall within the $\mu = 0$ interval, since the observed values (black dots) fall inside the grey lines. This figure also highlights one of the major advantages of the FNN over a simple non-fuzzy ANN: almost all of the fuzzy results intersect the 1:1 line.
whereas many of the crisp results are quite far from that line, especially at low DO values (which is marked at 6.5 mg/L on the figure). In other words, while the fuzzy number prediction may not predict the observed value exactly, they provide at least some possibility of the observed value within its various $\alpha$-cut intervals, but the crisp results do not provide this additional information. This figure illustrates that $E_{\text{NSE}}$ (listed in Table 8-2 for the $\mu = 1$ case only) is not representative of the entire fuzzy number predictions, since it does not capture the performance at different membership levels. Thus, there is a need to develop an equivalent performance metric when comparing crisp observations to fuzzy number predictions.

Figure 8-6 also demonstrates the benefit of the FNN approach compared to the crisp approach with respect to predicting low DO (i.e. when DO is less than 6.5 mg/L). The FNN predicts more of the low DO events within its intervals as compared to the crisp method. The figure demonstrates that both the crisp ($\mu = 1$) and fuzzy predictions tend to over predict the low DO events (since they fall above the 1:1 line), but the fuzzy intervals are closer to the observations (i.e. they intersect the 1:1 line for the majority of low DO events), and therefore predict some possibility (even if it is a low probability) the low DO events occur. Thus, generally speaking the ability of the FNN to capture nearly-all of the data within its predicted intervals guarantees that most of the low DO events are successfully predicted. This is a major improvement over conventional methods used to predict low DO.

Trend plots of observed minimum DO and predicted fuzzy minimum DO for the years 2004, 2006, 2007 and 2010 are illustrated in Figure 8-7 and Figure 8-8. These years were selected due to the high number of low DO occurrences in each year (as listed in Table 8-1 and highlight the utility of the proposed method to predict minimum DO using abiotic factors in the absence of a complete understanding of the physical mechanisms that govern DO in the Bow River. Note that for each year, 50% of the data are training data, 25% are validation and 25% are testing data. However, for clarity this difference is not individually highlighted for each data point in these figures.
Figure 8-6 A comparison of the predicted and observed minimum DO at the $\mu = 0$ interval (black line) and at $\mu = 1$ (black dots)

In Figure 8-7 and Figure 8-8, the predicted minimum DO at equivalent membership levels (e.g. $0^L$ or $0^R$) at different times steps are joined together creating bands representative of the predicted fuzzy numbers calculated at each time step. In doing so, it is apparent that all the observed values fall within the $\mu = 0$ interval for the years 2006, 2007 and 2010, and all but one observation in 2004. The width of each band represents the amount of uncertainty associated with each membership level. For example, the bands are the widest at $\mu = 0$, meaning the results have the most vagueness associated with it.
Narrower bands are seen as the membership level increases until $\mu = 1$. This reflects a decrease in vagueness, increase in credibility, or less uncertainty of the predicted value, as the membership level increases. Note that the majority of the predictions at $\mu = 1$ are single elements but some predictions are $\alpha$-cut intervals (e.g. $[a, b]_{\mu=1}$). This means that when not enough information is available, the fuzzy prediction collapse to trapezoidal membership functions.

In each of the years shown, the majority of the observations tend to fall within the $\mu = 0.2$ interval or higher, with only the low DO (i.e. $< 5 \text{ mg/L}$) falling within the $\mu = 0$ and $\mu = 0.2$ bands. This suggests that the low DO events are predicted with less certainty compared to the occasions when DO concentration is high. Also note that the interval at $\mu = 0$ is highly skewed towards the lower limit ($\mu = 0^L$), i.e. the modal value is not at the centre of the interval. This shows that the FNN has been trained to capture these low DO events, but predicts them with lower credibility. Compared to the crisp results (i.e. those at $\mu = 1$), for these low DO events, the proposed method provides some possibility of low DO, whereas the crisp results do not predict a possibility of low DO. Thus, the ability to capture the full array of DO observations within different intervals is an advantaged of the proposed method over existing methods.

The trend plot for 2004 shows that observed DO decreases rapidly from late June to late July, followed by a few days of missing data and near-zero observations, before increasing to higher concentrations. Details of this trend are shown in Figure 8-9 which shows magnified versions of important periods for each year. The reason for this rapid decrease in 2004 is unclear and may be related to problems with the real-time monitoring device which was in its first year of operation that year. However, it demonstrates that the efficacy of data-driven methods is dependent on the quality of the data. Since the proposed method was calibrated to capture nearly-all the observations (including outliers like those seen in 2004) within the least certain band at $\mu = 0$, the resulting network predicts results to include these outliers, but at low credibility levels. As the data length increases (i.e. the addition of more data and the FNN is subsequently updated), the number of these types of outliers included within the $\mu = 0$ band will decrease because the optimisation algorithm (Equation 8-12 to Equation 8-14) searches for the smallest width of the interval whilst including 99.5% of the data. Thus, with more data, it is expected
that these extreme events (i.e. the outliers seen in 2004) will no longer be captured within the $\mu = 0$ band.

The time series plot for 2006 shows that all the observations fall within the predicted intervals, and that the predicted trend generally follows the observed trend. The majority of the 25 low DO events ($< 5$ mg/L) occur from mid-July and continue occasionally until mid-September. Details of some of these low DO events are plotted in Figure 8-9. Figure 8-7 demonstrates these low DO events are captured between $\mu = 0$ and 0.2 intervals, similar to the 2004 case, meaning that the credibility of these predictions is the lowest. However, unlike the 2004 case, Figure 8-9 demonstrates that in 2006 the predicted intervals tend to follow the same trend as the observations for these low DO events, even if it is predicting them at a low credibility.

In contrast to the results from 2004 and 2006, the majority of observations are captured at higher membership levels (i.e. greater than $\mu = 0.2$) in 2007 as shown in Figure 8-8. That is, only a limited number of observations are captured within the lowest credibility band. More importantly, 26 out of the 27 low DO (<6.5 mg/L) events are captured at a membership level greater than 0.2. Meaning that the low DO predictions in 2007 for the 6.5 mg/L guideline are predicted with higher credibility than the 2004 and 2006 cases.

Another difference for the results from this year is that many of the low DO observations for 2007 are more evenly scattered around the $\mu = 1$ predictions (as seen in Figure 8-9) in contrast to the 2004 and 2006 cases, where the low DO events were always predicted to be closer to lower bounds of the intervals.
Figure 8-7 A comparison of the observed and predicted minimum DO trends for:

(top) 2004, and (bottom) 2006
Figure 8-8 A comparison of the observed and predicted minimum DO trends for three sample years: (top) 2007 and (bottom) 2010.
The trend plot for 2010 is shown in Figure 8-8 and it is clear that all observations fall within the $\mu = 0.2$ interval or at higher $\alpha$-level intervals, meaning that the predictions capture the observations with higher certainty. This is likely due to the lack of DO events below 5 or 6.5 mg/L in 2010. Also of note for this year is that all observations are less than 10 mg/L, and about 90% of all observations are below the 9.5 mg/L guideline (as listed in Table 8-1). The trend plot again illustrates that the FNN generally reproduces the overall trend of observed minimum DO. This can be seen in a period in early May where DO falls from a high of 10 mg/L to a low of 7 mg/L, and all the predicted intervals replicate the trend. This is an indication that the two abiotic input parameters are suitable parameters for predicting minimum DO in this urbanised watershed.

Figure 8-9 shows details of a low DO (<9.5 mg/L) event in 2010 in late July through late August. The bulk of low DO events are captured between the $\mu = 0.6^R$ and $0.2^R$ intervals – demonstrating that these values are predicted with higher credibility than the other low DO cases in 2004 and 2006, and are predicted closer to the upper end of the interval. All of the low DO (<9.5 mg/L) observations in this plot are underpredicted by the crisp method (though not with the FNN method since they are captured within a fuzzy interval). This shows that the crisp ANN results tend to overpredict extremely low DO events (i.e. < 5 mg/L) while under-predicting the DO < 9.5 mg/L events.

The analysis of the trend plots for these four sample years show that the proposed FNN method is extremely versatile in capturing the observed daily minimum DO in the Bow River using $Q$ and $T$ as inputs. The crisp case (at $\mu = 1$) cannot capture the low DO events (as shown in Figure 8-7 and Figure 8-8), however the FNN is able to capture these low DO events. Generally speaking the training method selected for the FNN has been successful in creating nested-intervals to represent the predicted fuzzy numbers. The widths of the predicted intervals correspond to the certainty of the predictions (i.e. larger intervals for more uncertainty). The utility of this method is further demonstrated in the proceeding section, where the risk of low DO is estimated using a possibility-probability (i.e. defuzzification) technique.
Figure 8-9 Zoomed in views of the trend plots for four sample year corresponding to important periods with low DO occurrences.
8.3.3 Risk analysis for low DO events

The utility of the FNN method is illustrated through an analysis of the ability of the model to predict low DO events, and then a possibility-probability transformation is used to assess the risk of these low DO events. The number of occasions when observed DO was below any of the three guidelines used for this research are summarised in Table 8-1. The FNN model was able to capture 100% of all low DO events (i.e. below 5, 6.5 or 9.5 mg/L) within the predicted intervals. In comparison, the crisp network (i.e. at $\mu = 1$) did not predict DO to be less than 5 mg/L on any of the 51 occasions. Similarly, it predicted DO to be less than the more conservative limit of 6.5 mg/L in only 53% of the 184 occurrences. For the last case, the 9.5 mg/L limit, the ANN method still trailed the FNN method, by predicting 96% of these low DO events. This illustrates that not only can the FNN method capture more low DO events within its predicted intervals, it performs exceptionally better for the highest risk case (DO < 5 mg/L). In general, more days were correctly identified when there was a risk of low DO using FNN rather than the typical ANN approach. This is one of the major advantages of using a fuzzy number based uncertainty analysis component to low DO prediction.

Once all the low DO events were identified, the inverse transformation (defuzzification) described in Sect. 2.4 was used to estimate the probability of low DO. The primary reason for converting from possibility to probability is to improve the communication of the risk of low DO. For each low DO event (i.e. either at the 5, 6 or 9.5 mg/L), the predicted membership function was used to determine the possibility of low DO, i.e. identify the membership level where the membership function intersects either of the low DO guidelines (some examples of low DO events are shown in Figure 8-10). Once these were identified, the defuzzification technique was used to predict the probability of low DO (e.g. $P(\{\text{DO}<5 \text{ mg/L}\})$).

For the first case, $P(\{\text{DO}<5 \text{ mg/L}\})$, the probability ranged between 11.5% and 16.6% for the 51 events, with a median value of 14%. This means that on days when DO was observed to be below 5 mg/L, the FNN results identified the possibility of low DO and the probability of DO to be below the 5 mg/L guideline was ~14%. Thus, the FNN method predicts a probability of low DO (even if it is relatively small) on days when the crisp ANN does not predict a low DO event. This value can be used as a threshold by
water resource managers for estimating the risk of low DO. For example, if forecasted water temperature and flow rate are used to predict minimum fuzzy DO using the calibrated model, if the risk of low DO reaches 14%, the event can be flagged. Appropriate defence mechanisms can then be implemented to prevent the occurrence of low DO.

For the 184 cases where DO was observed to be less than 6.5 mg/L, the probability-possibility transformation estimated the risk of low DO to be between 13.7% and 92.9%, with a median value of 73.4%. Compared to the first case, the probability of low DO for this threshold is higher and more variable. The low probabilities are associated with predictions of low DO at lower credibility levels at the lower limit of the intervals (i.e. L), whereas the higher probabilities are associated with predictions corresponding to the upper limits of the intervals (R). For 43 out of the 184 low DO events, the probability of low DO was less than 21% – these events correspond to predictions of low DO at low credibility levels at the lower limits. For the majority of events (107 out of 184), the risk was high, more than 65%. It is worth noting that the crisp network only predicted 53% of these low DO events, and of those correctly identified, the majority were over-predicted.

For the last, most conservative case, the probability of predicting DO to be less than 9.5 mg/L (1179 events) varied between 21.9% and 100%, with a median value of 98.1%. Only 46 out of the 1179 events had a probability of less than 70%; the majority of events had a high risk of low DO: more than 80% of the events had a risk of low DO of more than 90%. This shows that the FNN can predict with high probability, the events were minimum daily DO is observed to be below the 9.5 mg/L limit.

The predicted membership functions of minimum DO for nine examples are shown Figure 8-10, along with the observed minimum DO (the vertical dashed line). Three samples are shown for each low DO guideline: 5, 6.5, or 9.5 mg/L; along with the associated risk of low DO calculated using the defuzzification technique. Note that the membership functions of the predicted fuzzy numbers show that each is uniquely shaped, convex and normal, highlighting the fact that the proposed optimisation algorithm successfully produces nested intervals at each membership levels (as it does for the weights and biases shown in Figure 8-5). For the predicted fuzzy DO, the intervals are largest at $\mu = 0$, which decrease in size as the membership level increases. The shape of
the membership functions are not triangular shaped as assumed in many fuzzy number
based applications. This is of significance because it shows that the amount of uncertainty
(or credibility) of the model output does not change linearly with the magnitude of DO,
which has important implications regarding the risk of low DO.

For the 5 mg/L guideline, the intersection of the membership function and the guideline
occurs at low possibility levels (between $\mu = 0^L$ and 0.2$^L$), meaning that the
corresponding probability will be low as well, as illustrated by the probability values
shown in the figure. This again highlights that the risk of low DO ($< 5$ mg/L) is predicted
to be low by the FNN mostly due to the fact that the observations are captured at low
membership levels. Note that the crisp ANN results (at $\mu = 1$) always over predict low
DO, as shown in these three examples. The observed value falls within the predicted
interval for each case, also at low membership levels.

The examples for the 6.5 mg/L guideline (second row in Figure 8-10) show that the
intersection between the membership function and the guidelines occurs between $\mu = 0.4$
and 0.6 on 26 July 2006, between $\mu = 0.6$ and 0.8 on 8 August 2007, and at about $\mu = 0.6$
on 29 September 2004. This illustrates the broader trend with the 6.5 mg/L guideline
(which was discussed earlier and had a large range of risk predictions), which is that for
the full dataset, the possibility of low DO ($< 6.5$ mg/L) occurs at every interval, with the
majority occurring at higher intervals. This is in contrast to the 5 mg/L guideline where
the possibility of low DO only occurs only between $\mu = 0$ and 0.2.

The last row in Figure 8-10 show sample low DO results for the 9.5 mg/L guideline. As
discussed above, more than 80% of these events had a high (more than 90%) risk of low
DO. In the first example, on 23 September 2004, the guideline intersects the membership
function at $\mu = \sim0.2^R$, corresponding to a $\sim97\%$ risk of low DO. The 6 August 2008 has a
low DO prediction of 100% – this is because the predicted fuzzy number is entirely
below the guideline limit. A similar result can be seen for the last example. These
examples also illustrate that had only a triangular membership function been used (i.e. the
fuzzy numbers defined at two membership levels), the probability of low DO could not
be quantified as specifically as it has been here. The slight changes in membership
function shapes between intervals impact the final probability, and a linear function
would have not captured these changes.
Figure 8.10 Sample plots of low DO events and the corresponding risk of low DO calculated using a possibility-probability transformation for the (top) 5 mg/L, (middle) 6.5 mg/L, and (bottom) 9.5 mg/L guideline.

These examples are meant to illustrate the potential utility of the data-driven and abiotic input parameter DO model, that can be used to assess the risk of low DO. Given that it is a data-driven approach, the model can be continually updated as more data is available, further refining the predictions. Various combinations of input values can be
used to predict fuzzy minimum DO and defuzzification technique can be used to quantify the risk of low DO given the input values. The utility of this method is that a water-resource manager can use forecasted water temperature data and expected flow rates to quantify the risk of low DO events in the Bow River, and can plan accordingly. For example, if the risk of low DO reaches a specific numerical threshold or trigger, different actions or strategies (e.g. increasing flow rate in the river by controlled release from the upstream dams) can be implemented. The quantification of the risk to specific probabilities means that the severity of the response can be tuned to the severity of the calculated risk.

8.4 Chapter summary
A new method to predict DO concentration in an urbanised watershed is proposed. Given the lack of understanding of the physical system that governs DO concentration in the Bow River (in Calgary, Canada), a data-driven approach using fuzzy numbers is proposed to account for the uncertainty. Further, the model uses abiotic (non-living, physical and chemical attributes) factors as inputs to the model. Specifically water temperature and flow rate were selected which are routinely monitored and thus, a large dataset is available.

The data-driven approach proposed is a modification of an existing fuzzy neural network method that quantifies the total uncertainty in the model by using fuzzy number weights and biases. The proposed model refines the exiting model by (i) using possibility theory based intervals to calibrate the neural network (rather than arbitrarily selecting confidence intervals), and (ii) using fuzzy number inputs rather than crisp inputs. This research also proposes a new two-step method to construct these fuzzy number inputs using observations. First a bin-size optimising algorithm is used to find the optimum histogram (as an estimate of the underlying but unknown probably density function of the observations). Then a probability-possibility transformation is used to determine the shape of the fuzzy number membership function.

The results demonstrate the network training algorithm proposed can be successfully implemented. Model results demonstrate that low DO events are better captured by the fuzzy network as compared to a non-fuzzy network. A defuzzification technique is then used to calculate the risk of low DO events. Generally speaking, the method demonstrates
that a data-driven approach using abiotic inputs is a feasible method for predicting minimum daily DO. Results from this research can be implemented by water resource managers to assess conditions that lead to, and quantify the risk of low DO.
9. Conclusions

9.1 General conclusions
The overall objective of this dissertation was to research and develop methods to improve environmental prediction (specifically, dissolved oxygen and flow rate) using a combination of data-driven methods (linear regression and neural networks) and fuzzy numbers. Three specific objectives were researched: (i) develop methods to construct fuzzy numbers from high-frequency observations; (ii) develop data-driven models that use fuzzy numbers to quantify the uncertainty in the system; and (iii) develop a method to use fuzzy number model outputs for risk assessment. The major conclusions of each objective are provided below.

9.1.1 Constructing fuzzy numbers
In most applications using fuzzy set theory and fuzzy numbers, linear (or triangular shaped) membership functions are used to define fuzzy sets. In this chapter the utility of non-linear membership functions (e.g. generic Normal or Gumbel-based functions) was explored in relation to DO prediction. This research showed that it is more appropriate to use these non-linear functions. These functions were better able to represent the uncertainty in the observations and also performed better in predicting DO concentration using data-driven models. Triangular shaped functions are useful when minimal information of the variable is present – however, when additional data (e.g. real-time observations) are available, it should be incorporated into the fuzzy number framework.

Following this, an algorithm was developed to apply a probability-possibility transformation that converts the observed data (e.g. flow rate, water temperature or DO) into fuzzy numbers. The new method was able to generate distinct, non-linear membership functions for each variable with no assumption of the underlying distribution of the data. The performance of models using the non-linear membership functions for both the observed data and the model coefficients (either in the regression or neural network applications) demonstrated that the linear (and symmetric) representation of uncertainty may not be realistic for these types of applications.

The influence of changing the input resolution and the error value used to construct the fuzzy number was analysed to measure the robustness of the method. The analysis shows that the model performance did not decrease as the amount of data was reduced from 96
to 6 samples per day. An implication of this is that rather than collecting extremely high resolution data from one site, it may be preferable to collect data from a number of different sites at lower frequency. This would allow multiple models to be calibrated, improving site specific predictions. The analysis of the error value used for constructing fuzzy numbers showed that model performance improved as the error value is increased from representing only aleatory uncertainty to higher values representing epistemic uncertainty.

Lastly, the transformation algorithm to construct fuzzy numbers was generalised by adopting a two-step procedure. A bin-size optimisation algorithm was adapted to predict the optimum bin-size to create the histograms (as an estimator of the unknown density function of the observations). In doing so, the somewhat arbitrary selection of bin-size in the first approach was improved. Then, a generalised transformation (where multiple elements can have a membership level equal to 1) was introduced. This allows the construction of fuzzy numbers from observations without assumptions of the underlying distribution of the data and is flexible enough to handle different types of distributed data.

9.1.2 Data-driven modelling using fuzzy numbers

In this dissertation a number of different approaches to model DO concentration in the Bow River in Calgary, Alberta, Canada were explored. This included the use of abiotic factors (namely flow rate and water temperature) as the predictors. An autoregressive approach was also constructed, where lagged DO (e.g. from the previous day) was used as the predictor.

First, a new fuzzy linear regression method was proposed that used fuzzy number inputs, outputs and model coefficients. The use of the extension principle meant that the fuzzy numbers were reduced to a series of nested $\alpha$-cut intervals, allowing the regression to be generalised to a simple optimisation algorithm. This new method for fuzzy regression was then compared against two existing fuzzy regression methods, Bayesian linear regression, and error-in-variables regression. The new method was better able to predict DO concentration (or flow rate for the error-in-variables method) due to its ability to incorporate different sources of uncertainty (both random and non-random) in each component (input, output and coefficients). Output from these models was better able to predict low DO or high flow rate events. Also, results shows that a relatively small
dataset is needed for high performance, illustrating that the data-driven approach can represent the physical system with even a relatively small dataset. A number of model assessment metrics (including $d^2$ and a ranking method) were proposed to analyse the performance of a fuzzy predictions with respect to crisp observations.

The model was extended to predict DO across the cross-section of the Bow River showing that the intra-cross-section variability was lower than the one-point variability. This suggests that focusing on different locations along the Bow River is more important than focusing on the impact of bathymetry at one location. The use of a recursive algorithm to mimic a real-time updating model was also analysed. The results show that as new data is collected these models can be continually updated to reflect changes in the urban watershed and improve predictions.

In general, fuzzy linear regression methods tend to outperform probability-based methods for the prediction of DO or flow rate in a highly complex urban riverine environment. The proposed approach is useful for water resource managers in many jurisdictions, given the simplicity of the model and the ease of collecting real-time data. This can be easily extended to other water quality parameters.

A possibility theory based refinement to an existing fuzzy neural network model was developed and used to model minimum DO concentration in the Bow River. The refinements included the use of fuzzy number inputs in the model calibration procedure, and the use of possibility theory to define the amount of data captured by the predicted intervals. Further, a method to find the optimum network architecture was proposed to select the number of hidden neurons and the amount of data used for training, validation and testing. The performance of the updated fuzzy neural network was compared against the crisp results, and the method demonstrated the applicability of an abiotic input and data-driven based approach to predict DO concentration in highly complex and uncertain environments.

9.1.3 Risk analysis

The results of both modelling procedures, the fuzzy linear regression and fuzzy neural network methods, demonstrated that an improved ability to predict low DO concentration using abiotic factors compared to non-fuzzy techniques. This was validated by the use of various performance metrics. Similar results were seen for the peak flow rate model. The
models highlight the advantages of using a fuzzy numbers to represent uncertainty, as it was able to predict the possibility and probability of low DO or high peak flow rate, whereas a conventional data-driven modelling approach did not capture this information. The fuzzy methods captured more of these extreme events within its predicted interval compared to the other methods (e.g. Bayesian), and did so at a higher performance. Further, the use of non-linear membership functions demonstrated that more of the extreme events can be accurately predicted compared to linear functions. In general, the data-driven methods correctly identified the risk of the extreme events where other methods could not.

The fuzzy predictions were then use to quantify the risk (i.e. probability of low DO or high peak flow rate) using a possibility-probability transformation. A new possibility-probability transformation was used where the cumulative probability of an event was directly calculated using the given membership function rather than calculating individual probability measures of each element. The quantification of risk allows water resource managers to implement strategies to prevent the occurrence of low DO, or high flows, commensurate with the magnitude of risk.

The neural network model was used to calculate all the combination of input values that lead to some possibility, and hence some probability of low DO. This risk tool can be directly used by water resource managers to assess the risk of low DO for the selected conditions. This tool can be adapted to different locations along the Bow River, or updated as more data is available. The risk tool also demonstrates the utility of the non-linear membership functions, where the risk of low DO is shown to change non-linearly with the membership level.

**9.2 Novel contributions**

This dissertation has explored the use of data-driven methods and fuzzy numbers for environmental prediction. The novel contributions include:

1. a new method of fuzzy linear regression method where the inputs, output and model coefficients are all fuzzy numbers
2. an updated method for fuzzy neural networks where the inputs, output and model coefficients are all fuzzy numbers and a possibility theory based optimisation algorithm is used to train the network
3. A new, generalised method to construct fuzzy numbers from observed data is developed that uses bin-size optimisation and probability-possibility transformation.

4. An inverse-transformation to predict the cumulative probability was developed using the principles of the probability-possibility transformation; the method was used to create a low DO risk tool.

### 9.3 Future research directions

Based on the research results presented in this dissertation, future research should focus on:

1. **The use of alternative theories for uncertainty analysis**
   
   Use of fuzzy numbers and possibility theory is one of many uncertainty theories that are an alternative to the traditional probability based approach. This dissertation has shown the utility and advantages of exploring and applying one alternative for environmental prediction and risk analysis. There is significant potential to redevelop the approach used in this dissertation under different uncertainty frameworks, e.g., Dempster–Shafer theory of belief and plausibility, imprecise probability, and a coupled possibility-necessity framework. Specifically, the concept used to create fuzzy numbers should be extended to create alternative representations of uncertainty in these other frameworks, e.g. defining belief and plausibility functions. Many of these other frameworks have traditionally relied on “expert” input or linguistic variables. However, there is potential to include information from observations to influence the uncertainty quantification.

2. **Develop new model evaluation criteria for fuzzy numbers**
   
   Some attempts at establishing modelling evaluation criteria have been made in this dissertation. However, there is still a need to develop fuzzified versions of existing performance metrics; for example, the widely used Nash-Sutcliffe model efficiency and root mean square error. Developing these criteria will assist in comparing the performance of fuzzy models with crisp models and hence, assist in demonstrating their applicability in various applications. The overall objective of these criteria should be to conceptually mimic the crisp counterparts yet follow fuzzy arithmetic guidelines.
3. **Development of econometric and non-linear techniques for environmental modelling**

Results presented in this dissertation show the applicability of the coupled method of using two types of data-driven methods with fuzzy numbers to improve environmental prediction and uncertainty analysis. The developments presented can be seen as a general first-step in incorporating possibility theory into two simple and widely used data-driven methods. There is significant potential to apply the approach developed here to a wide array of data-driven methods.

Specifically, both developed methods (fuzzy regression and fuzzy neural networks) use one method to incorporate fuzziness into the models. For example, in fuzzy neural networks the fuzziness is included by forcing a predetermined amount of data within a predicted fuzzy interval. However, many conceptual alternatives exist – for example, a fuzzy neural network that minimises the distance between inputs and outputs (similar to the regression case). The impact of using these different definitions of fuzzy numbers in numerical modelling needs to be explored. The use of different frameworks may be determined by the overall objective (predicting a range or an exact mapping).

A promising extension of this approach is in econometric methods (time series based applications). This is because most econometric methods are designed for systems where the underlying mechanisms are highly complex are difficult to describe mathematically using conceptual methods and often rely on the high-resolution data which are now readily available for many environmental application. These methods typically rely on probability based methods for uncertainty analysis but the use of other methods has not been explored, especially for real-time, updating methods for environmental prediction. Thus, there is a need to develop conceptually equivalent methods of uncertainty quantification and propagation in econometric modelling techniques using fuzzy numbers.

Similarly there are a number of non-linear methods data-driven methods (e.g. genetic programming, support vector machines, chaos theory methods, evolutionary polynomial regression, instance-based learning, and model trees) that are popular for environmental applications. Many users prefer the use of non-linear methods due to the perceived advantage in modelling non-linear systems. The use of fuzzy numbers to propagate uncertainty in these systems has not been explored – specifically the impact of fuzzy
arithmetic (max min operators) on uncertainty quantification. Methods to include fuzzy numbers in these methods would be of great benefit for uncertainty and risk analyses.

4. **Selection and classification of fuzzy input parameters**

In crisp applications, selection of input parameters requires a sensitivity or correlation analysis to determine the suitability of input parameters. Given that fuzzy numbers represent all possible values, there is a need to include expanded variable (i.e. fuzzy intervals) in the pre-selection analysis. This will require the construction of new methods to determine equivalent correlation parameters for fuzzy numbers. It is not necessarily true that the fuzzy inputs follow the same relationships as the crisp inputs, due to the expanded possible values. In the same vein, many parametric, classification techniques are used to determine the best input parameters (e.g. self-organising maps) which require adaption to handle fuzzy numbers.

5. **Incorporating fuzzy numbers into physically-based models**

For cases where the physical system is well defined using a numerical model, or a pre-determined physical model is used for simplicity, there is still a need to incorporate the uncertainty in the input parameters (i.e. observations). Traditionally, probability based methods are used for these applications. The method to construct fuzzy numbers from observations presented in this research can thus be used to calibrate physically-based models. This would mean including fuzzy arithmetic in these physically-based models to account for the fuzzy number inputs.

6. **Re-evaluate existing DO guidelines to account for fuzziness in the system**

Given the difficulty in comparing fuzzy predictions with crisp guidelines, there is a need to develop low DO guidelines that are fuzzy numbers. The development of these guidelines can use the proposed fuzzy risk tool to generate the membership function of the guidelines. Given the fact that there are different guidelines for different conditions (i.e. for low flow season, or high temperature), there is significant potential to reframe the existing imprecise guidelines as a fuzzy number.
7. Expand DO monitoring locations and recalibrate the developed models

The use of real-time data is essential for the methods outlined in this research. Given the inexpensive deployment of continuous monitoring stations, it is recommended that a number of additional DO monitoring locations be identified along the reach (upstream and downstream). This will more beneficial than trying to characterise the bathymetry at one location. The new data should then be used to recalibrate the proposed models at different sites to analyse the risk of low DO along the reach.
Bibliography


BRBC (2010b). *Profile of the Bow River Basin.*


Appendix A
MATLAB code for Chapter 2

1. Constructing triangular shaped fuzzy numbers
Sample code to construct triangular fuzzy numbers for DO data for 2006

```matlab
% DO error
DOe1 = zeros(185,96);
DOe2 = zeros(185,96);
for i=1:185
    for j = 1:96
        if DO2006(i,j)<10
            DOe1(i,j)=DO2006(i,j) - 0.2;
            DOe2(i,j)=DO2006(i,j) + 0.2;
        end
        if DO2006(i,j)>10
            DOe1(i,j)=0.98*DO2006(i,j);
            DOe2(i,j)=1.02*DO2006(i,j);
        end
    end
end

%% Calculate median (formerly nanmean % DO data
DOm2006 = nanmedian(DO2006e,2);
%% Calculate min and max
DOMin2006 = min(DO2006e,[],2);
DOMax2006 = max(DO2006e,[],2);
%% create linear fuzzy matrices
%L = left R = right
DOf2006L = [DOMin2006 0.25*(DOMin2006-DOm2006) + DOMin2006
            0.50*(DOMin2006-DOm2006) + DOMin2006
            0.75*(DOMin2006-DOm2006) + DOMin2006
            0.75*(DOMax2006-DOm2006) + DOMax2006
            0.50*(DOMax2006-DOm2006) + DOMax2006
            0.25*(DOMax2006-DOm2006) + DOMax2006
            DOMax2006];
DOf2006 = [DOf2006L DOm2006 DOf2006R];
%fuzzy membership function matrix with assigned alpha cuts
U = [0 0.25 0.5 0.75 1 0.75 0.5 0.25 0];
U = repmat(U, 185,1);
```

2. Constructing Gaussian shaped fuzzy numbers
Sample code to construct Gaussian shaped fuzzy numbers for DO data for 2006

```matlab
% Calculate error
% DO error
DOe1 = zeros(185,96);
DOe2 = zeros(185,96);
for i=1:185
    for j = 1:96
        if DO2006(i,j)<10
            DOe1(i,j)=DO2006(i,j) - 0.2;
            DOe2(i,j)=DO2006(i,j) + 0.2;
        end
        if DO2006(i,j)>10
            DOe1(i,j)=0.98*DO2006(i,j);
            DOe2(i,j)=1.02*DO2006(i,j);
        end
    end
end
```
DOe1 (DOe1 == 0) = NaN;
DOe2 (DOe2 == 0) = NaN;
DO2006e = [DOe1 DO2006 DOe2];
% Calculate mean and var
DOM2006 = nanmean(DO2006e, 2);
DOS2006 = nanstd(DO2006e, [], 2);
% Calculate min and max
DOMin2006 = min(DO2006e, [], 2);
DOMax2006 = max(DO2006e, [], 2);
%% normpdf; calculate f for each years data
f2006 = zeros(185, 288);
for i = 1: 185
    f2006(i,:) = normpdf(DO2006e(i,:), DOM2006(i,1), DOS2006(i,1));
end
clearvars i
% calculate f(x=mean) for each day for each year
fmax2006 = 1./(sqrt(2*pi)*DOS2006);
% calc fs at alpha = 0.25, 0.5, 0.75 ("a" "b" "c")
fa2006 = 0.25*fmax2006;
fb2006 = 0.5*fmax2006;
fcc2006 = 0.75*fmax2006;
% calc corresponding x's ie support for each alpha cut (0 0.25 0.5 0.75 1)2006 DO
temp = sqrt(-2.*((DOS2006.^2).*log(fa2006.*DOS2006.*sqrt(2*pi))));
xa = [DOM2006-temp DOM2006+temp];
temp = sqrt(-2.*((DOS2006.^2).*log(fb2006.*DOS2006.*sqrt(2*pi))));
xb = [DOM2006-temp DOM2006+temp];
temp = sqrt(-2.*((DOS2006.^2).*log(fcc2006.*DOS2006.*sqrt(2*pi))));
xc = [DOM2006-temp DOM2006+temp];
clearvars temp
DOF2006 = [DOM2006-3*DOS2006 xa(:,1) xb(:,1) xc(:,1) DOM2006 xc(:,2)
xb(:,2) xa(:,2) DOM2006+3*DOS2006];
clearvars xa xb xc

3. Constructing Extreme Value III shaped fuzzy numbers
Sample code to construct Extreme Value III fuzzy numbers for Q data for 2006

%% Calculate error
Qe1 = zeros(185, 24);
Qe2 = zeros(185, 24);
for i=1:185
    for j = 1:24
        if Q2006(i,j)<80
            Qe1(i,j)=0.98*Q2006(i,j);
            Qe2(i,j)=1.02*Q2006(i,j);
        end
        if Q2006(i,j)>220
            Qe1(i,j)=0.93*Q2006(i,j);
            Qe2(i,j)=1.07*Q2006(i,j);
        end
        if Q2006(i,j)>80 && Q2006(i,j)<220
            Qe1(i,j)=0.97*Q2006(i,j);
            Qe2(i,j)=1.03*Q2006(i,j);
        end
    end
end
Qe1 (Qe1 == 0) = NaN;
Qe2 (Qe2 == 0) = NaN;
Q2006e = [Qe1 Q2006 Qe2];

% Calculate mean and var, location and scale factors
Qm2006 = nanmean(Q2006e,2);
Qs2006 = nanstd(Q2006e,[]); betas2006 = Qs2006*sqrt(6)/pi;
mu2006 = Qm2006 - (psi(1)*beta2006);
clearvars Qe1 Qe2 Q2006 Q2007 i j

%% evpdf; calculate f for each years data
z2006 = zeros(185,72);
f2006 = zeros(185,72);
for i = 1: 185
    z2006(i,:) = (Q2006e(i,:)-mu2006(i,1))./beta2006(i,1);
f2006(i,:) = (1/beta2006(i,:))*exp(-z2006(i,:)).*exp(-exp(-
z2006(i,:)));
end
clearvars i

%% calculate f(x=mean) for each day for each year
z = (Qm2006-mu2006).beta2006;
fmax2006 = (1/beta2006).*exp(-z).*exp(-exp(-z));
clearvars z

% calc fs at alpha = 0.25, 0.5, 0.75 ("a" "b" "c")
a2006 = 0.25*fmax2006;
b2006 = 0.5*fmax2006;
c2006 = 0.75*fmax2006;
% calc corresponding x's ie support for each alpha cut (0 0.25 0.5 0.75 1) 2006 flow
Q01 = mu2006 - 2*beta2006;
Q02 = mu2006 + 5*beta2006;

z1 = zeros(185,1);
zr = zeros(185,1);
for i=1:185
    z1(i,1) = fsolve(@(z)z-exp(-z)-log(a2006(i,1)*beta2006(i,1)),-3);
zr(i,1) = fsolve(@(z)z-exp(-z)-log(a2006(i,1)*beta2006(i,1))+3.5);
xla = z1.*beta2006+mu2006;
xra = zr.*beta2006+mu2006;
end

for i=1:185
    z1(i,1) = fsolve(@(z)z-exp(-z)-log(b2006(i,1)*beta2006(i,1)),-2);
zr(i,1) = fsolve(@(z)z-exp(-z)-log(b2006(i,1)*beta2006(i,1))+2.5);
xlb = z1.*beta2006+mu2006;
xrb = zr.*beta2006+mu2006;
end

for i=1:185
    z1(i,1) = fsolve(@(z)z-exp(-z)-log(c2006(i,1)*beta2006(i,1)),-1);
zr(i,1) = fsolve(@(z)z-exp(-z)-log(c2006(i,1)*beta2006(i,1))+1.5);
end
xlc = zl.*beta2006+mu2006;
xrc = zr.*beta2006+mu2006;
end

Qf2006 = [Q01 xla xlb xlc mu2006 xrc xrb xra Q02];
clearvars i xla xlb xlc xra xrb xrc zl zr

4. Multiple Linear Regression

% for membership = 1, 2006 & 2007 data to construct
DO = [DOf2006(:,5); DOf2007(:,5)];
T = [Tf2006(:,5); Tf2007(:,5)];
Q = [log10(Qf2006(:,5)); log10(Qf2007(:,5))];
I = ones(length(T),1);
X = [I T Q];

[b bint] = regress(DO, X);
% calculate r^2
DOp = b'*X';

%% for membership = 0.75 LEFT, 2006 & 2007 data to construct (same for
other levels)
DOaL = [DOf2006(:,4); DOf2007(:,4)];
TaL = [Tf2006(:,4); Tf2007(:,4)];
QaL = [log10(Qf2006(:,4)); log10(Qf2007(:,4))];
IaL = ones(length(T),1);
XaL = [IaL TaL QaL];
[baL bintaL] = regress(DOaL, XaL);
DOapL = baL'*XaL';

% for membership = 0.75 RIGHT, 2006 & 2007 data to construct (same for
other levels)
DOaR = [DOf2006(:,6); DOf2007(:,6)];
TaR = [Tf2006(:,6); Tf2007(:,6)];
QaR = [log10(Qf2006(:,6)); log10(Qf2007(:,6))];
IaR = ones(length(T),1);
XaR = [IaR TaR QaR];

[baR bintaR] = regress(DOaR, XaR);
DOapR = baR'*XaR';

% fuzzy arithmetic fix
for i = 1:370
    if DOapL(1,i)>DOapR(1,i)
        temp1 = DOapL(1,i);
        temp2 = DOapR(1,i);
        DOapL(1,i) = temp2;
        DOapR(1,i) = temp1;
    end
end
for i = 1:370
    if DOp(1,i)>DOapR(1,i)
        temp3 = DOapR(1,i);
        temp4 = DOp(1,i);
        DOapR(1,i) = temp4;
        DOp(1,i) = temp3;
    end
end
Appendix B
MATLAB code for Chapter 3

1. Constructing Gaussian shaped fuzzy numbers
Function to construct Gaussian shaped fuzzy numbers for DO data for 2006

```matlab
function [xf, mu] = normfuzzy(xbar, s, x)

    f = normpdf(x, xbar, s);
    fa = 0.75*f;
    fb = 0.50*f;
    fc = 0.25*f;

    temp = sqrt(-2*s^2*log(fa*s*sqrt(2*pi)));
    xa1 = x - temp;
    xa2 = x + temp;
    clear vars temp

    temp = sqrt(-2*s^2*log(fb*s*sqrt(2*pi)));
    xb1 = x - temp;
    xb2 = x + temp;
    clear vars temp

    temp = sqrt(-2*s^2*log(fc*s*sqrt(2*pi)));
    xc1 = x - temp;
    xc2 = x + temp;
    clear vars temp

    x1 = x-3*s;
    x2 = x+3*s;

    a = [x1 xc1 xb1 xa1];
    a = sort(a);

    b = [xa2 xb2 xc2 x2];
    b = sort(b);

    xf = [a b];
    mu = [0 0.2500 0.5000 0.7500 1.0000 0.7500 0.5000 0.2500 0];
end
```

2. Estimating 2D velocity profile

```matlab
m = 10; %m is between 1 and length(Q)
Q = Q(m,1);
% corresponding estimated WS and v_ave velocity estimate from stage-discharge
% xls
WS = 996.4;
vtop = 1.10; % actually v_ave, labelled as such for computation consistency
```
% introduce calculated average fudge factor relating vmax(top) to vave
vtop = 3.5*vtop; % calculated by avergae values

%% calc dely
ymin = min(coords(:,2));
dely = (WS-ymin)/length(coords(:,2));
clear vars ymin

%% calc mid point
xmax = max(coords(:,1));
xmin = min(coords(:,1));
xmid = (xmax-xmin)/2+xmin
xmid = int8(xmid)

% use result for xmid to manually look for row number in coords
mid = 44
clear var xmin xmax xmid

%% calc ys at mid point
ybot = coords(mid,2);
y = WS:-dely:ybot;
ymid=y';
c = length(ymid);
clear vars ybot y

% calculate velocity at mid point
vmax = vtop;
v = vmax:-(vmax-0)/(c-1):0;
vmid=v';
clear vars v c

% calculate vtop at each x
% first make loop to check if WS>y
a = zeros(1,mid);
for b = 1:mid
  if(WS<coords(b,2));
    a(b) = 1;
  end
end
c = sum(a);
clear vars a b

n = length(coords(:,1));
n = n-mid;
vxl = 0:(vmax-0)/(mid-1-c):vmax;
vxl = [zeros(1,c) vxl];
clear vars c

a = zeros(1,n);
for b = mid:n
  if(WS<coords(b,2));
    a(b) = 1;
  end
end
c = sum(a);
clear vars a b

vx2 = vmax:-(vmax-0)/(n-1-c):0;
vx2 = [vx2 zeros(1,c)];

vtopx = [vx1 vx2];
clear vars n vx1 vx2 c

% calculate v at each y as above?
ybot = coords(:,2);
n = length(coords(:,1));
WS = repmat(WS,length(ybot),1);
c = zeros(n,1);
y = nan(n,1);

for i = 1:n
    yy = WS(i,1):-dely:ybot(i,1);
    [~, b] = size(yy);
    if(b>n)
        yy(b)=[];
    end
    clear var a b
    k = length(yy);
    l = n - k;
    q = padarray(yy,[0 1],nan,'post');
y(i,:) = q;
c(i) = length(yy);
    clear var yy k l q
end
y  =y';
c(c>n)=n;
clear var i ybot

% matrix v: each column is list of v from max (at ws) to 0 (at bed)
v = nan(n,n);
vmax = vtopx;
n = length(vmax);

for i = 1:n
    x = vmax(1,i):-(vmax(1,i)-0)/(c(i)-1):0;
    k = length(x);
    l = n - k;
    q = padarray(x,[0 1],nan,'post');
v(i,:)= q;
    clear var x k l q
end
v=v';
clear var i mid

x = coords(:,1);
x = x';
x = repmat(x,length(coords(:,1)),1);
% use velocities to calculate flow, use Q abd A_Q
a = x(1,:);
\[
\text{n = length(x(:,1));}
\text{delx = zeros(1,n-1);} \\
\text{for } i= 1:n-1 \\
\text{\quad delx(i)=a(i)-a(i+1);} \\
\text{end} \\
\text{delx=-delx;} \\
\%\text{delx=-repmat(delx,length(coords(:,1)),1);} \\
\text{clear vars a i n;}
\]
\[
\text{n = length(y(:,1));}
\text{delyy = zeros(n-1,n);} \\
\text{for } i= 1:n-1 \\
\text{\quad for j = 1:n} \\
\text{\quad \quad delyy(i,j)=y(i,j)-y(i+1,j);} \\
\text{end}
\]
\[
\text{clear vars i j n;}
\]
\[
\%\text{calcualte area}
\text{Area = zeros(length(delx));} \\
\text{for } i = 1:length(delx) \\
\text{\quad Area(:,i) = delx(i)*delyy(:,i);} \\
\text{end}
\]
\[
\text{A_check = nansum(nansum(Area));}
\]
\[
\%\text{Flow calcs}
\text{v(length(coords(:,1)),:)=[];} \\
\text{v(:,length(coords(:,1)))=[];} \\
\text{q = v.*Area;} \\
\text{qcheck =nansum(nansum(q));}
\]
\[
\text{f=Q/qcheck} \\
\text{\% increase average velocity by this amount and redo calc}
\\]
\[
\%\text{pad q to make same size as x and y}
\text{z = NaN(1,length(coords(:,1))-1);} \\
\text{q = [q; z];} \\
\text{v = [v; z];} \\
\text{z = NaN(length(coords(:,1)),1);} \\
\text{q = [q z];} \\
\text{v = [v z];}
\]
\[
\text{clear vars z;}
\]
Appendix C
Some definitions for Chapter 4

1. Fuzzy numbers

Fuzzy numbers are defined as a fuzzy subset of the real line \( \mathbb{R} \), i.e. a continuous mapping from \( \mathbb{R} \) to the closed interval [0, 1]. A fuzzy number \( \tilde{A} \) is a convex, normal fuzzy set, that has a piecewise continuous membership function \( \mu(x) \), such that the elements \( x_1, x_2, ..., x_m, ..., x_n \) of \( \tilde{A} \), where \( x_1 < x_2 < ... < x_m < x_n \), and the following hold true:

- For the intervals \( x \in (-\infty, x_1) \) and \( (x_m, \infty) \), \( \mu(x) = 0 \), \( \mu(x) \) monotonically increases between \( x_1 \) and \( x_m \) and monotonically decreases between \( x_m \) and \( x_n \).
- \( \mu(x_1) = \mu(x_n) = 0 \), and there exists one element \( x_m \) where \( \mu(x_m) = 1 \); this element is referred to as the mode value or the kernel.

A normal fuzzy set, as mentioned above, simply means that there is at least one element \( x_i \), where the \( \mu(x_i) = 1 \). The convexity constraint means that for any three (or more) elements, \( x_a, x_b, \) and, \( x_c \), where \( x_a < x_b < x_c \), the following holds true (Zhang et al., 2009, Wang et al., 2011):

\[
\mu(x_b) \geq \min(\mu(x_a), \mu(x_c))
\]

Equation C-1

The underlying concept of a fuzzy number is to define an imprecise quantity, thus \( x_m \) (i.e. the element where \( \mu = 1 \)) is the value that is closest to the true value being measured, while the values \( x_1 \) and \( x_n \) (where \( \mu = 0 \)), can be considered the upper and lower limits, beyond which are values that cannot be considered possible values of \( \tilde{A} \).

The membership function \( \mu(x) \) describes the degree of membership or membership level that each element has in a fuzzy set (or number). The functions model the gradual change in membership of an element in a fuzzy set. Typically, triangular (or also referred to as “L-R”) membership functions are used. In this case, \( \mu(x) \) is typically defined at three \( x_i \): \( \mu = 0^L \) (the left or lower bound), \( \mu = 1 \) (the kernel) and \( \mu = 0^R \) (the right or upper bound). A common notation for triangular fuzzy numbers is shown in the equation below:

\[
\tilde{A} = \{x_0, x_a^L, x_c^R\}
\]

Equation C-2
where $\tilde{A}$ is a fuzzy number, $x_b$ is the kernel of $\tilde{A}$, and $x_a^L$ and $x_c^R$ are the left and right bounds of $\tilde{A}$. Other shapes for membership functions have also been used, including uniform, trapezoidal and Gaussian shaped functions (Duch, 2005; Khan et al., 2013).

2. The $\alpha$-cut and the extension principle

The $\alpha$-cut technique is a method used to reduce a fuzzy number into a series of nested intervals that contain crisp elements (Zhang et al., 2009; Zhang & Achari, 2010b; Wang et al., 2011; Wang et al., 2012). It provides a bridge to connect a fuzzy set to a crisp set. Formally, the crisp set $A_\alpha$ is defined as a set that includes all elements $x_i$ that belong to a fuzzy set $\tilde{A}$ with a membership grade greater than or equal to $\alpha$, where $\alpha \in [0, 1]$. This set $A_\alpha$ is known as the ‘$\alpha$-level cut’ or simply the $\alpha$-cut of $\tilde{A}$. In other words, the set $A_\alpha$ contains all $x_i$ that are members of $\tilde{A}$ at degree of membership of at least $\mu(\tilde{A}) = \alpha$. The ordinary set $A_\alpha$ consists of an interval $[x_i^L, x_i^R] \in A_\alpha$ at some degree of membership $\alpha$. The $\alpha$-cut is described as:

$$A_\alpha = \{ x \in X, \mu(x) \geq \alpha \}$$

Equation C-3

The extension principle was introduced by Zadeh (1975) and provides a method for extending crisp mathematical functions to deal with fuzzy quantities. Suppose that a function $f$ acts on a fuzzy number $\tilde{A}$, which is defined by the elements $x_1, ..., x_n$, and that their corresponding membership levels are $\mu_{\tilde{A}x_1}, ..., \mu_{\tilde{A}x_n}$. The fuzzy number $\tilde{N}$ is generated by $f$ acting on $\tilde{A}$ (i.e. $\tilde{N} = f(\tilde{A})$) and consists of elements $y_1, ..., y_n$, where $y_i = f(x_i)$. The corresponding $\mu$ describing the membership of elements $y_i$ in $\tilde{N}$ are unaffected by the mapping $f$, and are given by $\mu_{\tilde{A}x_1}, ..., \mu_{\tilde{A}x_n}$. This principle can be generalised to multiple input fuzzy numbers as:

$$\mu_{\tilde{N}}(y) = \sup \min \{ \mu_{\tilde{A}x_1}(x_1), \mu_{\tilde{A}x_2}(x_2), ..., \mu_{\tilde{A}x_n}(x_n) \}$$

Equation C-4

where $\tilde{A}_1$ to $\tilde{A}_n$ are a series of fuzzy numbers, $x_1$ to $x_n$ are their corresponding elements, each with a membership level of $\mu_{\tilde{A}x_1}(x_1)$ to $\mu_{\tilde{A}x_n}(x_n)$, and $\mu_{\tilde{N}}(y)$ is the membership level of $y$, which is an element in the fuzzy number $\tilde{N}$, and is generated from a mapping $f$ on the elements $x_1$ to $x_n$. The term “sup” is the supremum operator and “min” is the minimum (Zadeh, 1975; Dubois & Prade, 1980). This implies that the membership levels of the images of elements are not affected by the function $f$. This effectively reduces
mathematical operations on the fuzzy number \( \tilde{A} \), into a series of operations on the ordinary sets \( A_\alpha \) (the \( \alpha \)-cut of \( \tilde{A} \) at \( \mu = \alpha \)). A number of standardised mathematical operations have been defined to use with fuzzy numbers; the reader is referred to Kauffman & Gupta (1984) for more details.

3. Probability–possibility transformation

In order to represent crisp data as fuzzy numbers, the relationship between possibility and probability is used to transform the data. This transformation is referred to as a probability-possibility transformation (Dubois et al., 1993, 2004; Zhang et al., 2009). It converts a uni-modal probability density function (pdf) \( p(x) \) with bounded support to a possibility distribution \( \mu(x) \). The transformation relates the area under a pdf to a membership value. Details on the theoretical background relating the possibility and probability distributions of a variable are not included here, but can be found Dubois et al. (1993, 2004), amongst others. The transformation is given by:

\[
\mu(x) = u(f(x)) = \int_{-\infty}^{x} p(y)dy \int_{f(x)}^{\infty} p(y)dy
\]

where the kernel of \( \mu(x) \) is the mode \( M \) of \( p(x) \) and \( f(x) \) is a mapping where for any \( x \),

\[f(x) = y \geq M, \text{ such that } p(x) = p(y).\]

It is subject to the constraint:

\[f(x) = \max \{ y | p(y) \geq p(x) \}\]

In simple terms, consider a pdf of \( x, p(x) \) that is bounded by two values \( x_{\min} \) and \( x_{\max} \). A horizontal cut is made at \( p(x) = H \), resulting in two values of \( x, x^L = p^{-1}(H) \) and \( x^R = p^{-1}(H) \). The sum of the cumulative probability between \( x_{\min} \) and \( x^L \) and between \( x^R \) and \( x_{\max} \), is some value \( \alpha \). This \( \alpha \) is the minimum membership level for all values between \( x^L \) and \( x^R \), or in other words represents the interval or \( \alpha \)-cut given by \( A_\alpha = [x^L, x^R]_\alpha \). This procedure is repeated for all \( H \), between 0 and \( \max \{ p(x) \} \), resulting in a series of \( \alpha \)-cuts at different \( \alpha \). These values together define the membership function \( \mu \) of the fuzzy number \( \tilde{A} \):

\[\tilde{A} = \cup [x^L, x^R]_\alpha\]

where the symbol \( \cup \) represent the union of all \( \alpha \)-cuts.
Appendix D
MATLAB code for Chapter 4

1. Constructing linear fuzzy numbers

```matlab
function [mu, x] = muLinear(X)

delta = (max(X) - min(X))/length(X);
mini = min(X) - delta;
maxi = max(X) + delta;

[a, b] = hist(X, mini:delta:maxi);
% a are the counts
% b are the intervals; dependent on the c in hist(X, c)
X1 = [a; b]';

M = max(X1(:,1));
k = 1;
maxval = zeros(1,1);
    for i =1:length(a)
        if X1(i,1) == M
            maxval(k) = X1(i,2);
            k = k +1;
        end
    end
clear vars k i a b M X1

xmin = min(X);
xmax = max(X);
xmean = mean(maxval);
clear vars X

sL = (1-0)/(xmean - xmin);
c = -sL*(xmin);
y = [0.25; 0.5; 0.75];
xL = 1/sL.*(y - c*ones(length(y),1));
clear vars c y sL

sR = (1-0)/(xmean - xmax);
c = -sR*(xmax);
y = [0.75; 0.5; 0.25];
xR = 1/sR.*(y - c*ones(length(y),1));
clear vars c y sR

x = [xmin xL' xmean xR' xmax]';
mu = [0; 0.25; 0.5; 0.75; 1; 0.75; 0.5; 0.25; 0];
```

2. Constructing non-linear fuzzy numbers

```matlab
function [mu, x] = muPossible(X, e)

% X is a 1 dimensional vector of observations
% X is coverted into the correct format
% e is a vector (2,1) of upper and lower errors associated with X

[m, n] = size(X);
if m == 1
    X = X';
end

clear vars m n

% calc probabilities
delta = (max(X)-min(X))/length(X);
x = min(X)-delta-e(1):delta:max(X) + delta + e(2);
x = x';
n = histc(X,x);
p = n/sum(n);
X1 = [n x];
clear vars e

% find corresponding x max p
nmax = max(X1(:,1));
k = 1;
xmid = zeros(1,1);
for i = 1:length(n)
    if X1(i,1) == nmax
        xmid(k) = X1(i,2);
        k = k +1;
    end
end
xmid = mean(xmid);
clear vars k i X1 nmax

if isnan(xmid)
    x = [NaN NaN NaN NaN NaN NaN NaN NaN NaN];
    mu = [0 0.25 0.5 0.75 1 0.75 0.5 0.25 0];
else

% calculate values and corresponding index of points midway between xmin
% and xmid, and xmid and xmax

i = find(x == xmid);
if isempty(i)
    i = find(x<xmid);
i = i(end)+1;
end

xl = (x(i) + min(X))/2;
if xl<min(X)
    xl = min(X);
end

xr = (max(X) + x(i))/2;
clear vars i

i = find(x == xl);
if isempty(i)
    i = find(x<xl);
    i = i(end);
end

j = find(x == xr);
if isempty(j)
    j = find(x<xr);
    j = j(end);
end

% figure out xLs and xRs
xl = x(1:i);
pl = p(1:i);
Pl = sum(pl);
xr = x(j:end);
pr = p(j:end);
Pr = sum(pr);
clear vars i j pl pr

alpha1 = Pl + Pr;
clear vars Pl Pr nmax

% next level, midway between 0 and midway
n = length(xl);
n = round(n/2);
xl1 = xl(1:n);
pl = p(1:n);
Pl = sum(pl);
clear vars n

n = length(xr);
n = round(n/2);
xr1 = xr(end-(n-1):end);
pr = p(end-(n-1):end);
Pr = sum(pr);
clear vars n

alpha2 = Pl + Pr;
clear vars Pl Pr pl pr

% next level, midway between midway and max
a = (xmid + xl(end))/2;
xl2 = xl:delta:a;
pl = p(1:length(xl2));
Pl = sum(pl);
clear vars a pl

a = (xr(1) + xmid)/2;
xr2 = a:delta:xr(end);
n = length(xr2);
pr = p(end-n+1:end);
Pr = sum(pr);
clear vars a n pr
alpha3 = Pl + Pr;
clear vars Pl Pr

if alpha3 == 1
    alpha3 = alpha1;
end

% interpolation
L = [xl(1) xl1(end) xl(end) xl2(end) xmid];
R = [xmid xr2(1) xr(1) xr1(1) xr(end)];

[XL,XR] = muInterp(L, R, alpha1, alpha2, alpha3);
clear vars xl xl1 xl2 xr xr1 xr2 xmid p alpha1 alpha2 alpha3 L R delta

mu = [0 0.25 0.5 0.75 1 0.75 0.5 0.25 0];
x = [XL XR(2:end)];
clear vars XL XR
end

function [XL, XR] = muInterp(L, R, alpha1, alpha2, alpha3)

% this is a nested function in muPossible, used to interpolate values
% of
% the membership function, calculated via prob-possib tranformation
% four major rules are defined for when non-boundary alpha values are
equal

% if adjacent values in input L and R are equal, use this to prevent
% computational problems

% for i = 1:length(L)-1
%     if L(i) == L(i+1)
%         L(i) = L(i) - 0.0001;
%     end
% end
% for i = 1:length(R)-1
%     if R(i) == R(i+1)
%         R(i) = R(i) - 0.0001;
%     end
% end
% clear vars i

% RULE 1: alpha2 == alpha1 && alpha1 == alpha3
if alpha2 == alpha1 && alpha1 == alpha3
    if alpha3<=0.25
        muL = [alpha3 1];
        L1 = [L(4) L(5)];
        XL1 = interp1(muL, L1, 0.25);
        XL2 = interp1(muL, L1, 0.5);
        XL3 = interp1(muL, L1, 0.75);
clear vars muL L1
XL = [L(1) XL1 XL2 XL3 L(5)];
clear vars XL1 XL2 XL3
muR = [1 alpha3];
R1 = [R(1) R(2)];
XR1 = interp1(muR, R1, 0.25);
XR2 = interp1(muR, R1, 0.5);
XR3 = interp1(muR, R1, 0.75);
clear vars muR R1
XR = [R(1) XR3 XR2 XR1 R(5)];
clear vars XR1 XR2 XR3
else
if 0.25<alpha3 && alpha3<=0.5
muL = [alpha3 1];
L1 = [L(4) L(5)];
XL2 = interp1(muL, L1, 0.5);
XL3 = interp1(muL, L1, 0.75);
clear vars muL L1
muL = [0 alpha2];
L1 = [L(1) L(2)];
XL1 = interp1(muL, L1, 0.25);
clear vars muL L1
XL = [L(1) XL1 XL2 XL3 L(5)];
clear vars XL1 XL2 XL3
muR = [1 alpha3];
R1 = [R(1) R(2)];
XR2 = interp1(muR, R1, 0.5);
XR3 = interp1(muR, R1, 0.75);
clear vars muR R1
muR = [alpha2 0];
R1 = [R(4) R(5)];
XR1 = interp1(muR, R1, 0.25);
clear vars muR R1
XR = [R(1) XR3 XR2 XR1 R(5)];
clear vars XR1 XR2 XR3
else
if 0.5<alpha3 && alpha3<=0.75
muL = [alpha3 1];
L1 = [L(4) L(5)];
XL3 = interp1(muL, L1, 0.75);
clear vars muL L1
muL = [0 alpha2];
L1 = [L(1) L(2)];
XL1 = interp1(muL, L1, 0.25);
XL2 = interp1(muL, L1, 0.5);
clear vars L1 muL
XL = [L(1) XL1 XL2 XL3 L(5)];
clear vars XL1 XL2 XL3
muR = [1 alpha3];
R1 = [R(1) R(2)];
XR3 = interp1(muR, R1, 0.75);
clear vars muR LR
muR = [alpha2 0];
R1 = [R(4) R(5)];
XR1 = interp1(muR, R1, 0.25);
XR2 = interp1(muR, R1, 0.5);
clear vars LR muR
\[ XR = [R(1) \ XR3 \ XR2 \ XR1 \ R(5)]; \]
\[ clear \ vars \ XR1 \ XR2 \ XR3 \]

\textbf{else}

\begin{verbatim}
if \( \alpha_3 > 0.75 \)
    \[ muL = [0 \ \alpha_2]; \]
    \[ L1 = [L(1) \ L(2)]; \]
    \[ XL1 = \text{interp1}(muL, L1, 0.25); \]
    \[ XL2 = \text{interp1}(muL, L1, 0.5); \]
    \[ XL3 = \text{interp1}(muL, L1, 0.75); \]
    \[ clear \ vars \ muL \ L1 \]
    \[ XL = [L(1) \ XL1 \ XL2 \ XL3 \ L(5)]; \]
    \[ clear \ vars \ XL1 \ XL2 \ XL3 \]
    \[ muR = [\alpha_2 \ 0]; \]
    \[ R1 = [R(4) \ R(5)]; \]
    \[ XR1 = \text{interp1}(muR, R1, 0.25); \]
    \[ XR2 = \text{interp1}(muR, R1, 0.5); \]
    \[ XR3 = \text{interp1}(muR, R1, 0.75); \]
    \[ clear \ vars \ muR \ R1 \]
    \[ XR = [R(1) \ XR3 \ XR2 \ XR1 \ R(5)]; \]
    \[ clear \ vars \ XR1 \ XR2 \ XR3 \]
\end{verbatim}

\textbf{end}

\textbf{end}

\textbf{end}

\textbf{if} \( \alpha_3 == 0 \)

\begin{verbatim}
    XL(1) = L(4);
    XR(end) = R(end-3);
\end{verbatim}

\textbf{end}

\textbf{\% RULE 2:} \( \alpha_2 = \alpha_1, \ \alpha_3 > \alpha_1 \)

\textbf{if} \( \alpha_2 == \alpha_1 \ \&\& \ \alpha_3 > \alpha_1 \)

\begin{verbatim}
if \( \alpha_2 <= 0.25 \)
    \[ muL = [\alpha_1 \ \alpha_3 \ 1]; \]
    \[ L1 = [L(3) \ L(4) \ L(5)]; \]
    \[ XL1 = \text{interp1}(muL, L1, 0.25); \]
    \[ XL2 = \text{interp1}(muL, L1, 0.5); \]
    \[ XL3 = \text{interp1}(muL, L1, 0.75); \]
    \[ clear \ vars \ muL \ L1 \]
    \[ XL = [L(1) \ XL1 \ XL2 \ XL3 \ L(5)]; \]
    \[ clear \ vars \ XL1 \ XL2 \ XL3 \]
    \[ muR = [1 \ \alpha_3 \ \alpha_1]; \]
    \[ R1 = [R(1) \ R(2) \ R(3)]; \]
    \[ XR1 = \text{interp1}(muR, R1, 0.25); \]
    \[ XR2 = \text{interp1}(muR, R1, 0.5); \]
    \[ XR3 = \text{interp1}(muR, R1, 0.75); \]
    \[ clear \ vars \ muR \ R1 \]
    \[ XR = [R(1) \ XR3 \ XR2 \ XR1 \ R(5)]; \]
    \[ clear \ vars \ XR1 \ XR2 \ XR3 \]
\end{verbatim}

\textbf{if} \( \alpha_1 == 0 \)

\begin{verbatim}
    XL(1) = L(3);
    XR(end) = R(end-2);
\end{verbatim}

\textbf{end}

\textbf{else}

\begin{verbatim}
if \( 0.25<\alpha_2 \ \&\& \ \alpha_2<=0.5 \)
    \[ muL = [\alpha_1 \ \alpha_3 \ 1]; \]
    \[ L1 = [L(3) \ L(4) \ L(5)]; \]
    \[ XL2 = \text{interp1}(muL, L1, 0.5); \]
\end{verbatim}

\textbf{end}

\textbf{end}
XL3 = interp1(muL, L1, 0.75);
    clear vars muL L1
muL = [0 alpha2];
L1 = [L(1) L(2)];
XL1 = interp1(muL, L1, 0.25);
    clear vars muL L1
XL = [L(1) XL1 XL2 XL3 L(5)];
    clear vars XL1 XL2 XL3
muR = [1 alpha3 alpha1];
R1 = [R(1) R(2) R(3)];
XR2 = interp1(muR, R1, 0.5);
XR3 = interp1(muR, R1, 0.75);
    clear vars muR R1
muR = [alpha2 0];
R1 = [R(4) R(5)];
XR1 = interp1(muR, R1, 0.25);
    clear vars muR R1
XR = [R(1) XR3 XR2 XR1 R(5)];
    clear vars XR1 XR2 XR3
else
    if 0.5<alpha2 && alpha2<=0.75
        muL = [alphal alpha3 1];
        L1 = [L(3) L(4) L(5)];
        XL3 = interp1(muL, L1, 0.75);
            clear vars muL L1
        muL = [0 alpha2];
        L1 = [L(1) L(2)];
        XL1 = interp1(muL, L1, 0.25);
        XL2 = interp1(muL, L1, 0.5);
            clear vars muL L1
        XL = [L(1) XL1 XL2 XL3 L(5)];
            clear vars XL1 XL2 XL3
        muR = [1 alpha3 alpha1];
        R1 = [R(1) R(2) R(3)];
        XR3 = interp1(muR, R1, 0.75);
            clear vars muR R1
        muR = [alpha2 0];
        R1 = [R(4) R(5)];
        XR1 = interp1(muR, R1, 0.25);
        XR2 = interp1(muR, R1, 0.5);
            clear vars muR R1
        XR = [R(1) XR3 XR2 XR1 R(5)];
            clear vars XR1 XR2 XR3
    else
        if alpha2>0.75
            muL = [0 alpha2];
            L1 = [L(1) L(2)];
            XL1 = interp1(muL, L1, 0.25);
            XL2 = interp1(muL, L1, 0.5);
            XL3 = interp1(muL, L1, 0.75);
                clear vars muL L1
            XL = [L(1) XL1 XL2 XL3 L(5)];
                clear vars XL1 XL2 XL3
            muR = [alpha2 0];
            R1 = [R(4) R(5)];
            XR1 = interp1(muR, R1, 0.25);
            XR2 = interp1(muR, R1, 0.5);
XR3 = interp1(muR, R1, 0.75);
clear vars muR R1
XR = [R(1) XR3 XR2 XR1 R(5)];
clear vars XR1 XR2 XR3
end
end
end
end
end
end

% RULE 3: alpha1 = alpha3, alpha2 < alpha1
if alpha1 == alpha3 && alpha1 > alpha2
    if alpha1 <= 0.25
        muL = [alpha3 1];
        L1 = [L(4) L(5)];
        XL1 = interp1(muL, L1, 0.25);
        XL2 = interp1(muL, L1, 0.5);
        XL3 = interp1(muL, L1, 0.75);
        clear vars muL L1
        XL = [L(1) XL1 XL2 XL3 L(5)];
        clear vars XL1 XL2 XL3
        muR = [1 alpha3];
        R1 = [R(1) R(2)];
        XR1 = interp1(muR, R1, 0.25);
        XR2 = interp1(muR, R1, 0.5);
        XR3 = interp1(muR, R1, 0.75);
        clear vars muR R1
        XR = [R(1) XR3 XR2 XR1 R(5)];
        clear vars XR1 XR2 XR3
    end
    else
        if 0.25 < alpha1 && alpha1 <= 0.5
            muL = [alpha3 1];
            L1 = [L(4) L(5)];
            XL2 = interp1(muL, L1, 0.5);
            XL3 = interp1(muL, L1, 0.75);
            clear vars muL L1
            muL = [0 alpha2 alpha1];
            L1 = [L(1) L(2) L(3)];
            if alpha2 == 0
                muL = muL(2:end);
                L1 = L1(2:end);
            end
        end
        XL1 = interp1(muL, L1, 0.25);
        clear vars muL L1
        XL = [L(1) XL1 XL2 XL3 L(5)];
        clear vars XL1 XL2 XL3
        muR = [1 alpha3];
        R1 = [R(1) R(2)];
        XR2 = interp1(muR, R1, 0.5);
        XR3 = interp1(muR, R1, 0.75);
        clear vars muR R1
        muR = [alpha1 alpha2 0];
        R1 = [R(3) R(4) R(5)];
        if alpha2 == 0
            clear vars muR R1
            XR = [R(1) XR3 XR2 XR1 R(5)];
            clear vars XR1 XR2 XR3
        end
    end
end
muR = muR(1:end-1);
R1 = R1(1:end-1);
end
XR1 = interp1(muR, R1, 0.25);
clear vars muR R1
XR = [R(1) XR3 XR2 XR1 R(5)];
clear vars XR1 XR2 XR3
if alpha2 == 0
    XL(1) = L(2);
    XR(end) = R(end-1);
end
else
    if 0.5<alpha1 && alpha1<=0.75
        muL = [alpha3 1];
        L1 = [L(4) L(5)];
        XL3 = interp1(muL, L1, 0.75);
        clear vars muL L1
        muL = [0 alpha2 alpha1];
        L1 = [L(1) L(2) L(3)];
        if alpha2==0
            muL = muL(2:end);
            L1 = L1(2:end);
        end
        XL1 = interp1(muL, L1, 0.25);
        XL2 = interp1(muL, L1, 0.5);
        clear vars muL L1
        XL = [L(1) XL1 XL2 XL3 L(5)];
        clear vars XL1 XL2 XL3
        muR = [1 alpha3];
        R1 = [R(1) R(2)];
        XR3 = interp1(muR, R1, 0.75);
        clear vars muR R1
        muR = [alpha1 alpha2 0];
        R1 = [R(3) R(4) R(5)];
        if alpha2 == 0
            muR = muR(1:end-1);
            R1 = R1(1:end-1);
        end
        XR1 = interp1(muR, R1, 0.25);
        XR2 = interp1(muR, R1, 0.5);
        clear vars muR R1
        XR = [R(1) XR3 XR2 XR1 R(5)];
        clear vars XR1 XR2 XR3
        if alpha2 == 0
            XL(1) = L(2);
            XR(end) = R(end-1);
        end
    else
        if alpha1>0.75
            muL = [0 alpha2 alpha1];
            L1 = [L(1) L(2) L(3)];
            if alpha2==0
                muL = muL(2:end);
                L1 = L1(2:end);
            end
            XL1 = interp1(muL, L1, 0.25);
            XL2 = interp1(muL, L1, 0.5);
        else
            ...
XL3 = interp1(muL, L1, 0.75);
clear vars muL L1
XL = [L(1) XL1 XL2 XL3 L(5)];
clear vars XL1 XL2 XL3
muR = [alpha1 alpha2 0];
R1 = [R(3) R(4) R(5)];
if alpha2 == 0
muR = muR(1:end-1);
R1 = R1(1:end-1);
end
XR1 = interp1(muR, R1, 0.25);
XR2 = interp1(muR, R1, 0.5);
XR3 = interp1(muR, R1, 0.75);
clear vars muR R1
XR = [R(1) XR3 XR2 XR1 R(5)];
clear vars XR1 XR2 XR3
if alpha2 == 0
    XL(1) = L(2);
    XR(end) = R(end-1);
end

% RULE 4: alpha2 < alpha1 < alpha3
if alpha2 < alpha1 && alpha1 < alpha3
    if alpha2 == 0
        muL = [0 alpha1 alpha3 1];
        L1 = L(2:end);
        XL1 = interp1(muL, L1, 0.25);
        XL2 = interp1(muL, L1, 0.5);
        XL3 = interp1(muL, L1, 0.75);
        clear vars muL L1
        XL = [L(1) XL1 XL2 XL3 L(5)];
        clear vars XL1 XL2 XL3
        muR = [1 alpha3 alpha1 0];
        R1 = R(1:end-1);
        XR1 = interp1(muR, R1, 0.25);
        XR2 = interp1(muR, R1, 0.5);
        XR3 = interp1(muR, R1, 0.75);
        clear vars muR R1
        XR = [R(1) XR3 XR2 XR1 R(5)];
        clear vars XR1 XR2 XR3
    else
        if alpha3 == 1
            muL = [0 alpha2 alpha1 1];
            L1 = L(1:end-1);
            XL1 = interp1(muL, L1, 0.25);
            XL2 = interp1(muL, L1, 0.5);
            XL3 = interp1(muL, L1, 0.75);
            clear vars muL L1
            XL = [L(1) XL1 XL2 XL3 L(5)];
            clear vars XL1 XL2 XL3
            muR = [1 alpha1 alpha2 0];
            R1 = R(2:end);
            XR1 = interp1(muR, R1, 0.25);
XR2 = interp1(muR, R1, 0.5);
XR3 = interp1(muR, R1, 0.75);
clear vars muR R1
XR = [R(1) XR3 XR2 XR1 R(5)];
clear vars XR1 XR2 XR3
else
muL = [0 alpha2 alpha1 alpha3 1];
L1 = L;
XL1 = interp1(muL, L1, 0.25);
XL2 = interp1(muL, L1, 0.5);
XL3 = interp1(muL, L1, 0.75);
clear vars muL L1
XL = [L(1) XL1 XL2 XL3 L(5)];
clear vars XL1 XL2 XL3
muR = [1 alpha3 alpha1 alpha2 0];
R1 = R;
XR1 = interp1(muR, R1, 0.25);
XR2 = interp1(muR, R1, 0.5);
XR3 = interp1(muR, R1, 0.75);
clear vars muR R1
XR = [R(1) XR3 XR2 XR1 R(5)];
clear vars XR1 XR2 XR3
end
end
if alpha2 == 0
    XL(1) = L(2);
    XR(end) = R(end-1);
end
end

% if adjacent elements are equal, use this to "fix"...
% for i = 1:length(XL)-1
%     if abs(XL(i) - XL(i+1)) < 0.0001
%         XL(i) = XL(i) - 0.0001;
%     end
% end
% for i = 1:length(XR)-1
%     if abs(XR(i) - XR(i+1)) < 0.0001
%         XR(i) = XR(i) - 0.0001;
%     end
% end
% clear vars i
End

3. Fuzzy linear regression – Tanaka

function [A0, A1, A2] = tanaka(x, y, z, e, H)
% calculates possibility fuzzy regression using the Tanaka et al. (1982).
% Input includes crisp X and Y, fuzzy Z, where Z is the modal value and
% e is the half width (i.e. symmetrical trinangular fuzzy number).
% H a value between 0 and 1, known as the h-certain factor. 1 encompasses
% all possibilities, 0 is most stringent (might not find a solution)
% objective is to find minimum A0, A1 and A2, which are triangular fuzzy
% numbers and coefficients to the equation:
\[
(Z, e) = (A_0, a_0) + (A_1, a_1)X + (A_2, a_2)Y
\]
\[
data = \{x \ y \ z \ e\};
\]
% check for NaN
% c1 = sum(isnan(x));
% c2 = sum(isnan(y));
% c3 = sum(isnan(z));
% c4 = sum(isnan(e));

% if NaN exist remove entire row from 'data' matrix:
data = data(isfinite(data(:, 3)), :);

clear vars x y z e

n = length(data(:,1));
f = [0; 1; 0; 1; 0; 1];
% f = [A_0, a_0, A_1, a_1, A_2, a_2]
% here 0s associates with the centre values of A0, A1 and A2,
% the 1s are associated with a0, a1 and a2, the fuzzy half widths
% objective is to minimize f, given A\*x < b

% matrix A
A = zeros(2\*n, 6);
A(1:2:end,1) = -1;
A(2:2:end,1) = 1;
A(1:2:end,2) = -(1 - H);
A(2:2:end,2) = -(1 - H);
A(1:2:end,3) = -data(:,1);
A(2:2:end,3) = +data(:,1);
A(1:2:end,4) = -(1 - H)*data(:,1);
A(2:2:end,4) = -(1 - H)*data(:,1);
A(1:2:end,5) = -data(:,2);
A(2:2:end,5) = +data(:,2);
A(1:2:end,6) = -(1 - H)*data(:,2);
A(2:2:end,6) = -(1 - H)*data(:,2);

% matrix b
b = zeros(2\*n, 1);
b(1:2:end) = -data(:,3) - (1 - H)*data(:,4);
b(2:2:end) = +data(:,3) - (1 - H)*data(:,4);

clear vars data n
% lower and upper bounds for A0, a0, A1, a1, A2 and a2
% only constraint is that the half-widths cannot be negative
l = [-Inf; 0; -Inf; 0; -Inf; 0];
u = [Inf; Inf; Inf; Inf; Inf; Inf];
% run linprog(f, A, b, [], [], lb, up); this solves f'*x, such that
A*x<=b
% [] and [] are empty, since there are no equality constraints
% lb, up and lower and upper bounds respectively
% default optimization solver used by linprog is an interior-point
% method,
% suitable for this application
X = linprog(f,A,b,[],[],l,u);

clear vars f A b l u

% Output of X gives the values of f = [A0, a0, A1, a1, A2, a2]
% the values are used to calculate the following fuzzy numbers
A0 = [X(1)-X(2) X(1) X(1)+X(2)];
A1 = [X(3)-X(4) X(3) X(3)+X(4)];
A2 = [X(5)-X(6) X(5) X(5)+X(6)];
clear vars X
end

4. Fuzzy linear regression – Diamond

function [E, EL, ER, b, c] = diamond(Y, YL, YR, X1, X1L, X1R, X2, X2L, X2R)

% solves the minimization problem F2 adapted from Diamond (1988):
% min r(Ef, b, c) = d^2(Yf, Ef + bX1f + cX2f); where f denotes fuzzy-number
% inputs include values for Y, X1 and X2 as well as their left (L) and
% right (R) extremes

% input values be column vectors
N = length(Y); % N = number of observations

% calculate average values
X1bar  = sum(X1)/N;
X1Lbar = sum(X1L)/N;
X1Rbar = sum(X1R)/N;
X2bar  = sum(X2)/N;
X2Lbar = sum(X2L)/N;
X2Rbar = sum(X2R)/N;
Ybar  = sum(Y)/N;
YLbar = sum(YL)/N;
YRbar = sum(YR)/N;

% calulate sqaured averages
X12bar  = sum(X1.^2)/N;
X1L2bar = sum(X1L.^2)/N;
X1R2bar = sum(X1R.^2)/N;
X22bar  = sum(X2.^2)/N;
X2L2bar = sum(X2L.^2)/N;
X2R2bar = sum(X2R.^2)/N;
% calculate variables alpha to mu
alpha = X12bar + X1L2bar + X1R2bar;
beta = sum(X1.*X2)/N + sum(X1L.*X2L)/N + sum(X1R.*X2R)/N;
gamma = sum(Y.*X1)/N + sum(YL.*X1L)/N + sum(YR.*X1R)/N;
delta = X1bar.*X1bar + X1Lbar.*X1Lbar + X1Rbar.*X1Rbar;
epsilon = X2bar.*X1bar + X2Lbar.*X1Lbar + X2Rbar.*X1Rbar;
etta = Ybar.*X1bar + YLbar.*X1Lbar + YRbar.*X1Rbar;
theta = X2bar + X2Lbar + X2Rbar;
kappa = sum(Y.*X2)/N + sum(YL.*X2L)/N + sum(YR.*X2R)/N;
lambda = X2bar.*X2bar + X2Lbar.*X2Lbar + X2Rbar.*X2Rbar;
mu = Ybar.*X2bar + YLbar.*X2Lbar + YRbar.*X2Rbar;

clear vars X12bar X1L2bar X1R2bar X22bar X2L2bar X2R2bar N

% create matrix to solve for b and c
% matrix notation used is Ax = B, where x is a vector [b; c]
A = [alpha - delta beta - epsilon; beta - epsilon theta - lambda];
B = [gamma - eta; kappa - mu];
x = A\B; % or x = inv(A)*B
b = x(1);
c = x(2);

clear vars A B x

E = Ybar - b*X1bar - c*X2bar;
EL = YLbar - b*X1Lbar - c*X2Lbar;
ER = YRbar - b*X1Rbar - c*X2Rbar;

clear vars X1bar X1Lbar X1Rbar X2bar X2Lbar X2Rbar Ybar YLbar YRbar
end

5. Fuzzy linear regression - Khan

function [A0, A1, A2] = usman02(X1f, X2f, Yf)
% Input are fuzzy numbers X1, X2 and Yf, each N by 9 elements in size,
% corresponding to the membership function:
% mu = [0 0.25 0.5 0.75 1 0.75 0.5 0.25 0]
% Output are the three parameters resulting from least-squares curve fitting

% sort input matrices into vectors
X1_aL = X1f(:,1);
X1_bL = X1f(:,2);
X1_cL = X1f(:,3);
X1_dL = X1f(:,4);
X1_l  = X1f(:,5);
X1_dR = X1f(:,6);
X1_cR = X1f(:,7);
X1_bR = X1f(:,8);
X1_aR = X1f(:,9);
clear vars X1f

X2_aL = X2f(:,1);
X2_bL = X2f(:,2);
X2_cL = X2f(:,3);
X2_dL = X2f(:,4);
X2_1  = X2f(:,5);
X2_dR = X2f(:,6);
X2_cR = X2f(:,7);
X2_bR = X2f(:,8);
X2_aR = X2f(:,9);

clear vars X2f

Y_aL = Yf(:,1);
Y_bL = Yf(:,2);
Y_cL = Yf(:,3);
Y_dL = Yf(:,4);
Y_1  = Yf(:,5);
Y_dR = Yf(:,6);
Y_cR = Yf(:,7);
Y_bR = Yf(:,8);
Y_aR = Yf(:,9);

clear vars Yf

% least squares

N = length(X1_1);
X_aL = [ones(N,1) X1_aL X2_aL];
X_bL = [ones(N,1) X1_bL X2_bL];
X_cL = [ones(N,1) X1_cL X2_cL];
X_dL = [ones(N,1) X1_dL X2_dL];
X_1  = [ones(N,1) X1_1 X2_1 ];
X_dR = [ones(N,1) X1_dR X2_dR];
X_cR = [ones(N,1) X1_cR X2_cR];
X_bR = [ones(N,1) X1_bR X2_bR];
X_aR = [ones(N,1) X1_aR X2_aR];
XX = zeros(9*N, 9*3);
XX(1:N,1:3) = X_aL;
XX(N+1:2*N,4:6) = X_bL;
XX(2*N+1:3*N,7:9) = X_cL;
XX(3*N+1:4*N,10:12) = X_dL;
XX(4*N+1:5*N,13:15) = X_1 ;
XX(5*N+1:6*N,16:18) = X_dR;
XX(6*N+1:7*N,19:21) = X_cR;
XX(7*N+1:8*N,22:24) = X_bR;
XX(8*N+1:9*N,25:27) = X_aR;
YY = [Y_aL; Y_bL; Y_cL; Y_dL; Y_1 ; Y_dR; Y_cR; Y_bR; Y_aR];
BETA = (XX'*XX) \ XX'*YY;
A0 = BETA(1:3:end);
A1 = BETA(2:3:end);
A2 = BETA(3:3:end);
end
Appendix E
Bayesian linear regression explained

1. Introduction to Bayesian linear regression

Bayesian linear regression is an approach to ordinary linear regression (OLR) within a Bayesian framework. The OLR model is typically denoted as:

\[ y = \beta_0 + \beta_1 x + \varepsilon \]  \hspace{1cm} \text{Equation E-1}

where \( x \) are the explanatory variables, \( y \) the dependent variables, \( \beta_0 \) and \( \beta_1 \) the regression coefficients, and \( \varepsilon \) are the regression error, assumed to be independent and identically distributed, expressed as \( \sim N(0, \sigma^2) \). Another assumption of OLR is that the \( x \) are independent of \( \varepsilon \). The objective of OLR is to calculate values of \( \beta_0 \) and \( \beta_1 \) that minimise (in the least–square sense) the error between the observed values of \( y \) and \( \beta_0 \) and \( \beta_1 x \). In a Bayesian framework, this model structure can also be represented as a Normal probability model:

\[ y \sim N(X\beta, \sigma^2 I) \]  \hspace{1cm} \text{Equation E-2}

where \( X \) is matrix of dependent variables, limited to \([x_1 \ x_2]\) here with \( x_1 = I \) to allow for an intercept, and \( x_2 \) are the observations, the \( \beta \) are the corresponding regression coefficients, \( \sigma^2 \) is the variance of the dependent variable \( y \) given the observations (i.e. \( y|\beta, X \), and lastly, \( I \) is the identity matrix. In a Bayesian framework, this function \( (y|\beta, \sigma^2, X) \) is the likelihood function. Note that Bayes rule states that the posterior distribution is proportional to the product of the likelihood function and the prior distribution. Thus, OLR when viewed through a Bayesian framework can be written as:

\[ p(\beta, \sigma^2 | y, X) \propto p(y|\beta, \sigma^2, X)p(\beta, \sigma^2) \]  \hspace{1cm} \text{Equation E-3}

where the term on the left hand side represents the posterior, and the right hand side is the product of the likelihood and the prior. The likelihood in this case is a multivariate Normal, and is often more convenient to be parameterised using the following terms (the derivation is not shown here but available in many econometric textbooks including
Zellner, 1971; Birkes & Dodge, 1993; Koop, 2003; Greenberg, 2008; Gelman et al., 2014):

\[
\hat{\beta} = (X^T X)^{-1}X^T y \quad \text{Equation E-4}
\]
\[
s^2 = (y - X\hat{\beta})^T (y - X\hat{\beta})/\nu \quad \text{Equation E-5}
\]

These values are identical to the least-squares estimates of degrees of freedom \((\nu = N - 2)\) (with \(N\) the total number of observations), \(\hat{\beta}\) and the standard error, \(s^2\).

2. Natural conjugate priors

In Bayesian analysis, the prior is meant to reflect any previously available data or information about the parameters of interest. The choice of prior used determines whether or not there is an analytical solution for the posterior distribution. In the case where no analytical solution is possible, numerical integration techniques, using Monte Carlo simulations, such as the Gibb’s sampler are needed. A particular class of priors, known as natural conjugate priors, are typically used because when they are combined with the likelihood function, these yield a posterior of the same class of distributions (Birkes & Dodge, 1993; Koop, 2003). This ensures that an analytical solution for the posterior is possible. For the Normal regression model, the natural conjugate prior can be stated as:

\[
p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2) \quad \text{Equation E-6}
\]
\[
p(\beta, \tau) = p(\beta|\tau)p(\tau) \quad \text{Equation E-7}
\]

where \(\tau = 1/\sigma^2\), and \((\beta|\tau) \sim N(\bar{\beta}, \tau^{-1}V)\) and \(\tau \sim \Gamma(s^{-2}, \nu)\), and the underlined terms represent the prior hyper-parameters. This means that the joint prior is Normal–Gamma: \((\beta, \tau) \sim N\Gamma(\bar{\beta}, V, s^{-2}, \nu)\). Multiplying this prior with the likelihood function results in the joint posterior of \(\beta\) and \(\tau\) that is also Normal–Gamma (as expected because of the choice of natural conjugate priors):

\[
(\beta, \tau|y) \sim N\Gamma(\bar{\beta}, \bar{V}, \bar{s}^{-2}, \bar{\nu}) \quad \text{Equation E-8}
\]

where the bar above parameter denotes the “posterior” value of the hyper-parameters. These hyper-parameters can then be estimated as follows:

\[
\bar{V} = (V^{-1} + X^T X)^{-1} \quad \text{Equation E-9}
\]
\[ \bar{\beta} = \bar{V}(V^{-1}\beta + X^TX\beta) \]  
Equation E-10

\[ \bar{v} = \nu + N \]  
Equation E-11

And \( \bar{s}^{-2} \) can be calculated from:

\[ \bar{v}\bar{s}^{-2} = \nu s^{-2} + s^{-2} (\bar{\beta} - \beta)^T [\bar{V} + (X^TX)^{-1}]^{-1} (\bar{\beta} - \beta) \]  
Equation E-12

In essence, this form of BLR results in model estimates (of \( \beta \) and \( \tau \)) that are weighted averages of the provided priors and the least–square estimates (West, 1984). This formulation is the standard BLR form with a natural conjugate prior.

3. Non–informative priors

It can be readily seen from the above formulation that the impact of the selection of priors can be extremely significant, and can severely influence the posterior values of the model parameters. Thus, often a single value for a prior is rarely used; instead an upper or lower limit is used to get an estimate of parameters. However, in many cases there is no prior information available, and thus this type of prior is unrealistic or subjective. Thus, an objective prior is needed that can represent complete uncertainty regarding the model parameters – this type of prior is known as a non–informative prior. To represent this in the formulation above, the values of \( \nu \) and \( V^{-1} \) are both set at 0. This results in posterior estimates of the hyper–parameters as follows:

\[ \bar{V} = (X^TX)^{-1} \]  
Equation E-13

\[ \bar{\beta} = \hat{\beta} \]  
Equation E-14

\[ \bar{v} = N \]  
Equation E-15

And \( \bar{s}^{-2} \) can be calculated from:

\[ \bar{v}\bar{s}^{-2} = \nu s^{-2} \]  
Equation E-16

Using this formulation the predictive distribution of \( y \) can be found analytically and is multivariate \( t \):

\[ (y^*|y) \sim t(X^*\bar{\beta}, \bar{s}^2[I + X^*\bar{V}X^*^T], \bar{v}) \]  
Equation E-17
where \(X^*\) are new observation used for prediction, and \(y^*\) are the predicted values.

4. Independent priors

Apart from the natural conjugate prior, and the non-informative prior described above, another option for the form of prior used in BLR is the independent prior. Under this prior the parameters \(\beta\) and \(\tau\) are independent, i.e. \(p(\beta, \tau) = p(\beta) p(\tau)\), which compared to the original relationship between the two parameters removes the conditional on \(\beta\) (n.b. the variance of \(\beta\) was listed as \(\tau^{-1}V\), again showing the dependence of \(\beta\) on \(\tau\)). By removing this conditional, the independent priors can be expresses as: \(\beta \sim N(\beta, V)\) and \(\tau \sim \Gamma(\alpha, \nu)\). Unlike the first case using the natural conjugate priors, multiplying these priors with the likelihood function results in an expression for \((\beta, \tau|y)\) that is no longer \(N\Gamma\), or in the form of a known density function. Thus, an analytical solution for the joint distribution for both posterior parameters and the predictive distribution for \(y^*\) is not possible. However, a numerical solution using a Monte Carlo integration method along with the conditional posterior distributions of the parameters is possible. The conditional posterior distribution for \(\beta\) and \(\tau\) can be calculated as:

\[
(\beta|\tau, y) \sim N(\bar{\beta}, \bar{V}) \quad \text{Equation E-18}
\]

\[
(\tau|\beta, y) \sim \Gamma(\bar{s}^{-2}, \bar{\nu}) \quad \text{Equation E-19}
\]

where the hyper-parameters can be calculated as follows:

\[
\bar{V} = (V^{-1} + \tau X^TX)^{-1} \quad \text{Equation E-20}
\]

\[
\bar{\beta} = \bar{V}(V^{-1} \beta + \tau X^Ty) \quad \text{Equation E-21}
\]

\[
\bar{\nu} = \nu + N \quad \text{Equation E-22}
\]

\[
\bar{s}^{-2} = \frac{(y - X\beta)^T(y - X\beta) + s^2}{\bar{\nu}} \quad \text{Equation E-23}
\]

A Gibb’s sampling algorithm can be used to sample from the posterior conditionals to find the joint posterior distribution \((\beta, \tau)\) numerically (adapted from Greenberg, 2008). First a random value of \(\tau\) is selected; this value is used to calculate one realisation of \(\bar{V}\) then \(\bar{\beta}\). A value of \(\beta\) is then sampled from \((\beta|\tau, y) \sim N(\bar{\beta}, \bar{V})\), and this value is used to sample \((\tau|\beta, y) \sim \Gamma(\bar{s}^{-2}, \bar{\nu})\). This process is repeated until convergence and typically a number of samples considered to be “burn–in” are discarded (selected as 10%, or 100
samples for this research). Each pair of sampled values of $\beta$ and $\tau$ represents an approximation of the joint posterior. To predictive probability of $y$ can be estimate numerically; for each pair of $\beta$ and $\tau$, a sample for $y^*$ can be drawn from the predictive distribution:

$$(y^*) \sim N(X^* \beta, 1/\tau)$$

Equation E-24

where $X^*$ are new observation used for prediction and $y^*$ are the predicted values. The posterior conditional distributions from the Gibb’s sampling routine can then be used as the priors for any subsequent models.
Appendix F
Summary of results for Chapter 5

Figure F-1 Results for M01 to M09 using BLR and FLR models for the 1 day lag at all resolutions.

Figure F-2 Results for M01 to M09 using BLR and FLR models for the 2 day lag at all resolutions.
Figure F-3 Results for M01 to M09 using BLR and FLR models for the 3 day lag at all resolutions

Figure F-4 Results for M01 to M09 using BLR and FLR models for the 7 day lag at all resolutions
1. Constructing fuzzy numbers

clear all
close all
clc

% call mu_possible to convert observations to fuzzy numbers for daily DO

for mm = 1:9
    M = ['M0' num2str(mm) '\'];
    for lag = [1, 2, 3, 7]
        Lag = ['Lag_0' num2str(lag) '\'];
        for resolution = [96, 24, 6]
            Res = ['Res_' num2str(resolution) '\'];

            path = 'C:\Users\Usman\Documents\MATLAB\DO_BLRvsFLR\Model_data\';
            file = [path M Lag Res 'Data_CALIB.mat'];
            load(file)

            % fuzzify: mean daily DO & DOL
            [m, ~] = size(DO); e = zeros(m,2); eL = zeros(m,2);
            for k = 1:m
                ea = 0.1 * nanmin(DO(k,:)); eb = 0.1 * nanmax(DO(k,:));
                eaL = 0.1 * nanmin(DOL(k,:)); ebL = 0.1 * nanmax(DOL(k,:));
                e(k,:) = [ea eb]; eL(k,:) = [eaL ebL];
            end
            clear vars k ea eb eaL ebL

            DOf = zeros(m,9); DOfL = zeros(m,9);

            for j = 1:m
                [~, DOf(j,:)] = muPossible(DO(j,:),e(j,:));
                [~, DOfL(j,:)] = muPossible(DOL(j,:),eL(j,:));
            end
            clear vars j e eL

            file = [path M Lag Res 'Data_CALIB_F.mat'];
            save(file,'Date','DOF','DOFL')
            clear vars DOF DOFL Date DO DOL file

            % validation
            file = [path M Lag Res 'Data_VALID.mat'];
            load(file)
\[ [m, \sim] = \text{size}(\text{DO}); \ e = \text{zeros}(m,2); \ eL = \text{zeros}(m,2); \]

\begin{verbatim}
for k = 1:m
  ea = 0.1 * \text{nanmin}(\text{DO}(k,:)); \ eb = 0.1 * \text{nanmax}(\text{DO}(k,:));
  eaL = 0.1 * \text{nanmin}(\text{DOL}(k,:)); \ ebL = 0.1 * \text{nanmax}(\text{DOL}(k,:));
  e(k,:) = [ea \ eb]; \ eL(k,:) = [eaL \ ebL];
end
clear vars k ea eb eaL ebL
\end{verbatim}

\[ \text{DOf} = \text{zeros}(m,9); \ \text{DOfL} = \text{zeros}(m,9); \]

\begin{verbatim}
for j = 1:m
  [~, \text{DOf}(j,:)] = \text{muPossible} \left( \text{DO}(j,:), e(j,:) \right);
  [~, \text{DOfL}(j,:)] = \text{muPossible} \left( \text{DOL}(j,:), eL(j,:) \right);
end
clear vars j e eL
\end{verbatim}

\[ \text{file} = \left[ \text{path} \ M \ \text{Lag} \ \text{Res} \ 'Data_VALID_F.mat' \right]; \]

\begin{verbatim}
save(file, 'Date', 'DOf', 'DOfL')
clear vars DOf DOfL Date DO DOL file
end
\end{verbatim}

\begin{verbatim}
function [mu, x] = \text{muPossible} (X, e)

% X is a 1 dimensional vector of observations
% X is converted into the correct format
% e is a vector (2,1) of upper and lower errors associated with X

[m, n] = \text{size}(X);
if m == 1
  X = X';
end
clear vars m n

% calc probabilities
delta = (\text{max}(X) - \text{min}(X))/\text{length}(X);

% if entire array is the same value, we have problems...so this is a
% workaround, technically speaking this means the value should be
% crisp,
% but this doesn't account for the error e
if delta == 0
  delta = (e(1) + e(2))/2;
end
x = \text{min}(X) - delta - e(1):delta:\text{max}(X) + delta + e(2);
x = x';
n = \text{histc}(X,x);
p = n/\text{sum}(n);
X1 = [n x];
clear vars e

% find corresponding x max p
nmax = \text{max}(X1(:,1));
\end{verbatim}
k = 1;
xmid = zeros(1,1);
for i = 1:length(n)
    if X1(i,1) == nmax
        xmid(k) = X1(i,2);
        k = k +1;
    end
end
xmid = nanmean(xmid);
clear vars k i X1 nmax

if isnan(xmid)
    x = [NaN NaN NaN NaN NaN NaN NaN NaN NaN];
    mu = [0 0.25 0.5 0.75 1 0.75 0.5 0.25 0];
else

    % calculate values and corresponding index of points midway between xmin % and xmid, and xmid and xmax
    i = find(x == xmid);
    if isempty(i)
        i = find(x<xmid);
        i = i(end)+1;
    end
    xl = (x(i) + min(X))/2;
    if xl<min(X)
        xl = min(X);
    end

    xr = (max(X) + x(i))/2;
    clear vars i
    i = find(x == xl);
    if isempty(i)
        i = find(x<xl);
        i = i(end);
    end
    j = find(x == xr);
    if isempty(j)
        j = find(x<xr);
        j = j(end);
    end

    % figure out xLs and xRs
    xl = x(1:i);
    pl = p(1:i);
    Pl = sum(pl);
    xr = x(j:end);
    pr = p(j:end);
    Pr = sum(pr);
clear vars i j pl pr

alpha1 = Pl + Pr;
clear vars Pl Pr nmax

% next level, midway between 0 and midway
n = length(xl);
n = round(n/2);
xl1 = xl(1:n);
pl = p(1:n);
Pl = sum(pl);
clear vars n

n = length(xr);
n = round(n/2);
xr1 = xr(end-(n-1):end);
pr = p(end-(n-1):end);
Pr = sum(pr);
clear vars n

alpha2 = Pl + Pr;
clear vars Pl Pr pl pr

% next level, midway between midway and max
a = (xmid + xl(end))/2;
xl2 = xl:delta:a;
pl = p(1:length(xl2));
Pl = sum(pl);
clear vars a pl

a = (xr(1) + xmid)/2;
xr2 = a:delta:xr(end);
n = length(xr2);
pr = p(end-n+1:end);
Pr = sum(pr);
clear vars a n pr

alpha3 = Pl + Pr;
clear vars Pl Pr

if alpha3 == 1
    alpha3 = alpha1;
end

% interpolation
L = [xl(1) xl1(end) xl(end) xl2(end) xmid];
R = [xmid xr2(1) xr(1) xr1(1) xr(end)];

[XL,XR] = muInterp(L, R, alpha1, alpha2, alpha3);
clear vars x1 xl1 xl2 xr xr1 xr2 xmid p alpha1 alpha2 alpha3 L R delta

mu = [0 0.25 0.5 0.75 1 0.75 0.5 0.25 0];
x = [XL XR(2:end)];
clear vars XL XR
end

end

2. Fuzzy linear regression
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clear all
close all
clc
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% file path to save results at the end:
path = 'C:\Users\Usman\Documents\MATLAB\DO_BLRvsFLR\Model_data\';

% conduct fuzzy linear regression
for mm = 1:9
    for lag = [1, 2, 3, 7]
        for resolution = [96, 24, 6]

            % convert model/run number to string to access file via
            path:
            M = ['M0' num2str(mm) '\'];
            Lag = ['Lag_0' num2str(lag) '\'];
            Res = ['Res_' num2str(resolution) '\'];

            %%% calibration %%%
            file = [path M Lag Res 'Data_CALIB_F.mat'];
            load(file)

            %%% remove nans
            AA = [DOf DOfL Date]; AB = AA(~any(isnan(AA),2),:);

            DOf = AB(:,1:9);
            DOfL = AB(:,10:18);
            Date = AB(:,19);
            clear vars AA AB

            %%% do flr and reorg coefficients
            [B0, B1] = usmanl(DOfL, DOf);

            %%%% predictions: calibration %%%%
            DOfh = zeros(length(DOf),9);
            errorC = zeros(6,9);

            for i = 1:9
                DOfh(:,i) = B0(i) + B1(i)*DOfL(:,i);
            end

            for i = 1:length(DOfh)
                DOfh(i,:) = reorgminmax(DOfh(i,:));
            end

```
%%% errors %%%
for i = 1:9
    errorC(:,i) = ALLERROR(DOf(:,i), DOfh(:,i));
end
clear vars i

RSR = zeros(9,1); NSE = zeros(9,1); PBIAS = zeros(9,1);
DateC = Date;
file = [path M Lag Res 'Data_CALIB.mat'];
load(file)

K = ismember(Date, DateC);
DO = DO(K,:);
DOm = nanmean(DO,2);
clear vars K DOL Date DO

for i = 1:9
    NSE_in = errorC(2,i);
    RMSE = errorC(3,i);
    Yobs = DOf(:,i);
    Yhat = DOfh(:,i);
    [RSR(i,1), NSE(i,1), PBIAS(i,1)] = getstats(RMSE, DOm, NSE_in, Yobs, Yhat);
end
clear vars i NSE_in RMSE Yobs Yhat DOm

file = [path M Lag Res 'Results_CALIB_F.mat'];
save(file,'B0', 'B1', 'DateC', 'DOf', 'DOfL', 'DOfh', 'errorC', 'NSE', 'RSR', 'PBIAS')
clear vars DateC DOf DOfL DOfh
clear vars errorC NSE RSR PBIAS
clear vars file

%%% validation %%%
file = [path M Lag Res 'Data_VALID_F.mat'];
load(file)

%%% remove nans
AA = [DOf DOfL Date]; AB = AA(~any(isnan(AA),2),:);

DOf = AB(:,1:9);
DOfL = AB(:,10:18);
Date = AB(:,19);
clear vars AA AB

%%% predictions: validation %%%
DOfh = zeros(length(DOf),9);
errorV = zeros(6,9);

for i = 1:9
    DOfh(:,i) = B0(i) + B1(i)*DOfL(:,i);
end

for i = 1:length(DOfh)
    DOfh(i,:) = reorgminmax(DOfh(i,:));
end
for i = 1:9
    errorV(:,i) = ALLERROR(DOf(:,i), DOfh(:,i));
end

RSR = zeros(9,1); NSE = zeros(9,1); PBIAS = zeros(9,1);
DateV = Date;
file = [path M Lag Res 'Data_VALID.mat'];
load(file)

K = ismember(Date,DateV);
DO = DO(K,:);
DOm = nanmean(DO,2);
clear vars K DOL Date DO

for i = 1:9
    NSE_in = errorV(2,i);
    RMSE = errorV(3,i);
    Yobs = DOf(:,i);
    Yhat = DOfh(:,i);
    [RSR(i,1), NSE(i,1), PBIAS(i,1)] = getstats(RMSE, DOm, NSE_in, Yobs, Yhat);
end
clear vars i NSE_in RMSE Yobs Yhat DOm

file = [path M Lag Res 'Results_VALID_F.mat'];
save(file, 'B0', 'B1', 'DateV', 'DOf', 'DOfL', 'DOfh', 'errorV', 'NSE', 'RSR', 'PBIAS')
clear vars DateV DOf DOfL DOfh
clear vars errorV NSE RSR PBIAS
clear vars file B0 B1
end

function [A0, A1] = usman1(X1f, Yf)
% Input are fuzzy numbers X1f, and Yf, each N by 9 elements in size,
% corresponding to the membership function:
% mu = [0 0.25 0.5 0.75 1 0.75 0.5 0.25 0]

% Output are the three parameters resulting from least-squares linear
% regression

% sort input matrices into vectors
X1_aL = X1f(:,1);
X1_bL = X1f(:,2);
X1_cL = X1f(:,3);
X1_dL = X1f(:,4);
X1_1  = X1f(:,5);
X1_dR = X1f(:,6);
X1_cR = X1f(:,7);
X1_bR = X1f(:,8);
X1_aR = X1f(:,9);

clear vars X1f

Y_aL = Yf(:,1);
Y_bL = Yf(:,2);
Y_cL = Yf(:,3);
Y_dL = Yf(:,4);
Y_1  = Yf(:,5);
Y_dR = Yf(:,6);
Y_cR = Yf(:,7);
Y_bR = Yf(:,8);
Y_aR = Yf(:,9);

clear vars Yf

% least squares

N = length(X1_1);

X_aL = [ones(N,1) X1_aL];
X_bL = [ones(N,1) X1_bL];
X_cL = [ones(N,1) X1_cL];
X_dL = [ones(N,1) X1_dL];
X_1  = [ones(N,1) X1_1];
X_dR = [ones(N,1) X1_dR];
X_cR = [ones(N,1) X1_cR];
X_bR = [ones(N,1) X1_bR];
X_aR = [ones(N,1) X1_aR];

XX = zeros(9*N, 9*2);

XX(1:N,1:2) = X_aL;
XX(N+1:2*N,3:4) = X_bL;
XX(2*N+1:3*N,5:6) = X_cL;
XX(3*N+1:4*N,7:8) = X_dL;
XX(4*N+1:5*N,9:10) = X_1;
XX(5*N+1:6*N,11:12) = X_dR;
XX(6*N+1:7*N,13:14) = X_cR;
XX(7*N+1:8*N,15:16) = X_bR;
XX(8*N+1:9*N,17:18) = X_aR;

YY = [Y_aL; Y_bL; Y_cL; Y_dL; Y_1; Y_dR; Y_cR; Y_bR; Y_aR];

BETA = (XX'*XX) \ XX'*YY;

A0 = BETA(1:2:end);
A1 = BETA(2:2:end);

End

function Anew = reorgminmax(A)
% this program reorganized regression coefficients to obey fuzzy rules
% using min/max algorithms

[m, ~] = size(A);

% rearrange input matrix for computation in this function
% need to use column vector rather than row vector
if m == 1
    A = A';
end

clear vars m

% modal value stays the same
Anew(5,1) = A(5,1);

% at membership level mu = 0.75, we make the minimum of three values at
% A(4) A(5) and A(6) to be the LEFT limit and the maximum of the three
% values to be the RIGHT limit

Anew(4,1) = min([A(4,1) A(6,1) Anew(5,1)]);
Anew(6,1) = max([A(4,1) A(6,1) Anew(5,1)]);

% the above processess is continued until mu = 0. The resulting vector is
% the fuzzy number of the coefficient of fuzzy regression, ensuring that it
% is normal and convex as required

Anew(3,1) = min([A(3,1) A(7,1) Anew(4,1)]);
Anew(7,1) = max([A(3,1) A(7,1) Anew(6,1)]);
Anew(2,1) = min([A(2,1) A(8,1) Anew(3,1)]);
Anew(8,1) = max([A(2,1) A(8,1) Anew(7,1)]);
Anew(1,1) = min([A(1,1) A(9,1) Anew(2,1)]);
Anew(9,1) = max([A(1,1) A(9,1) Anew(8,1)]);

end

3. Bayesian linear regression – non-informative prior

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clear all
close all
clc
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% file path to save results at the end:
path = 'C:\Users\Usman\Documents\MATLAB\DO_BLRvsFLR\Model_data\';

mm = 1; M = ['M0' num2str(mm) ' 
'];
for lag = [1, 2, 3, 7]
    for resolution = [96, 24, 6]
        % convert model/run number to string to access file via path:
        Lag = ['Lag_0' num2str(lag) ' 
'];
        Res = ['Res_' num2str(resolution) ' 
'];
%%% calibration %%%

file = [path M Lag Res 'Data_CALIB.mat'];
load(file)

% calculate daily means
DO = nanmean(DO,2); DOL = nanmean(DOL,2);

%%% remove nans
AA = [DO DOL Date];
AB = AA(~any(isnan(AA),2),:);

DO = AB(:,1);
DOL = AB(:,2);
Date = AB(:,3);
clear vars AA AB

X = [ones(length(DOL),1) DOL];
% beta0 is the mean of BETA (see below); this is the standard non-Bayesian
% estimate of beta
[beta0, beta0int] = regress(DO, X);

V_beta0 = inv(X'*X);

% B0 is the variance of beta
B0(1,1) = (sqrt(length(DOL)) * (max(beta0int(1,:)) -
  min(beta0int(1,:)))/3.92)^2;
B0(2,1) = (sqrt(length(DOL)) * (max(beta0int(2,:)) -
  min(beta0int(2,:)))/3.92)^2;

clear vars beta0int

% Bayesian analysis
% Bayesian estimate of sigma2 sampled from below:
% note: this is an inverse-scaled-chi2 dist with "df" as shape
and "sgamm2" as scale
n_sim = 1000;
burn_in = 0.1*n_sim;

tau = gamrnd(df/2,1/(df/2*s2),n_sim,1);
sigma2 = 1./tau;
mean_sigma2 = mean(sigma2);

% Bayesian estimate of beta can be sampled from below:
% these are primary for reference to show evolution of coefficients
% not used for posterior predictive distribution simulation
% (beta0 is used)
% the mean of beta should be the same as beta0

beta1 = zeros(length(sigma2),1); beta2 = zeros(length(sigma2),1);
for ii = 1:length(sigma2)
    beta1(ii,:) = normrnd(beta0(1,1),V_beta0(1,1)*sqrt(sigma2(ii,1)));
    beta2(ii,:) = normrnd(beta0(2,1),V_beta0(2,2)*sqrt(sigma2(ii,1)));
end
clear vars ii

beta = [beta1 beta2]';
clear vars beta1 beta2

%%%%%%%%%%%%%%%%%%%% predictions: calibration
%%%%%%%%%%%%%%%%%%%%%
DOhALL = zeros(n,n_sim);
for j = 1:n
    A = X(j,:)*beta0;
    B = sqrt(mean_sigma2*(1+(X(j,:)*V_beta0*X(j,:)'')));
    C = df;
    % predicted y using random sampling
    DOhALL(j,:) = random('tlocationscale',A,B,C,[1,n_sim]);
end
clear vars j A B C df n k X

% remove burn_in
    beta(:,1:burn_in)=[];
    tau(1:burn_in,:)=[];
    sigma2(1:burn_in,:)=[];
    DOhALL(:,1:burn_in)=[];

%%%%% erros %%%%%
errorC = ALLERROR(DO, mean(DOhALL,2));
% errorC = ALLERROR(DO, max(DOhALL,[],2))
DateC = Date;
DOC = DO;
file = [path M Lag Res 'Data_CALIB.mat'];
load(file,'DO','Date')
clear vars file

K = ismember(Date,DateC);
DO = DO(K,:);
DOM = nanmean(DO,2);
clear vars K Date DO

DO = DOC; clear vars DOC

NSE_in = errorC(2,1);
RMSE = errorC(3,1);
Yobs = DO;
Yhat = mean(DOhALL,2);
[RSR, NSE, PBIAS] = getstats(RMSE, D0m, NSE_in, Yobs, Yhat);

clear vars NSE in RMSE Yobs Yhat D0m
% % save variables to be used as priors for Model 2:
a_prior = beta0;
bPrior = [B0(1,1) 0; 0 B0(2,1)];
c_prior = mean(tau)*mean(tau)/var(tau);
d_prior = var(tau)/mean(tau);

% save priors for Model 2
file = [path M Lag Res 'Priors_B_1.mat'];
save(file,'a_prior','b_prior','c_prior','d_prior')
clear vars a_prior b_prior c_prior d_prior B0 beta file

% save calib results
file = [path M Lag Res 'Results_CALIB_B.mat'];
save(file,'beta0','s2', 'tau',
'sigma2','mean_sigma2','V_beta0','DateC','DO','DOL','DOhALL','errorC','
NSE','RSR','PBIAS')
clear vars DateC DO DOL DOhALL errorC NSE RSR PBIAS file

% % validation %%%
file = [path M Lag Res 'Data_VALID.mat'];
load(file)

% calculate daily means
DO = nanmean(DO,2); DOL = nanmean(DOL,2);

% % remove nans
AA = [DO DOL Date];
AB = AA(~any(isnan(AA),2),:);

DO = AB(:,1);
DOL = AB(:,2);
Date = AB(:,3);
clear vars AA AB

X = [ones(length(DOL),1) DOL];

% % predictions: validation
[n, k] = size(X);
df = n - k;

DOhALL = zeros(n,n_sim - burn_in);
for j = 1:n
A = X(j,:)*beta0;
B = sqrt(mean_sigma2*(1+(X(j,:)*V_beta0*X(j,:)'')));
C = df;
% predicted y using random sampling
DOhALL(j,:) = random('tlocationscale',A,B,C,[1,n_sim -
burn_in]));
end
clear vars j A B C df n k X
%%%%% erros %%%%%
errorV = ALLERROR(DO, mean(DOhALL,2));
% errorV = ALLERROR(DO, max(DOhALL,[],2))
DateV = Date;
DOV = DO;
file = [path M Lag Res 'Data_VALID.mat'];
load(file,'DO','Date')
clear vars file
K = ismember(Date,DateV);
DO = DO(K,:);
DOm = nanmean(DO,2);
clear vars K Date DO
DO = DOV; clear vars DOV
NSE_in = errorV(2,1);
RMSE = errorV(3,1);
Yobs = DO;
Yhat = mean(DOhALL,2);
[RSR, NSE, PBIAS] = getstats(RMSE, DOm, NSE_in, Yobs, Yhat);
clear vars NSE_in RMSE Yobs Yhat DOm

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% save valid results
file = [path M Lag Res 'Results_VALID_B.mat'];
save(file,'beta0','s2', 'tau',
'sigma2','mean_sigma2','V_beta0','DateV','DO','DOL','DOhALL','errorV','NSE','RSR','PBIAS')
clear vars DateV DO DOL DOhALL errorV NSE RSR PBIAS file

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clear vars beta0 V_beta0 sigma2 tau s2 mean_sigma2
end
end
clear vars lag mm resolution Lag M Res path n_sim burn_in

4. Bayesian linear regression – conjugate prior
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
set(0,'DefaultFigureWindowStyle','docked')
addpath ('C:\Users\Usman\Documents\MATLAB\Function Library')
addpath ('C:\Users\Usman\Documents\MATLAB\DO_BLRvsFLR')
addpath ('C:\Users\Usman\Documents\MATLAB\WQ_Data_load')
clear all
close all
clc

% file path to save results at the end:
path = 'C:\Users\Usman\Documents\MATLAB\DO_BLRvsFLR\Model_data\';

for mm = 2:9;
    for lag = [1, 2, 3, 7]
        for resolution = [96, 24, 6]
            % convert model/run number to string to access file via path:
            M = ['M0' num2str(mm) '\'];
            Lag = ['Lag_0' num2str(lag) '\'];
            Res = ['Res_' num2str(resolution) '\'];

            %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
            %%%% calibration %%%%
            file = [path M Lag Res 'Data_CALIB.mat'];
            load(file)
            clear vars file

            % get prior:
            MM = ['M0' num2str(mm-1) '\'];
            prior = ['Priors_B_' num2str(mm-1) '.mat'];
            file = [path MM Lag Res prior];
            load(file)
            clear vars file MM prior

            % calculate daily means
            DO = nanmean(DO,2); DOL = nanmean(DOL,2);

            %%% remove nans
            AA = [DO DOL Date];
            AB = AA(~any(isnan(AA),2),:);

            DO = AB(:,1);
            DOL = AB(:,2);
            Date = AB(:,3);
            clear vars AA AB

            X = [ones(length(DOL),1) DOL];
            y = DO;
            %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% estimate posteriors
            n = length(y); % length of dataset
            n_sim = 1000; % number of simulations
            burn_in = 0.1*n_sim; % burn in is 10%

            tau = zeros(n_sim, 1); % initialize vectors for Gibb's sampling
            beta = zeros(size(X,2), n_sim); % initialize vectors for Gibb's sampling

            tau(1,1) = gamrnd(c_prior, d_prior); % guess first value of tau
            D_prior = 1/d_prior;

            for i = 2:n_sim
                b_post = inv((tau(i-1,1)*X'*X)) + inv(b_prior));
a_post = b_post*(tau(i-1,1)*(X'*y) +
inv(b_prior)*a_prior);
beta(:,i) = normrnd(a_post, sqrt(diag(b_post)));
c_post = c_prior + n/2;
D_post = D_prior + ((y - X*beta(:,i))'*(y -
X*beta(:,i)))/2;
d_post = 1/D_post;
tau(i,1) = gamrnd(c_post, d_post);
end
clear vars i
clear vars b_post a_post c_post d_post y D_post D_prior

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% simulate posterior predictive: Calibration
beta;
sigma2 = 1./tau';
DOhALL = zeros(n, n_sim);

for i = 1:n_sim
  % right now, generating one realisation for each beta, sigma2, X; but
  % possible to get an array of realisations, averaged at each beta etc.
  temp = normrnd(X*beta(:,i),sqrt(sigma2(:,i)));
  DOhALL(:,i) = temp;
end
clear vars i n X temp

% remove burn in
beta(:,1:burn_in)=[];
tau(1:burn_in,:)=[];
sigma2(:,1:burn_in)=[];
DOhALL(:,1:burn_in)=[];

% B01_BLR_Figs % potential to run figures here

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% erros %%%%
errorC = ALLERROR(DO, mean(DOhALL,2));
% errorC = ALLERROR(DO, max(DOhALL,[],2))
DateC = Date;
DOC = DO;
file = [path M Lag Res 'Data_CALIB.mat'];
load(file, 'DO', 'Date')
clear vars file

K = ismember(Date,DateC);
DO = DO(K,:);
DOM = nanmean(DO,2);
clear vars K Date DO

DO = DOC; clear vars DOC

NSE_in = errorC(2,1);
RMSE = errorC(3,1);
Yobs = DO;
Yhat = mean(DOhALL,2);
[RSR, NSE, PBIAS] = getstats(RMSE, DOm, NSE_in, Yobs, Yhat);

clear vars NSE_in RMSE Yobs Yhat DOm

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% % save variables to be used as priors for Model mm+1:
a_prior = mean(beta,2);
b_prior = [var(beta(1,:)) 0; 0 var(beta(2,:))];
c_prior = mean(tau)*mean(tau)/var(tau);
d_prior = var(tau)/mean(tau);

% save priors for Model mm+1
file = [path M Lag Res 'Priors_B_' num2str(mm) '.mat'];
save(file,'a_prior','b_prior','c_prior','d_prior')
clear vars a_prior b_prior c_prior d_prior file

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% save calib results
file = [path M Lag Res 'Results_CALIB_B.mat'];
save(file,'beta','tau','sigma2','DateC','DO','DOL','DOhALL','errorC','NSE','RSR','PBIAS')
clear vars DateC DO DOL DOhALL errorC NSE RSR PBIAS file

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% validation
file = [path M Lag Res 'Data_VALID.mat'];
load(file)
clear vars file

% calculate daily means
DO = nanmean(DO,2); DOL = nanmean(DOL,2);

%%% remove nans
AA = [DO DOL Date];
AB = AA(~any(isnan(AA),2),:);

DO = AB(:,1);
DOL = AB(:,2);
Date = AB(:,3);
clear vars AA AB

X = [ones(length(DOL),1) DOL];

%%%%%%%%%%%%%%%% simulate posterior predictive: Validation
beta;
sigma2 = 1./tau';
DOhALL = zeros(length(X), n_sim - burn_in);

for i = 1:n_sim - burn_in
  % right now, generating one realisation for each beta, sigma2, X; but
% possible to get an array of realisations, averaged at
each beta etc.

temp = normrnd(X*beta(:,i),sqrt(sigma2(:,i)));
DOhALL(:,i) = temp;
end
clear vars i n X temp

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%% errros %%%%%
errorV = ALLERROR(DO, mean(DOhALL,2));
% errorV = ALLERROR(DO, max(DOhALL,[],2))
DateV = Date;
DOV = DO;
file = [path M Lag Res 'Data_VALID.mat'];
load(file,'DO','Date')
clear vars file

K = ismember(Date,DateV);
DO = DO(K,:);
DOm = nanmean(DO,2);
clear vars K Date DO

DO = DOV; clear vars DOV

NSE_in = errorV(2,1);
RMSE = errorV(3,1);
Yobs = DO;
Yhat = mean(DOhALL,2);
[RSR, NSE, PBIAS] = getstats(RMSE, DOm, NSE_in, Yobs, Yhat);
clear vars NSE_in RMSE Yobs Yhat DOm

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% save valid results
file = [path M Lag Res 'Results_VALID_B.mat'];
save(file,'beta','tau','sigma2','DateV','DO','DOL','DOhALL','errorV','NSE','RSR','PBIAS')
clear vars DateV DO DOL DOhALL errorV NSE RSR PBIAS file

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clear vars beta sigma2 tau
end
end
clear vars lag mm resolution Lag M Res path n_sim burn_in

5. Possibility-Probability transformation

function Px = inversetransform(Xobs, XFuzzy)

%%% this function calculates the probability of exceedance of a fuzzy
%%% number using the inverse tranformation used by Zhang 2009 pg 73.
%%% Given a value of Xobs that is within the mu = 0 alpha cut interal
%%% of the corresponding fuzzy number, interpolate to get the
membership
level corresponding to Xobs. Then interpolate to get the mirror (Xprime) of Xobs, i.e. other values in the fuzzy number with equal membership level.

Then the membership level of these two values is sum of the probability of exceedance: P(X>Xob) + P(X<Xprime)

Then use the ratio of the length between Xobs and X@mu=0R, and X@mu=0L to get area/probability corresponding to P(X>Xob)

Xobs is observed value 1 x1
Xfuzzy is fuzzy value, 1 x 9

Check if observed value is larger or smaller than modal value of fuzzy number

if Xobs >= XFuzzy(1,9)
    Px = 0;
elseif Xobs < XFuzzy(1,1)
    Px = 100;
elseif Xobs >= XFuzzy(1,5)
    XR = XFuzzy(1,5:9);
    muR = 1:-0.25:0;
    get membership level of Xobs
    mu_obs = interp1(XR, muR, Xobs);
    using membership level of Xobs, get value of mirror X (Xprime)
    muL = 0:0.25:1;
    XL = XFuzzy(1,1:5);
    Xprime = interp1(muL, XL, mu_obs);

    total area undercurve between (-inf, Xprime] + [Xobs, inf) =
    or realistically: (Xfuzzy(1,1), Xprime] + [Xobs, Xfuzzy(1,9)) =

    % ratio of areas Al and AR:
    AL = AR*(Xprime-Xfuzzy(1,1))/(Xfuzzy(1,9)-Xobs))
    AL = AR*dXL/dXR
    We know that AL + AR = A (A == mu_obs), then:
    AR = mu_obs/(1 + dXL/dXR)
    dXR = XFuzzy(1,9) - Xobs;
    dXL = Xprime - XFuzzy(1,1);
    AR = mu_obs/(1 + dXL/dXR)*100;
    Px = AR;
else
    XL = XFuzzy(1,1:5);
    muL = 0:0.25:1;
    get membership level of Xobs
    mu_obs = interp1(XL, muL, Xobs);
    using membership level of Xobs, get value of mirror X (Xprime)
    muR = 1:-0.25:0;
    XR = XFuzzy(1,5:9);
    Xprime = interp1(muR, XR, mu_obs);

    dXR = XFuzzy(1,9) - Xprime;
    dXL = Xobs - XFuzzy(1,1);
    AL = mu_obs/(1 + dXR/dXL)*100;
\[ P_x = 100 - A_L; \]
end
end
Appendix H
Input variable selection for Chapter 6


Figure H-1 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 0 day lag

Figure H-2 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 1 day lag

Figure H-3 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 2 day lag
Figure H-4 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 3 day lag

Figure H-5 Correlation coefficients for peak flow versus mean daily flow and precipitation for the Bow River at Calgary for (from top to bottom): 7 day lag
Appendix I
Model performance results for Chapter 6

Figure I-1 RSR, NSE and PBIAS values for the validation dataset for $e = 10\%$ proof-of-concept models: the markers (circle, square, rhombus, and triangle) represent result for different lags (1, 2, 3 and 7 days, respectively): Banff (top), Calgary (middle) and Bragg Creek (bottom).

Figure I-2 RSR, NSE and PBIAS values for the validation dataset for $e = 5\%$ proof-of-concept models: the markers (circle, square, rhombus, and triangle) represent result for different lags (1, 2, 3 and 7 days, respectively): Banff (top), Calgary (middle) and Bragg Creek (bottom).
Figure I-3 RSR, NSE and PBIAS values for the validation dataset for all ten recursive models: the markers (circle, square, rhombus, and triangle) represent result for different lags (1, 2, 3 and 7 days, respectively). Error values of $e = 10\%$ (left), 10\% (middle), and 20\% (right) were used, for each of the three sites: Banff (top), Calgary (middle) and Bragg Creek (bottom).

Figure I-4 RSR, NSE and PBIAS values for the validation dataset for all ten recursive models: the markers (circle, square, rhombus, and triangle) represent result for different lags (1, 2, 3 and 7 days, respectively). Error values of $e = 10\%$ (left), 10\% (middle), and 20\% (right) were used, for each of the three sites: Banff (top), Calgary (middle) and Bragg Creek (bottom).
Appendix J
MATLAB code for Chapter 6

1. Error-in-variables regression

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
set(0,'DefaultFigureWindowStyle','docked')
addpath ('C:\Users\Usman\Documents\MATLAB\Function Library')
addpath (genpath('C:\Users\Usman\Documents\MATLAB\Flood_Risk'))
addpath (genpath('C:\Users\Usman\Documents\MATLAB\Flow_data_load'))

clear all
close all
clc

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% Regression type 1: 11 years data, 6 for calibration 5 for validation

%%% Load calibration data set
CALIB = 11:-2:1;
VALID = 10:-2:1;

for LAG = [1, 2, 3, 7]
    file0 = 'C:\Users\Usman\Documents\MATLAB\Flood_Risk\Regression\Lag_Q\Non_fuzzy' ;
    file2 = '_Qp_Qd_Lag_' ;

    for J = 1:3;
        close all
        if J == 1
            file1 = '05BB001';
            file3 = [file0 file1 file2 num2str(LAG) '.mat'];
            load(file3)
            disp('Hola J = 1')
            Location = 'BOW RIVER AT BANFF';
        else if J == 2
            file1 = '05BH004'; % used for filename save at end
            file3 = [file0 file1 file2 num2str(LAG) '.mat'];
            load(file3)
            disp('Hola J = 2')
            Location = 'BOW RIVER AT CALGARY';
        else if J == 3
            file1 = '05B004'; % used for filename save at end
            file3 = [file0 file1 file2 num2str(LAG) '.mat'];
            load(file3)
            disp('Hola J = 3')
            Location = 'ELBOW RIVER AT BRAGG CREEK ';
        end
    end
end
% calibration
Xs = []; Ys = []; u = [];

for i = 1:length(CALIB)
    Xs = [Xs; Qp_Qd_Lag{CALIB(1,i),1}];
    Ys = [Ys; Qp_Qd_Lag{CALIB(1,i),2}];
    u = [u; u_Lag{CALIB(1,i),1}];
end
% do regression
[B0, B1, ~, ~] = regress_EIV(Y, X, var_u);

% prediction
Yh = [ones(length(X),1) X]*[B0;B1];

% validation
Xvs = []; Yvs = []; u = [];

for i = 1:length(VALID)
    Xvs = [Xvs; Qp_Qd_Lag{VALID(1,i),1}];
    Yvs = [Yvs; Qp_Qd_Lag{VALID(1,i),2}];
    u = [u; u_Lag{VALID(1,i),1}];
end
Xv = Xvs(:,2); Yv = Yvs(:,2); Datev = Yvs(:,1);

% prediction
Yvh = [ones(length(Xv),1) Xv]*[B0;B1];
rv = Yv - Yvh; % recalc residual
[~, ~] = testregress(rv, [ones(length(Xv),1) Xv], Xv);

 calculate CI of reg line and predicted Yvs
[YvhL, YvhU, YvhLi, YvhUi] = confint(Xv, Yvh, s_r, meanX, Sxx);

 % error stats
% Rsq; Rsq NS; rmse; mae; aic; bic
ErrorV = ALLERROR(Yv, Yvh);

 %%% Plot all Figures %%%%%%%%
Regression_method_03_LAG_FIGS

 %%%%%%%%%%%%%%%%%%%%%%%%%%%%% Save files %%%%%%%%%%%%%%%%%%%%%%%%%%%
coefficients{1,1} = [B0; B1];
clear vars B0 B1

calib_data{1,1} = Xs; calib_data{2,1} = Ys;
calib_data{3,1} = Yh; calib_data{4,1} = r;
calib_data{5,1} = YhLi; calib_data{6,1} = YhUi;
calib_data{7,1} = YhL; calib_data{8,1} = YhU;
clear vars Xs Ys Yh YhL YhU YhLi YhUi

valid_data{1,1} = Xvs; valid_data{2,1} = Yvs;
valid_data{3,1} = Yvh; valid_data{4,1} = rv;
valid_data{5,1} = YvhLi; valid_data{6,1} = YvhUi;
valid_data{7,1} = YvhL; valid_data{8,1} = YvhU;
clear vars Xvs Yvs Yvh YvhL YvhU YvhLi YvhUi

errors{1,1} = ErrorC; errors{2,1} = ErrorV;
clear vars ErrorC ErrorV

 save_path0 =
 ['C:\Users\Usman\Documents\MATLAB\Flood_Risk\Regression\Error_in_Var\METHOD_03_EIV_SPLIT\'...
 'LAG ' num2str(LAG) ']
 mkdir(save_path0)
 filename = [save_path0 file1 '_RESULTS' '.mat'];

toload(filename, 'coefficients', 'calib_data', 'valid_data', 'errors')
clear vars filename

clear vars coefficients calib_data valid_data errors

 filename = [save_path0 file1 '_Lag_' num2str(LAG) '.fig'];
savefig(Fig, filename)
end

clear vars file0 file2 file3
end
Appendix K
MATLAB code for Chapter 7

1. Fuzzy neural network with crisp inputs

```matlab
% Load input and output data for calibration
load('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy\Data\Q_FINAL_RED.mat')
input1C = Q_FINAL_RED(:,4); % [1,3,5,7,9]
load('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy\Data\WQ_FINAL.mat')
input2C = WT_FINAL(:,4);
Output = DO_FINAL(:,4);
Input = [input1C, input2C];
clear vars DO_FINAL Q_FINAL_RED WT_FINAL

% convert cells to matrices
Input = cell2mat(Input); Output = cell2mat(Output);
clear vars input1C input2C

% Choose a Training Function
% For a list of all training functions type: help nntrain
% 'trainlm' is usually fastest.
trainFcn = 'trainlm'; % Levenberg-Marquardt backpropagation.

% Create a Fitting Network
hiddenLayerSize = 10;
global Nhidden
Nhidden = hiddenLayerSize;
% fitnet uses sigmoid, then linear transfer function; check using:
% net.layers{1} or net.layers{2}
global net
net = fitnet(hiddenLayerSize,trainFcn);
clear vars hiddenLayerSize trainFcn
```

% Choose Input and Output Pre/Post-Processing Functions
% For a list of all processing functions type: help nnprocess
net.input.processFcns = {'removeconstantrows','mapminmax'};
net.output.processFcns = {'removeconstantrows','mapminmax'};

% Setup Division of Data for Training, Validation, Testing
% For a list of all data division functions type: help nndivide
net.divideFcn = 'dividerand';  % Divide data randomly
net.divideMode = 'sample';  % Divide up every sample
net.divideParam.trainRatio = 50/100;
net.divideParam.valRatio = 25/100;
net.divideParam.testRatio = 25/100;

% Choose a Performance Function
% For a list of all performance functions type: help nnperformance
net.performFcn = 'mse';  % Mean Squared Error

% Choose Plot Functions
% For a list of all plot functions type: help nnplot
net.plotFcns = {'plotperform','plottrainstate','ploterrhist', ...
'plotregression','plotfit'};

% Train the Network
[net,tr] = train(net,x,t);

% Test the Network
y = net(x);
e = gsubtract(t,y);
performance = perform(net,t,y);
clear vars e performance

% Recalculate Training, Validation and Test Performance
trainTargets = t .* tr.trainMask{1};
valTargets = t .* tr.valMask{1};
testTargets = t .* tr.testMask{1};

trainPerformance = perform(net,trainTargets,y);
valPerformance = perform(net,valTargets,y);
testPerformance = perform(net,testTargets,y);

clear vars trainTargets valTargets testTargets

clear vars trainPerformance valPerformance testPerformance

%%% set(0,'DefaultFigureWindowStyle','docked')
addpath ('C:\Users\Usman\Documents\MATLAB\Function Library')
addpath (genpath('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy'))

% get weights and biases for optimization
global Nhidden % relisting here to call global value
global inp outp % used for fmincon and SCEUA
  % inp and outp only use training subset...
inp = x(:,tr.trainInd); outp = t(:,tr.trainInd);

global NInp Noutp nWIH nBH nWHO nBO nWB
% NInp and Noutp are the number of input variables (2) and output variables
NInp=size(inp,1); Noutp=size(outp,1);
% nWIH = number of weights between INPUT & HIDDEN layer
nWIH=NInp*Nhidden;
% nBH = number of biases between INPUT & HIDDEN layer
nBH=Nhidden;
% nWHO = number of weights between HIDDEN & Output layer
nWHO=Nhidden*Noutp;
% nBO = number of biases between HIDDEN & Output layer
nBO=Noutp;
% total Number of Weights and Biases
nWB= nWIH + nBH + nWHO + nBO;

% get values of weights and biases from each layer
PesiINP1=net.IW{1,1}; Pesi12=net.LW{2,1};
Bias1=net.b{1}; Bias2=net.b{2};
ntotPB=2*Nhidden;
PesiBias = [reshape(PesiINP1,1,ntotPB), Bias1', Pesi12, Bias2];
clear vars PesiINP1 Pesi12 Bias1 Bias2
clear vars ntotPB
global percInclusi % for alpha cuts
% for alpha cut = 0, 100% of data within final interval
percInclusi = 99.5;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% define delta to start upper and lower limits of pesibias
delta = 0.5;

% define best guess (xi) and upper and lower limits (bu, bl) of
% pesibias
xi(1,1:2:nWB*2)=PesiBias'+delta/2;
xi(1,2:2:nWB*2)=PesiBias'--delta/2;

bu(1,1:2:nWB*2)=PesiBias'+delta;
bu(1,2:2:nWB*2)=PesiBias'--10^-6;

bl(1,1:2:nWB*2)=PesiBias'+10^-6;
bl(1,2:2:nWB*2)=PesiBias'--delta;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% define properties for sce ua algorithm:
ngs=5;
maxn=20000;
kstop=50;
pcento=0.1;
peps=0.001;
iseed=-1;
iniflg=1;

[bestx,bestf] =
sceuauST('FO',xi,bl,bu,maxn,kstop,pcento,peps,ngs,iseed,iniflg);
clear vars ngs maxn kstop pcento peps iseed iniflg
disp(bestx);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% define properties for fmincon refinement
options = optimoptions('fmincon','Display','iter-detailed','MaxIter',20000,...
'MaxFunEvals',20000,' TolX',1e-20,' TolFun',1e-20,' TolCon',1e-20);
[bestx,bestf,exitflag]=fmincon(@FO,bestx,[],[],[],[],bl,bu,[],options);
disp(bestx);
Ngrey(1,:)=bestx(1:2:nWB*2);
Ngrey(2,:)=bestx(2:2:nWB*2);
disp(Ngrey)
% check convexity:
if (PesiBias<Ngrey(1,:) & PesiBias>Ngrey(2,:))
disp('OK')
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% plot figures
BV2_FUZZY_ANN_100_4FIGURES
FINALX(1,:) = bestx;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% get next alpha level
for i = 2:5
 if i ==2
   percInclusi = 80;
 elseif i == 3
   percInclusi = 60;
 elseif i == 4
   percInclusi = 40;
 elseif i == 5
   percInclusi = 20;
 end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
temp = FINALX(i-1,:);
Ngrey(1,:)=temp(1:2:nWB*2);
Ngrey(2,:)=temp(2:2:nWB*2);
clear vars temp
bu(1,1:2:nWB*2)=Ngrey(1,:);
bu(1,2:2:nWB*2)=PesiBias'-10^-6;

bl(1,1:2:nWB*2)=PesiBias'+10^-6;
bl(1,2:2:nWB*2)=Ngrey(2,:);

xi(1,1:2:nWB*2)=PesiBias'+delta/2;
xi(1,2:2:nWB*2)=PesiBias'-delta/2;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% define properties for sce ua algorithm:
ngs=5;
maxn=10000;
kstop=50;
pcento=0.1;
peps=0.001;
iseed=-1;
iniflg=1;
356

[bestx,~] = 
scceuEST('FO',xi,bl,hu,maxn,kstop,pcento,peps,ngs,iseed,iniflg);
clear vars ngs maxn kstop pcento peps iseed iniflg
disp(bestx);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% 
% define properties for fmincon refinement
options = optimoptions('fmincon', 'Display', 'iter-detailed', 'MaxIter', 10000,...
'MaxFunEvals', 10000, 'TolX', 1e-20, 'TolFun', 1e-20, 'TolCon', 1e-20);
[bestx, bestf, exitflag] = fmincon(@FO, bestx, [], [], [], [], bl, bu, [], options);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% 

FINALX(i,:) = bestx;
end

templ(1,1:2:nWB*2)=PesiBias';
templ(1,2:2:nWB*2)=PesiBias';

FINALX(6,:) = templ;
clear vars templ

save('Hidden_10.mat', 'net', 'tr', 'x', 't', 'y', ...
'delta', 'Nhidden', 'FINALX', 'PesiBias')

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% 

2. Cost function

function Gks=FO(x)
% % originally called KHAN_FO_2.m

global inp outp percInclusions

global NInp Nhidden Noutp nWIH nBH nWHO nBO nWB

global net

% Characterized NUMBERS GREY
% split incoming vector of best guess (x) into upper and lower parts in a
% two row vector called Ngrey ("Number grey")
Ngrey(1,:) = x(1:2:nWB*2);
Ngrey(2,:) = x(2:2:nWB*2);
Ngrey = sort(Ngrey,1,'descend');

% isolate weights from Input-Hidden layer:
W1 = Ngrey(:,1:nWIH);
% split into max(upper) and min (lower) matrices:
% upper
W1u = reshape(W1(1,:), Nhidden, NInp);
% lower
W1l = reshape(W1(2,:), Nhidden, NInp);
% isolate biases from Input-Hidden layer:
B1 = Ngrey(:,nWIH+1:(nWIH+nBH));
% split into max(upper) and min (lower) matrices:
% upper
Blu = B1(1,:);
Bll = B1(2,:);'

% isolate weights from Hidden-Output layer:
W2 = Ngrey(:,nWIH+nBH+1:(nWIH+nBH+nWHO));
% split into max(upper) and min (lower) matrices:
% upper
W2u = reshape(W2(1,:),Noutp,Nhidden);
% lower
W2l = reshape(W2(2,:),Noutp,Nhidden);
% isolate biases from Input-Hidden layer:
B2 = Ngrey(:,nWIH+nBH+nWHO+1:(nWIH+nBH+nWHO+nBO));
% split into max(upper) and min (lower) matrices:
% upper
B2u = B2(1,:)';
B2l = B2(2,:)';
clear vars Ngrey W1 B1 W2 B2

% normalize input and output (why are the all negative?):
INn = feval( 'mapminmax', 'apply', inp,
net.inputs{1}.processSettings{1});
OUTn = feval( 'mapminmax', 'apply', outp,
net.outputs{2}.processSettings{1});
% OUTn = outp;

% get total number of data points in input (length of inp):
n = size(inp,2);

% calculate the amount of data-points allowed to be outside the interval
%(e.g. for mu = 0,100% data should be within interval so n should be 0
nfuori = n*(1-percInclusi/100); %(maximum allowed outside)
% Initialization FO
% Gks = [];
% counter start for viol(?)
viol = 1;
% some sort of violation metric...
deltaV(viol) = 0;

% calculate output for upper:
y_upper = zeros(n,1);
y_lower = zeros(n,1);

for i = 1:n
% if negative than swap upper and lower weights
neg = find(INn(:,i)<0);
% first calculate uppers at hidden
Wl_temp = Wlu;
Wl_temp(:,neg) = Wll(:,neg);
temp_upper = tansig(Wl_temp*INn(:,i) + Blu);
clear vars Wl_temp

% then calculate lowers at hidden
Wl_temp = Wll;
Wl_temp(:,neg) = Wlu(:,neg);
temp_lower = tansig(Wl_temp*INn(:,i) + Bll);
clear vars Wl_temp

% now calculate upper and lower before output:
u = max([temp_upper.*W2u', temp_upper.*W2l', ...
    temp_lower.*W2u', temp_lower.*W2l'],[],2);
l = min([temp_upper.*W2u', temp_upper.*W2l', ...
    temp_lower.*W2u', temp_lower.*W2l'],[],2);

y_upper(i,:) = purelin(sum(u)+B2u);
y_lower(i,:) = purelin(sum(l)+B2l);
clear vars u l neg temp_upper temp_lower
end

% calculate the distance between upper and lower (to minimize):
Gks = nansum(y_upper-y_lower);

y_upper = y_upper';
y_lower = y_lower';

% It verifies if the observed values are contained
fuorisopra=find(OUTn>y_upper);
deltaVup=abs(OUTn(fuorisopra)-y_upper(1,fuorisopra));

fuorisotto=find(OUTn<y_lower);
deltaVdown=abs(OUTn(fuorisotto)-y_lower(1,fuorisotto));

deltaV=[deltaVup deltaVdown];
viol=length(deltaV)+1;

% if not within interval...
if (viol-1)>nfuori
    % from lower to higher
    deltaVcresc=sort(deltaV);
    Gks = Gks*prod(exp(deltaVcresc(floor(nfuori)+1:(viol-1))));
end
end

3. Inverse transformation

function Px = inversetransform2(Xobs, XFuzzy)
%%%% this function calculates the 1 - (probability of exceedance) of a fuzzy
%%%% number using the inverse tranformation used by Zhang 2009 pg 73.
%%%% Given a value of Xobs that is within the mu = 0 alpha cut interal
%%%% of the corresponding fuzzy number, interpolate to get the membership
%%%% level corresponding to Xobs. Then interpolate to get the mirror
%%%% (Xprime) of Xobs, i.e. other values in the fuzzy number with equal
%%%% membership level.
%%%% Then the membership level of these two values is sum of the
%%%% probabilty of exceedance: P(X>Xob) + P(X<Xprime)
%%%% Then use the ratio of the length between Xobs and X@mu=0R, and
%%%% X@mu=0L
%%%% to get area/probability corresponding to P(X>Xob)
%%%% Xobs is observed value 1 x 1
%%%% Xfuzzy is fuzzy value, 1 x 12
%%%% Check if observed value is larger or smaller than modal value of
%%%% fuzzy number
if sum(isnan(XFuzzy))/length(XFuzzy) == 1
    Px = NaN;
else
    if Xobs >= XFuzzy(1,12)
        Px = 100;
    elseif Xobs < XFuzzy(1,1)
        Px = 0;
    elseif Xobs >= XFuzzy(1,6)
        XR = XFuzzy(1,7:12);
        muR = 1:0.2:0;
        mu_obs = interp1(XR, muR, Xobs);
        XL = XFuzzy(1,1:6);
        Xprime = interp1(muL, XL, mu_obs);
        dXR = XFuzzy(1,12) - Xobs;
        dXL = Xprime - XFuzzy(1,1);
        AR = mu_obs/(1 + dXL/dXR)*100;
        Px = 100 - AR;
    else
        XL = XFuzzy(1,1:6);
        muL = 0:0.2:1;
        mu_obs = interp1(XL, muL, Xobs);
        XR = XFuzzy(1,7:12);
        Xprime = interp1(muR, XR, mu_obs);
        dXR = XFuzzy(1,12) - Xprime;
        dXL = Xobs - XFuzzy(1,1);
        AL = mu_obs/(1 + dXR/dXL)*100;
        Px = AL;
    end
end
end
1. Optimum bin-size and transformation algorithms

```matlab
set(0,'DefaultFigureWindowStyle','docked')
addpath ('C:\Users\Usman\Documents\MATLAB\Function Library')
addpath (genpath('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy'))
addpath (genpath('C:\Users\Usman\Documents\MATLAB\WQ_Data_load'))

clearvars
close all
clc

%%% load data:
% select data by unhighlighting required set; manually save final answer
% "temp" back into original cell (e.g. "Q_FINAL_RED") in third column.

load('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy\Data\Q_FINAL_RED.mat')
    temp = Q_FINAL_RED;

% load('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy\Data\WT_FINAL.mat','WT_FINAL')
%     temp = WT_FINAL;

% load('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy\Data\WQ_FINAL.mat','DO_FINAL')
%     temp = DO_FINAL;

for ii = 1:length(temp)
    for jj = 1:length(temp{ii,2})
        X = temp{ii,2}(jj,:);
        if sum(isnan(X))/length(X) ~= 1
            % Fix measurement accuracy by using correct sig digs (3)
            for each
                % measurement
                X = num2str(X,3);
                X = str2num(X);

            end
        end

        % get theoretical optimum nbin size, which is nbin = 0.2
        for all
            % measurements, except for flow, Q>99.9, where nbin = 2.
            % for flow
            if max(X)>=99.9
                nbin = 2;
            else
                nbin = 0.2;
```
x_min = floor(min(X)); x_max = ceil(max(X));
if x_min ~= x_max
    N_MIN = 3; % has to be greater than one
    buf = abs(diff(sort(X)));  
    dx = min(buf(logical(buf == 0)));  
    N_MAX = min(floor((x_max - x_min)/(2*dx)),50); % another option, perhaps too many: N_MAX = 100;
    N = N_MIN:N_MAX;
    SN = 100; % basically number of iterations to consider...
    D = (x_max - x_min) ./ N;
    D(D < nbin)=[];
    if ~isempty(D)
        % Computation of the Cost Function
        Cs = zeros(length(D),SN);
        for i = 1:length(D)
            shift = linspace(0,D(i),SN);
            for p = 1:SN
                edges = linspace(x_min+shift(p)-D(i)/2,...
                                x_max+shift(p)-D(i)/2,N(i)+1); % Bin edges
                ki = histc(X,edges); % Count # of events in bins
                ki = ki(1:end-1);
                k = mean(ki); % Mean of event count
                v = sum((ki-k).^2) / N(i); % Variance of event count
                Cs(i,p) = ( 2*k - v ) / D(i)^2; % The Cost Function
            end
        C = mean(Cs,2);
    clear vars N_MIN N_MAX buf dx SN i p k v shift
    else

    end

    % Optimal Bin Size Selection
    [Cmin, idx] = min(C);
    optN = N(idx); % Optimal number of bins
    optD = D(idx); % *Optimal binwidth
    optD = num2str(optD,1);
    optD = str2double(optD);
edges = x_min-optD:optD:x_max+optD;
clear vars N D Cmin idx C x_min x_max
else
    optD = nbin;
    edges = x_min-optD:optD:x_max+optD;
disp(['D is empty; ii = ' num2str(ii) '; jj = ' num2str(jj)])
end
clearvars edges optN
else
    optD = nbin;
    edges = x_min-optD:optD:x_max+optD;
disp(['x_min == x+max; ii = ' num2str(ii) '; jj = ' num2str(jj)])
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
mu = (0:0.2:1)';
nbin = optD;
clearvars optD
PI = muUSMAN(X, nbin, mu);
clear vars mu X nbin
temp{ii,3}(jj,:) = PI(:,2)';
clear vars PI
else
    mu = (0:0.2:1)';
    PI = [mu; flipud(mu)] nan(2*length(mu),1)];
clear vars mu X
    temp{ii,3}(jj,:) = PI(:,2)';
disp (['All X === NaN; ii = ' num2str(ii) '; jj = ' num2str(jj)])
end
clear vars PI
end
end
clear vars ii jj
%%% function [PI] = muUSMAN(X, nbin, mu)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% this function is a probability-possibility transformation that converts
% discrete, observed data into a fuzzy number. More details on the
% background of the transformation can be found in Dubois et al 2004
% The transformation obeys the two important principles:
% 1st is the consistency principle which means that the possibility
% must be
% higher than or equal to the probability: ensuring that something is
% possible
% before it is probable.
% 2nd is order preservation which means, for a discrete system if
% p(x1)>px2) than pi(x1)>pi(x2)
%
% SUMMARY: this function takes a generic array X (observations) and
% calculates the probability (p(x)) using a histogram approach. The
% histogram is
% defined at nbin intervals (user selected, or set at 1). Once the p(x)
% known pi(x) is calculated: first they are ordered highest to lowest,
% then separated into L and R sections, % then cumsum from x_i to x_n.
% A few caveats along the way to create trapezoidal fuzzy numbers
% (e.g. when max(p(x)) occurs more than once.
% once final pi(x) array is known, an interp1 scheme is created to find
% values at predefined my functions (provided by user, eg mu = 0:0.25:1
% or
% mu = 0:0.1:1).                                                                

if sum(isnan(X))/length(X) == 1
    PI = [[mu; flipud(mu)] nan(2*length(mu),1)];
else
    % define start and end of array used for histogram, round up or
    % down min max value of observed data
    x_start = floor(min(X)); x_end = ceil(max(X));

    % calculates the array for which the histogram is calculated:
    xbins = x_start:nbin:nbin:x_end+nbin;
    clear vars x_start x_end nbin

    % calculate histogram edges and numbers
    [N,~] = histcounts(X,xbins);
    % need to append 0 at end of N to make equal vector sizes
    N = [N 0];
    % calculate discrete pdf
    px = (N')/sum(N);
    clear vars N X
endif

% create matrix to house bins and corresponding probs
A = [px xbins'];
% now order matrix from highest to lowest
B = flipud(sortrows(A,1));
    clear vars A
% now find location in xbins for highest p(x)
C = find(B(:,1)==max(px));
    clear vars px

if length(C)==1 % this means that it will be trapezoidal
    % C gives the location of the xbins with highest p(x); if
    % multiple exist, need to figure out which is minimum x and
    % maximum x
    CL = find(B(:,2)==min(B(C,2)));
    % find min x correpsonding to
    max(px)
    CR = find(B(:,2)==max(B(C,2)));
    % find max x corresponding to
    max(px)

    % find all values of x between CL and CR: these are the
    % "modal" values of the function, where mu is forced to be
    % highest even though p(x) might be lowest, due to fuzzy
    % interval construction
    M = find(B(:,2)>B(CL,2) & B(:,2)<B(CR,2));
else
% if C is singleton, proceed as normal (all definitions as above)
CL = C;
CR = C;
M = []; % M is a null set, left in for consistency
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% calculate membership level (pi) using special_cumsum
for i = 1:length(B)
  % temp contains ordered (highest to lowest) possibility values,
  % corresponding xbin values are stored in B
  temp(i,1) = sum(B(i:end,1));
end
clear vars i CL CR

% manually assign all x with highest membership level =1
temp(C,:) = 1;
% append membership level to matrix B (probs and corresponding xbins)
D = [temp B];
clear vars temp C B

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% now all areas with equal p(x) must have equal pi(x). Because D is % sorted from highest to lowest wrt to pi and p, assign all x with % equal p(x), equal pi(x) corresponding to max(pi(x)) for that % sub-array
[n, ~] = size(D);
for i = 1:n-1
  if D(i,2) == D(i+1,2)
    % for two consecutive equal p(x), replace pi(x) to be equal
    D(i+1,1) = D(i,1);
  end
end
clear vars i n

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% assign all M with pi = 1 (modal values between extreme cases % where pi = 1)
D(M,1) = 1;
% sort D highest to lowest wrt to pi
D = flipud(sortrows(D,1));
clear vars M
% this D is the final observed membership function, however the % membership levels generated are non-standard and dependent on % observations: so an interpolation scheme is devises to calculate % the membership function at predefined mu-levels (see input mu).

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% find unique values of pi in D
id = unique(D(:,1));
% intialize membership function with zeros
mus = zeros(length(id),3);
% for each unique pi level, find correpsonding min and max values % of xbins
for i = 1:length(id)
    mus(i,:) = [id(i,1) ...
        min(D(D(:,1)==id(i),3)) max(D(D(:,1)==id(i),3))];
end
clear vars i
% sort membership levels wrt to mu
mus = flipud(mus);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% use fuzzy arithmetic to make sure that fuzzy number generated is
% convex: principle is that if something is true at higher
% membership level, it must be true at lower membership level:
[m, ~] = size(mus);
for i = 1:m-1
    if mus(i+1,2) > mus(i,2)
        % for Left side, if xbin is greater at lower membership
        % level make it equal to higher level value
        mus(i+1,2) = mus(i,2);
    end
    if mus(i+1,3) < mus(i,3)
        % for Right side, if xbin is lower at lower membership
        % level, make it equal to higher level value
        mus(i+1,3) = mus(i,3);
    end
end
clear vars i m id
clear vars xbins
% rearrange mus low to high wrt mu: this is the final empirical
% distribution
mus = flipud(mus); % flipping needed for interpolation that follows

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% interpolation scheme %%%

n = length(mu);
PI_final = zeros(n,3);
PI_final(1,:) = mus(1,:);
for i = 1:n-1
    % get value of given mu (user defines, see input)
    mul = mu(1+i,1);
    % find location of empirical mus smaller and greater than
    % assigned mu
    int1 = find(mus(:,1)<mul);
    int2 = find(mus(:,1)>=mul);

    % create matrix for 1D interpolation:
    temp = mus([int1(end,1); int2(1,1)],:);
    clear vars int1 int2
    % interpolate and store values in PI_final
    PI_final(i+1,:) = [mul interp1(temp(:,1),temp(:,2),mul) ...
        interp1(temp(:,1),temp(:,3),mul)];
end
clear vars i n mu temp mul

% final function:
PI(:,1) = [PI_final(:,1); flipud(PI_final(:,1))];
PI(:,2) = [PI_final(:,2); flipud(PI_final(:,3))];
clear vars D mus PI_final
end
end

2. Fuzzy neural network with fuzzy inputs

set(0,'DefaultFigureWindowStyle','docked')
addpath ('C:\Users\Usman\Documents\MATLAB\Function Library')
addpath (genpath('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy'))
clearvars close all
clc

% Load input and output data for calibration
load('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy\Data\Q_FINAL_RED.mat')
input1C = Q_FINAL_RED(:,3); % [1,3,5,7,9]
InputQ = cell2mat(input1C);
load('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy\Data\WQ_FINAL.mat')
input2C = WT_FINAL(:,3);
InputWT = cell2mat(input2C);

Output = DO_FINAL(:,4); % minimum crisp DO
Output = cell2mat(Output);

clear vars DO_FINAL Q_FINAL_RED WT_FINAL
clear vars input1C input2C input1V input2V
rng('default');
rng(3);
global Nhidden

mu = 1; seed to begin optimisation
% get value of fuzzy inputs at mu = 1
x1 = InputQ(:,[6 7]);
x2 = InputWT(:,[6 7]);

% calculate middle value to conduct regular ANN to get starting weights
% and biases for fuzzy ANN optimization; resize as crisp ANN
x = [nanmean(x1,2) nanmean(x2,2)]';
t = Output';

% Choose a Training Function
% For a list of all training functions type: help nntrain
% 'trainlm' is usually fastest.

trainFcn = 'trainlm'; % Levenberg-Marquardt backpropagation.

% Create a Fitting Network

hiddenLayerSize = 5;
Nhidden = hiddenLayerSize;

% fitnet uses sigmoid, then linear transfer function; check using:
% net.layers{1} or net.layers{2}
net = fitnet(hiddenLayerSize, trainFcn);
clear vars hiddenLayerSize trainFcn

% Choose Input and Output Pre/Post-Processing Functions
% For a list of all processing functions type: help nnprocess
net.input.processFcns = {'removeconstantrows', 'mapminmax'};
net.output.processFcns = {'removeconstantrows', 'mapminmax'};

% Setup Division of Data for Training, Validation, Testing
% For a list of all data division functions type: help nndivide
net.divideFcn = 'dividerand'; % Divide data randomly
net.divideMode = 'sample'; % Divide up every sample
net.divideParam.trainRatio = 50/100;
net.divideParam.valRatio = 25/100;
net.divideParam.testRatio = 25/100;

% Choose a Performance Function
% For a list of all performance functions type: help nnperformance
net.performFcn = 'mse'; % Mean Squared Error

% Choose Plot Functions
% For a list of all plot functions type: help nnplot
net.plotFcns = {'plotperform', 'plottrainstate', 'ploterrhist', ...
                'plotregression', 'plotfit'};

% Train the Network
[net, tr] = train(net, x, t);

% Test the Network
y = net(x);
e = gsubtract(t, y);
performance = perform(net, t, y);
clear vars e performance

% Recalculate Training, Validation and Test Performance
trainTargets = t .* tr.trainMask{1};
valTargets = t .* tr.valMask{1};
testTargets = t .* tr.testMask{1};

trainPerformance = perform(net, trainTargets, y);
valPerformance = perform(net, valTargets, y);
testPerformance = perform(net, testTargets, y);

clear vars trainTargets valTargets testTargets
clear vars trainPerformance valPerformance testPerformance

delete vars x1 x2 which are the values \mu = 1 interval
clear vars x1 x2
set(0, 'DefaultFigureWindowStyle', 'docked')
addpath ('C:\Users\Usman\Documents\MATLAB\Function Library')
addpath (genpath('C:\Users\Usman\Documents\MATLAB\ANN_fuzzy'))

% get weights and biases for optimization
global Nhiddenn % relisting here to call global value
global inpx1 inpx2 outp % used for fmincon and SCEUA

% start with \mu = 0 for PercInclusi = 99.5%
x1 = InputQ(:,[1 12]);
x2 = InputWT(:,[1 12]);

% inpx1, inpx2 and outp only use training subset:
inpx1 = x1(tr.trainInd,:);
inpx2 = x2(tr.trainInd,:);
outp = t(:,tr.trainInd);

global NInp Noutp nWIH nWHO nBO nWB
% NInp and Noutp are the number of input variables (2) and output variables
NInp=size(inpx1,1); Noutp=size(outp,1);
% nWIH = number of weights between INPUT & HIDDEN layer
nWIH=NInp*Nhidden;
% nBH = number of biases between INPUT & HIDDEN layer
nBH=Nhidden;
% nWHO = number of weights between HIDDEN & Output layer
nWHO=Nhidden*Noutp;
% nBO = number of biases between HIDDEN & Output layer
nBO=Noutp;
% total Number of Weights and Biases
nWB= nWIH + nBH + nWHO + nBO;

% get values of weights and biases from each layer
PesiINP1=net.IW{1,1}; Pesi12=net.LW{2,1};
Bias1=net.b{1}; Bias2=net.b{2};
ntotPB=2*Nhidden;
PesiBias = [reshape(PesiINP1,1,ntotPB), Bias1', Pesi12, Bias2];
clear vars PesiINP1 Pesi12 Bias1 Bias2
clear vars ntotPB

global percInclusi % for alpha cuts
% for alpha cut = 0, 100% of data within final interval
percInclusi = 99.5;
% define delta to start upper and lower limits of pesibias
delta = 0.5;

% define best guess (xi) and upper and lower limits (bu, bl) of
% pesibias
xi(1,1:2:nWB*2)=PesiBias' + delta/4;
xi(1,2:2:nWB*2)=PesiBias' - delta/4;
bu(1,1:2:nWB*2)=PesiBias' + delta;
bu(1,2:2:nWB*2)=PesiBias' + 10^-6; % changed - to +
bl(1,1:2:nWB*2)=PesiBias' - 10^-6;
bl(1,2:2:nWB*2)=PesiBias' - delta; % changed + to -

% define properties for sce ua algorithm:
ngs=5;
maxn=20000;
kstop=50;
pcento=0.1;
peps=0.001;
iseed=-1;
iniflg=1;

[bestx,bestf] = 
sceuauest('FOF',xi,bl,bu,maxn,kstop,pcento,peps,ngs,iseed,iniflg);
clear vars maxn kstop pcento peps iseed iniflg
disp(bestx);
disp(bestf);

% define properties for fmincon refinement
% options = optimoptions('fmincon','Display','iter-detailed','MaxIter',20000,...
% 'MaxFunEvals',20000,'TolX',1e-20,'TolFun',1e-20,'TolCon',1e-20);
% [bestx,bestf,exitflag]=fmincon(@FOf,bestx,[],[],[],[],bl,bu,[],options)

% disp(bestx);
% disp(bestf);
Ngrey(1,:) = bestx(1:2:nWB*2);
Ngrey(2,:) = bestx(2:2:nWB*2);
disp(Ngrey);

% check convexity:
% if (PesiBias<Ngrey(1,:) & PesiBias>Ngrey(2,:))
% disp('OK')
% end

FINALX(1,:) = bestx;
FINALF(1,:) = bestf;

% get next alpha level
for i = 2:5
    if i == 2
        toc
        percInclusi = 80;
        x1 = InputQ(:,[2 11]);
        x2 = InputWT(:,[2 11]);
        inpX1 = x1(tr.trainInd,:);'
        inpX2 = x2(tr.trainInd,:);
    elseif i == 3
        toc
        percInclusi = 60;
        x1 = InputQ(:,[3 10]);
        x2 = InputWT(:,[3 10]);
elseif i == 4
toc
percInclusi = 40;
x1 = InputQ(:,[4 9]);
x2 = InputWT(:,[4 9]);
inpx1 = x1(tr.trainInd,:)';
inpx2 = x2(tr.trainInd,:)';
elseif i == 5
toc
percInclusi = 20;
x1 = InputQ(:,[5 8]);
x2 = InputWT(:,[5 8]);
inpx1 = x1(tr.trainInd,:)';
inpx2 = x2(tr.trainInd,:)';
% else i == 6
% percInclusi = 0.5;
% x1 = InputQ(:,[6 7]);
% x2 = InputWT(:,[6 7]);
% inpx1 = x1(tr.trainInd,:)';
% inpx2 = x2(tr.trainInd,:)';
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
temp = FINALX(i-1,:);
Ngrey(1,:)=temp(1:2:nWB*2);
Ngrey(2,:)=temp(2:2:nWB*2);
clear vars temp

bu(1,1:2:nWB*2)=Ngrey(1,:);
bu(1,2:2:nWB*2)=PesiBias'-10^-6;
bl(1,1:2:nWB*2)=PesiBias'+10^-6;
bl(1,2:2:nWB*2)=Ngrey(2,:);
% xi(1,1:2:nWB*2)=PesiBias'+delta/2;
% xi(1,2:2:nWB*2)=PesiBias'-delta/2;
xi(1,1:2:nWB*2)=(Ngrey(1,:)'-PesiBias')/4;
xi(1,2:2:nWB*2)=(PesiBias'-Ngrey(2,:)')/4;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% define properties for sce_ua algorithm:
ngs=5;
maxn=20000;
kstop=50;
pcento=0.1;
peps=0.001;
iseed=-1;
iniflg=1;

[bestx,bestf] = 
sceuaEST('FOf',xi,bl,bu,maxn,kstop,pcento,peps,ngs,iseed,iniflg);
clear vars ngs maxn kstop pcento peps iseed iniflg
disp(bestx);
disp(bestf);
% define properties for fmincon refinement
options = optimoptions('fmincon','Display','iter-detailed','MaxIter',20000,...
   'MaxFunEvals',20000,'TolX',1e-20,'TolFun',1e-20,'TolCon',1e-20);

[bestx,bestf,exitflag]=fmincon(@FOf,bestx,[],[],[],[],bl,bu,[],options);

% disp(bestx);
% disp(bestf);
FINALX(i,:) = bestx;
FINALF(i,:) = bestf;
end

% get next alpha level
toc
temp = FINALX(5,:);
tempU = temp(1,1:2:nWB*2);
tempL = temp(1,2:2:nWB*2);
PesiBias_Final = mean([tempU; tempL]);
disp(PesiBias_Final);

FINALX(6,1:2:nWB*2) = PesiBias_Final;
FINALX(6,2:2:nWB*2) = PesiBias_Final;
clear vars PesiBias_Final temp tempU tempL i

save('Hidden_5f_test_02.mat','net','tr','t','y','x',...
     'Nhiddenn','delta','FINALX','FINALF','PesiBias')
toc

3. Cost function

function Gks=FOf(x)
% original called KHAN_FO_2.m
global inpx1 inpx2 outp percInclusi
global NInp Nhiddenn Noutp nWIH nWHO nBO nWB

% Characterized NUMBERS GREY
% split incoming vector of best guess (x) into upper and lower parts in a
% two row vector called Ngrey ("Number grey")
Ngrey(1,:) = x(1:2:nWB*2);
Ngrey(2,:) = x(2:2:nWB*2);
Ngrey=sort(Ngrey,1,'descend');

% isolate weights from Input-Hidden layer:
W1=Ngrey(:,1:nWIH);
% split into max(upper) and min (lower) matrices:
% upper
W1u=reshape(W1(1,:),Nhiddenn,NInp);
% lower
W1l=reshape(W1(2,:),Nhiddenn,NInp);
% isolate biases from Input-Hidden layer:
B1=Ngrey(:,nWIH+1:(nWIH+nBH));
% split into max(upper) and min (lower) matrices:
% upper
B1u = B1(1,:);’;
B1l = B1(2,:);’;

% isolate weights from Hidden-Output layer:
W2=Ngrey(:,nWIH+nBH+1:(nWIH+nBH+nWHO));
% split into max(upper) and min (lower) matrices:
% upper
W2u=reshape(W2(1,:),Noutp,Nhidden);
% lower
W2l=reshape(W2(2,:),Noutp,Nhidden);
% isolate biases from Input-Hidden layer:
B2=Ngrey(:,nWIH+nBH+nWHO+1:(nWIH+nBH+nWHO+nBO));
% split into max(upper) and min (lower) matrices:
% upper
B2u = B2(1,:)’;
B2l = B2(2,:)’;

clear vars Ngrey W1 B1 W22

% normalize input and output (why are the all negative?):
INnL = feval( ‘mapminmax’, ’apply’, [inpx1(1,:); inpx2(1,:)],
net.inputs{1}.processSettings{1});
INnR = feval( ‘mapminmax’, ’apply’, [inpx1(2,:); inpx2(2,:)],
net.inputs{1}.processSettings{1});

x1 = [INnL(1,:)' INnR(1,:)’];
x2 = [INnL(2,:)' INnR(2,:)’];

OUTn = feval( ‘mapminmax’, ’apply’, outp,
net.outputs{2}.processSettings{1});

% get total number of data points in input (length of inp):
n = size(inpx1,2);

% calculate the amount of data-points allowed to be outside the
% interval (%e.g. for mu = 0, 100% data should be within interval so n should be 0
nfuori = n*(1-percInclusi/100); % (maximum allowed outside)

% Initialization FO
Gk = [ ];
% counter start for viol(?)
viol = 1;
% some sort of violation metric...
deltaV(viol) = 0;

% calculate output for upper:
y_upper = zeros(n,1);
y_lower = zeros(n,1);

inter = zeros(Nhidden,2);
templ = zeros(Nhidden,2);
temp2 = zeros(Nhidden,2);
temp3 = zeros(Nhidden,2);

for j = 1:n
    for i = 1:Nhidden
        % for x1: calculate: W1(:,1)*x1
        W_tempx1=[W1l(i,1) W1u(i,1)];
        temp1(i,:) = ... 
            [min(...
                [x1(j,1)*W_tempx1(1,1);
                 x1(j,1)*W_tempx1(1,2);
                 x1(j,2)*W_tempx1(1,1);
                 x1(j,2)*W_tempx1(1,2))] ... 
            max(...
                [x1(j,1)*W_tempx1(1,1);
                 x1(j,1)*W_tempx1(1,2);
                 x1(j,2)*W_tempx1(1,1);
                 x1(j,2)*W_tempx1(1,2)]);
        % for x2: calculate: W1(:,1)*x2
        W_tempx2=[W1l(i,2) W1u(i,2)];
        temp2(i,:) = ... 
            [min(...
                [x2(j,1)*W_tempx2(1,1);
                 x2(j,1)*W_tempx2(1,2);
                 x2(j,2)*W_tempx2(1,1);
                 x2(j,2)*W_tempx2(1,2))] ... 
            max(...
                [x2(j,1)*W_tempx2(1,1);
                 x2(j,1)*W_tempx2(1,2);
                 x2(j,2)*W_tempx2(1,1);
                 x2(j,2)*W_tempx2(1,2)]);
        temp_B = [B1l(i,1) B1u(i,1)];
        % intermediate value:
        inter(i,:) = tansig((temp1(i,:)+temp2(i,:)) + temp_B);
    end
    for i=1:Nhidden
        temp3(i,:) = ... 
            [min(...
                [inter(i,1)*W21(1,i);
                 inter(i,1)*W2u(1,i);
                 inter(i,2)*W21(1,i);
                 inter(i,2)*W2u(1,i))] ... 
            max(...
                [inter(i,1)*W21(1,i);
                 inter(i,1)*W2u(1,i);
                 inter(i,2)*W21(1,i);
                 inter(i,2)*W2u(1,i)]);
    end
    clear vars tempt1 tempt2 temp_B inter
    final = purelin(sum(temp3) + [B21 B2u]);
y_lower(j,:) = final(:,1);
y_upper(j,:) = final(:,2);
end

% calculate the distance between upper and lower (to minimize):
Gks = nansum(y_upper-y_lower);

y_upper = y_upper';
y_lower = y_lower';

% It verifies if the observed values are contained
fuorisopra=find(OUTn>y_upper);
deltaVup=abs(OUTn(fuorisopra)-y_upper(1,fuorisopra));

fuorisotto=find(OUTn<y_lower);
deltaVdown=abs(OUTn(fuorisotto)-y_lower(1,fuorisotto));

deltaV=[deltaVup deltaVdown];
viol=length(deltaV)+1;

% if not within interval...
if (viol-1)>nfuori
    % from lower to higher
    deltaVcresc=sort(deltaV);
    Gks = Gks*prod(exp(deltaVcresc(floor(nfuori)+1:(viol-1))));
end
end