

Comparing A Bayesian and Fuzzy Number Approach to Uncertainty Quantification in Short-Term Dissolved Oxygen Prediction

U. T. Khan and C. Valeo*

Mechanical Engineering, University of Victoria, Victoria, British Columbia V8W 2Y2, Canada

Received 31 March 2015; revised 22 October 2015; accepted 24 November 2015; published online 19 September 2017

ABSTRACT. A new autoregressive-type, updating fuzzy linear regression method is proposed to predict daily dissolved oxygen (DO) concentration in a highly urbanized riverine environment. Results of this model are compared to results from an updating Bayesian regression model. Both methods use lagged daily DO (at four different lags) as the independent variable. Uncertainty in the models is represented by a fuzzy number based approach in the first case, and by a Bayesian framework in the second. Real-time data from the Bow River in Calgary, Canada is used to calibrate the models sequentially to mimic a real-time updating model. Four different performance metrics were used to measure the performance of each model. Lastly, the input data resolution is reduced to measure the impact on model performance. Results show that the physical system can be adequately characterized using only one year of data. Both approaches can capture the general trend of daily DO, but the fuzzy number based method can better capture the changes in observed variability. The metrics for both models are comparable, with the one-day lag case categorized as “very good”; however, the performance reduces at higher lags. The fuzzy number method captures more low DO events than the Bayesian approach, with a much lower mean squared error. A possibility to probability transformation is used to highlight the risk of low DO days for the fuzzy case. Lastly, reducing the input data resolution from 96 to 6 points per day has a minimal impact on model performance, suggesting the limited efficacy or utility in increasing sampling rates.

Keywords: dissolved oxygen, Bayesian linear regression, fuzzy linear regression, uncertainty analysis, river water quality

1. Introduction

The dissolved oxygen (DO) concentration of a water body is a key indicator of overall aquatic ecosystem health (Dorfman and 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 (Pogue and Anderson, 1995; Hauer and Hill, 2007). 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.

2. An Overview of Modelling DO Using Data-Driven Approaches

The issues mentioned in the previous section 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 and Valeo, 2014a). 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; cannot accommodate the interactions between various system parameters; require a complete, or near-complete understanding of all the factors affecting DO; or that the underlying processes governing DO are unknown (Cox, 2003; Xu et al., 2012; Khan and Valeo, 2014a). Secondly, given the rapid changes in urban areas, any long-term inferences made using static (i.e. non-updating) physically-based or conceptual models, do not capture the impacts of these short-

* Corresponding author. Tel.: +1 250-721-8623; fax: +1 250-721-6051.

E-mail address: valeo@uvic.ca (C. Valeo).

term changes (Radwan et al., 2009). Uncertainty in measured data and temporal and spatial variability compound these problems (Robinson et al., 2009; Khan and Valeo, 2014a) 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 research, 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 generalized links between input and output datasets (Solomatine and 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 and 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 utilize 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. (2013) 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. Hedd am (2014) for a river in Oregon, USA, Ay and Kisi (2011) 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.

2.1. Modelling DO in the Bow River in Calgary, Canada

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., 2011; Khan et al., 2013; Khan and Valeo, 2014a; Khan and Valeo, 2014b). 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, 2010). The river has an average annual discharge of 90 m³/s, an average width and depth of 100 m and 1.5 m, respectively (Khan and Valeo, 2014a). Given its size, and the rate of development in the watershed – the river is one of the most regulated rivers in Alberta (BRBC, 2010) – the Bow River is approaching its 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 water body. 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 and Valeo (2014a) has shown that data-driven models, particularly those that use abiotic factors (such as river discharge, or water temperature) as inputs, and regression models, 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 discharge rate downstream of a treatment plant) could improve DO at times when the risk of low DO was high.

2.2. Uncertainty Analysis Using Bayesian and Fuzzy Number Frameworks

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 theoretic)

tical 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 and Solomatine, 2008). Scarcity in data and measurement errors can contribute to uncertainty in modelling projections (Zhang et al., 2009; El-Baroudy and Simonovic 2006). Additionally, uncertainty may be increased when data from multiple sources are used, integrated and propagated (Porter et al., 2000; Shrestha and Nestmann 2009; Shrestha and Simonovic 2010; Suo et al., 2013). The proper identification and propagation of this uncertainty is critical for understanding and evaluating model prediction (Shrestha and 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 and Mannina et al., 2010; Tyrallis and Koutsoyiannis, 2014; Gelman et al., 2014). The use of Bayesian based methods has increased 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 and Dodge, 1993; Shrestha and 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). In contrast to this, non-Bayesian techniques only provide deterministic results, typically the mean value rather than the entire probability distribution.

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 and 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 and 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 and Prade, 1997; Ozbek and 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 and Valeo, 2014a; Huang, et al., 2010; Tan et al., 2015). They are more suitable when data is missing, incomplete, combined from multiple sites, to represent uncertainty that is not purely random in nature, or when not enough information is available to develop a probability distribution function (Bárdossy et al., 1990; Guyonnet et al., 2003; Zhang and Achari, 2010; Huang et al., 2010; Xu et al., 2014). A fuzzy number is a convex and normal fuzzy set (Zadeh, 1965). A fuzzy number A , with elements $[a_1, a_2, \dots, a_n]$ is defined by its membership function, μ , which assigns a membership level for each element of A . For a fuzzy number, μ is bounded in the interval $[0, 1]$. A linear (or triangular shaped) membership function is typically used to define fuzzy numbers. This type of function is selected due to 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 $\mu = 1$), resulting in a triangular shaped function. Following this, in probability-possibility frameworks, a triangular membership function corresponds to a uniform probability distribution – the least specific distribution since any value is equally probable and hence, represents the most uncertainty (Dubois et al., 2004; Dubois and Prade, 2015). However, these types of functions are not the most suitable for environmental data (Khan et al., 2013; Khan and Valeo, 2014a). 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 utilize 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., 2013; Khan and Valeo, 2014a). Since fuzzy numbers are based on possibility theory, it means that some of the strict assumptions in many probability models (e.g. the Normality assumption for linear regression) 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 for uncertainty analysis (e.g. Li *et al.*, 2013; Hu *et al.*, 2014; Xu *et al.*, 2014; Tan *et al.*, 2015). 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 (2014a).

2.3. Objectives

The above review shows that there is considerable opportunity to use both Bayesian and fuzzy number based methods, along with data-driven models (particularly regression based methods) to improve the prediction of environmental parameters such as DO. In this research, 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 autoregressive 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.

3. Methods

3.1. Data Collection and Site Description

The Bow River basin, located in southern Alberta, Canada, is approximately 25,123 km² in size and the river is 645 km long (BRBC, 2010). 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 the Bearspaw reservoir) and downstream (currently at Highwood, see Figure 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 and Valeo, 2014b). 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 and Valeo, 2014a). 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 Figure 1) 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.

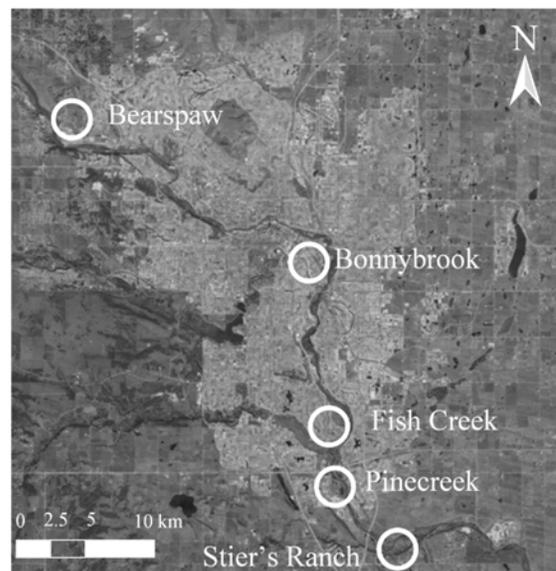


Figure 1. Aerial image of the City of Calgary showing the locations of the water and wastewater treatment plants, and monitoring locations.

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 $\pm 2\%$ 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.

3.2. Model Implementation

A preliminary analysis of the DO concentration (Khan and Valeo, 2014b) 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:

$$DO(t) = f(DO(t-d)) \tag{1}$$

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:

$$DO = \beta_0 + \beta_1 DO_L \tag{2}$$

where DO is the observed mean daily DO concentration, DO_L is the lagged mean DO concentration, and β_0 and β_1 are the regression coefficients. Typically, in a numerical modelling setup, 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 year’s worth of data (2004) was used to calibrate the regression model (either Bayesian or fuzzy), while data from the following year (2005) was used 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 1. Details of Data Used to Calibrate and Validate each Model in the Recursive Modelling Scheme

Model no.	Calibration data	Validation data
M01	2004	2005
M02	2004-05	2006
M03	2004-06	2007
M04	2004-07	2008
M05	2004-08	2009
M06	2004-09	2010
M07	2004-10	2011
M08	2004-11	2012
M09	2004-12	2013

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.

3.3. 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 and Martin, 2009; Wakefield, 2013).

In this research, two Bayesian frameworks are adopted to predict DO concentration: one uses an objective (or *non-informative*) prior and the other uses independent and informative priors. Details of the mathematical development of both these methods (stemming from the Normal regression model) are included in the Supplementary Material. The first case is used when prior information may not be available, making it difficult to describe the prior. Using a non-objective prior, the Bayesian regression development of the predictive distribution (\mathbf{y}^*) can be found analytically and is multivariate t :

$$(\mathbf{y}^* | \mathbf{y}) \sim t(X^* \bar{\beta}, \bar{s}^2 [I + X^* \bar{V} X^{*T}], \bar{\nu}) \tag{3}$$

where X^* are new observations used for prediction, \mathbf{y}^* are the

predicted values, $\bar{\beta}$ are the posterior estimate of regression coefficients with posterior variance \bar{V} , \bar{s}^2 are the posterior estimates of the standard error, I is the identity matrix, and $\bar{\nu}$ are the posterior degrees of freedom. This formulation was used to model daily mean DO (i.e. y^*) using DO_L (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 1). This is because for this case, in 2004, there is no information available to choose a so-called informative, natural conjugate prior.

The second case (independent and informative priors) is used when sufficient information is available to define the prior. Thus, for subsequent models (i.e. M02 onwards), independent and informative priors were used. For example, in M02, two years of data are used to calibrate the Bayesian regression model (2004 and 2005) while data from 2006 is used to validate the model. For the calibration procedure, the posterior estimates the model coefficients from M01 are used as the priors for the coefficient estimates for the M02 model. 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 the Supplementary Material). The predictive distribution of y can be estimated numerically: for each pair of β and τ (from the Gibb's sampler), a sample for y^* can be drawn from the predictive distribution:

$$(y^*) \sim N(X^* \beta, 1/\tau) \quad (4)$$

where X^* (DO_L) are new observations used for prediction and y^* (DO) are the predicted values, and τ is the inverse of the variance. 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.

3.4. 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 and Valeo, 2014a). 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 and 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 (Maddala, 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; Maddala, 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 and Valeo (2014a) and is unique in that it uses fuzzy inputs, outputs and fuzzy coefficients to parameterize 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 and Valeo, 2014a).

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

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x} + \tilde{\varepsilon} \quad (5)$$

where the coefficients $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are fuzzy numbers. The objective is to solve the following least-squares optimization problem:

$$\min r(\tilde{\beta}_0, \tilde{\beta}_1) = \sum_{i=1}^n d^2(\tilde{y}_i, \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i) \quad (6)$$

where $d^2(\tilde{y}_i, \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i) = \int_{\mu} [\tilde{y}_i - \tilde{\beta}_0 - \tilde{\beta}_1 \tilde{x}_i]_{\mu}^2$ for $i = 1, 2, \dots, n$ and $\mu = 0$ to 1. The metric d measures the sum of the squared-deviations of the observed (\tilde{y}_i) and predicted ($\tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i$) intervals $[\dots]_{\mu}$, for all α -cuts between $\mu = 0$ and $\mu = 1$. Using fuzzy arithmetic ensures that the coefficients $\tilde{\beta}_0$ and $\tilde{\beta}_1$ 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 α -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.

3.4.1 Creating Fuzzy Numbers

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 and Valeo (2014 a) to create non-linear discrete fuzzy numbers. There is consi-

derable literature on how to construct fuzzy numbers from observed data (e.g. Mauris, 2013), and the transformation used in this research is as follows:

For a discrete system, this can be represented as:

$$\text{if } p(x_1) > p(x_2) > \dots > p(x_n)$$

Then the possibility distribution of x ($\pi(x)$), follows the same order, that is:

$$\pi(x_1) > \pi(x_2) > \dots > \pi(x_n)$$

The transformation is then given by:

For $p(x_1) > p(x_2) > \dots > p(x_n)$:

$$\pi(x_1) = 1$$

$$f(x) = \sum_{j=1}^n P_j, \text{ if } P_{i-1} > P_i \quad (7)$$

$$f(x) = \sum_{j=1}^n \pi(x_{i-1}), \text{ else}$$

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 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 α -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 and Zhou, 2010). 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. This transformation follows the consistency principle, i.e., an event must be possible before it is probable, and order preservation, i.e., 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 .

3.4.2 Risk Analysis

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:

$$\mu(x) = A_L + A_R \quad (8)$$

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:

$$A_L = P(x < X) = \frac{\mu(x)}{1 + \frac{b-x'}{x-a}} \quad (9)$$

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 (CCME, 1999). In this research, 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 μ) of low DO.

3.5. 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:

$$NSE = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

$$RSR = \frac{\sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (11)$$

$$PBIAS = \frac{\sum_{i=1}^n y_i - \hat{y}_i}{\sum_{i=1}^n y_i} \times 100\% \quad (12)$$

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. These three metrics were calculated in part to use a model rating system proposed by Moriasi et al. (2007) to facilitate direct comparison with existing models.

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

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (13)$$

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.

4. Results and Discussion

4.1. 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 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 σ^2 (i.e. $1/\tau$) decreases with each subsequent year of added data, while β_0 increases and remains relatively unchanged. Figure 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. Note that this would not be possible using a non-Bayesian, i.e. ordinary linear regression, approach.

Figure 3 shows the fuzzy regression coefficients calculated using FLR. The results are similar to the BLR coefficient results: β_0 increases with more data (Figure 3a), while β_1 decreases (Figure 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 β_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 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

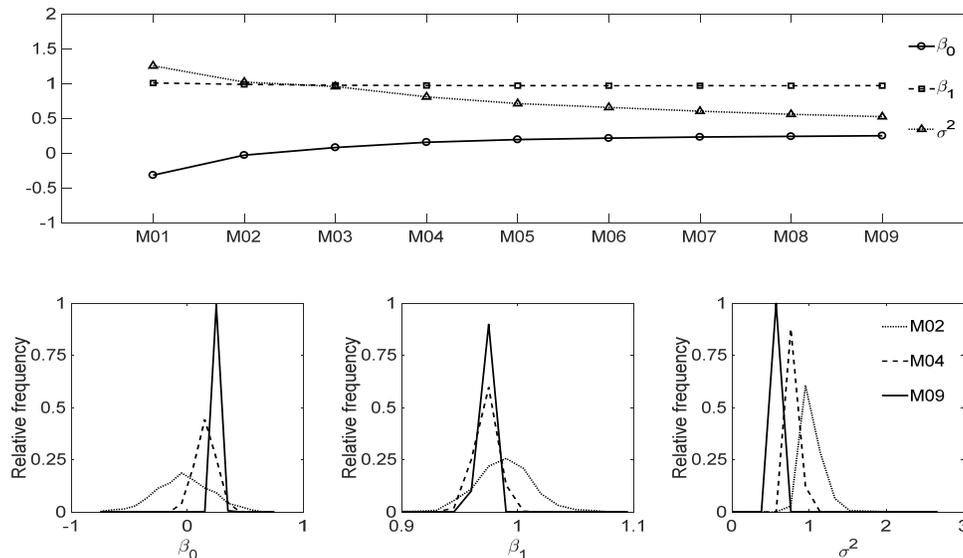


Figure 2. Evolution of the BLR parameters: (top) the change in mean values of the two coefficients and the variance; (bottom) change in the approximate pdf of the variables for three cases.

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 characterized using only one year's worth of data.

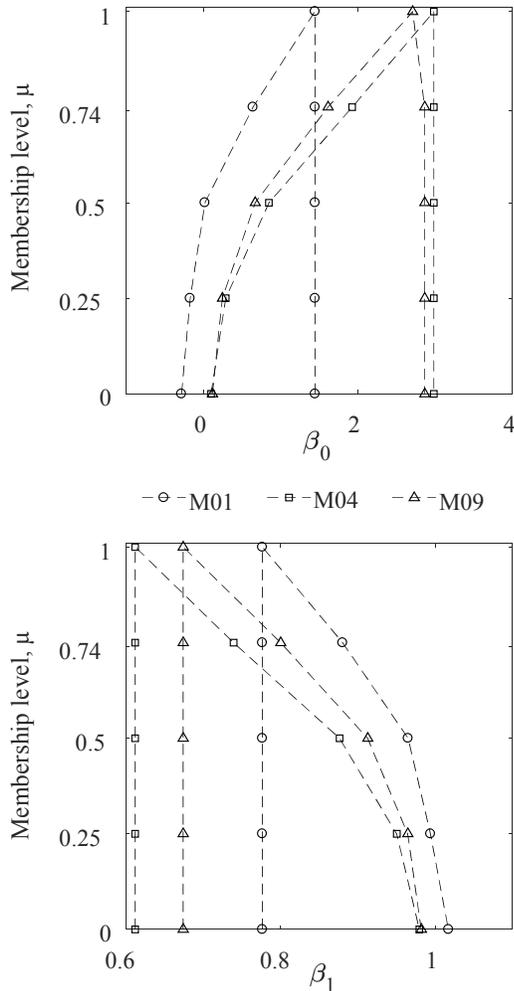


Figure 3. Evolution of the FLR coefficients for the 1-day lag and resolution = 96 case, shown here for the three cases.

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 inter-

val 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.

The impact of selecting a non-informative prior for M01 and independent priors for M04 and M09 is also demonstrated in Figure 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 summarized via the performance metrics discussed later.

Figure 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 5a -c). This figure also shows a plot of lagged-DO versus DO for the observed, BLR and FLR predicted datasets (see Figure 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.

The error metrics for each model are summarized in Figure 6 to assess model performance. Figure 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 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.

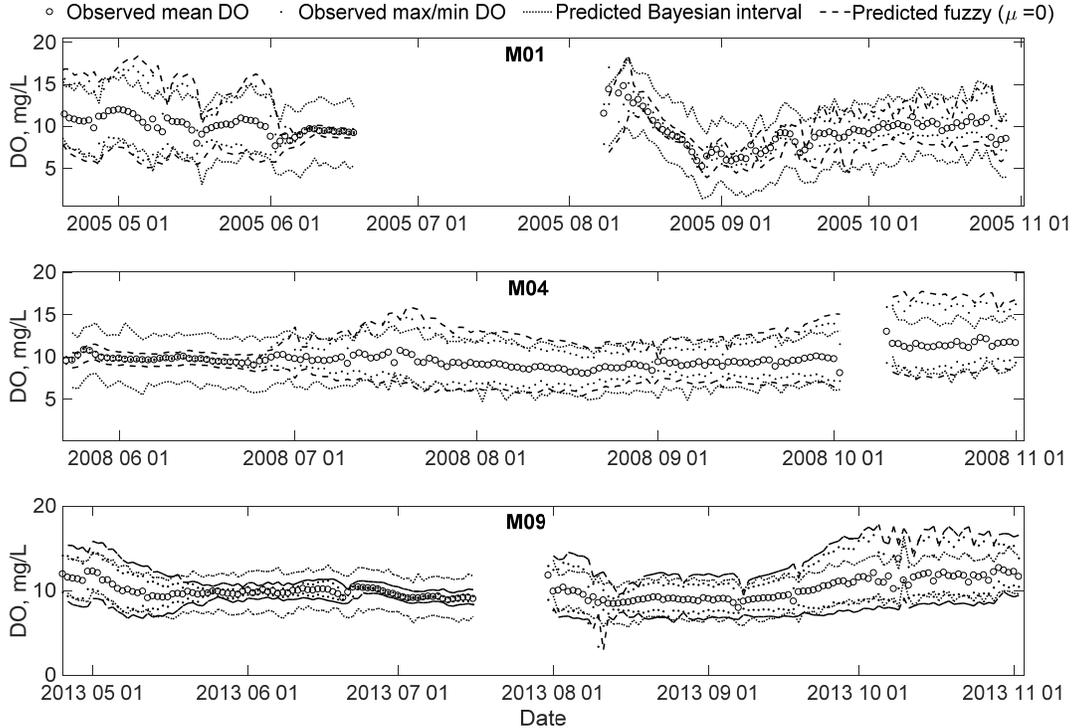


Figure 4. Trend plots for validation results for M01, M04 and M09 for the 1-day lag, resolution = 96 case.

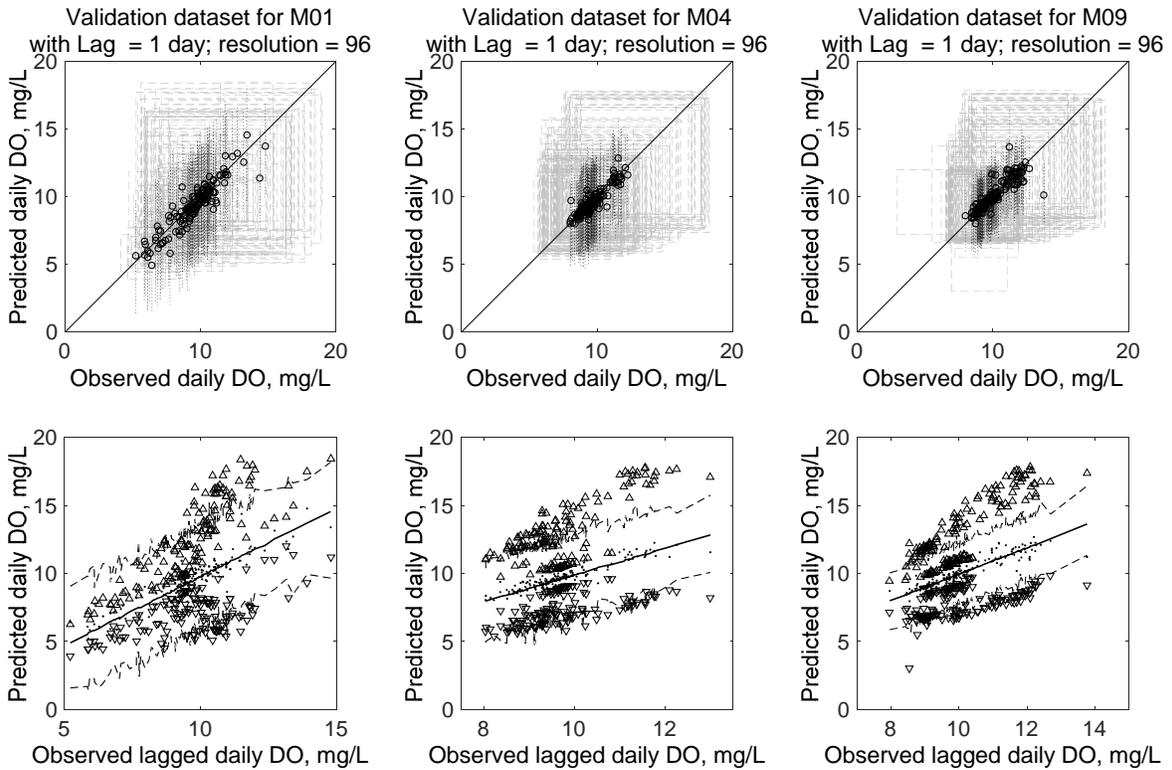


Figure 5. (top): A comparison of (top) observed vs. predicted daily DO concentration for the BLR and FLR cases; and (bottom) observed (dots), BLR (black lines) and FLR results (triangles for $\mu=0$).

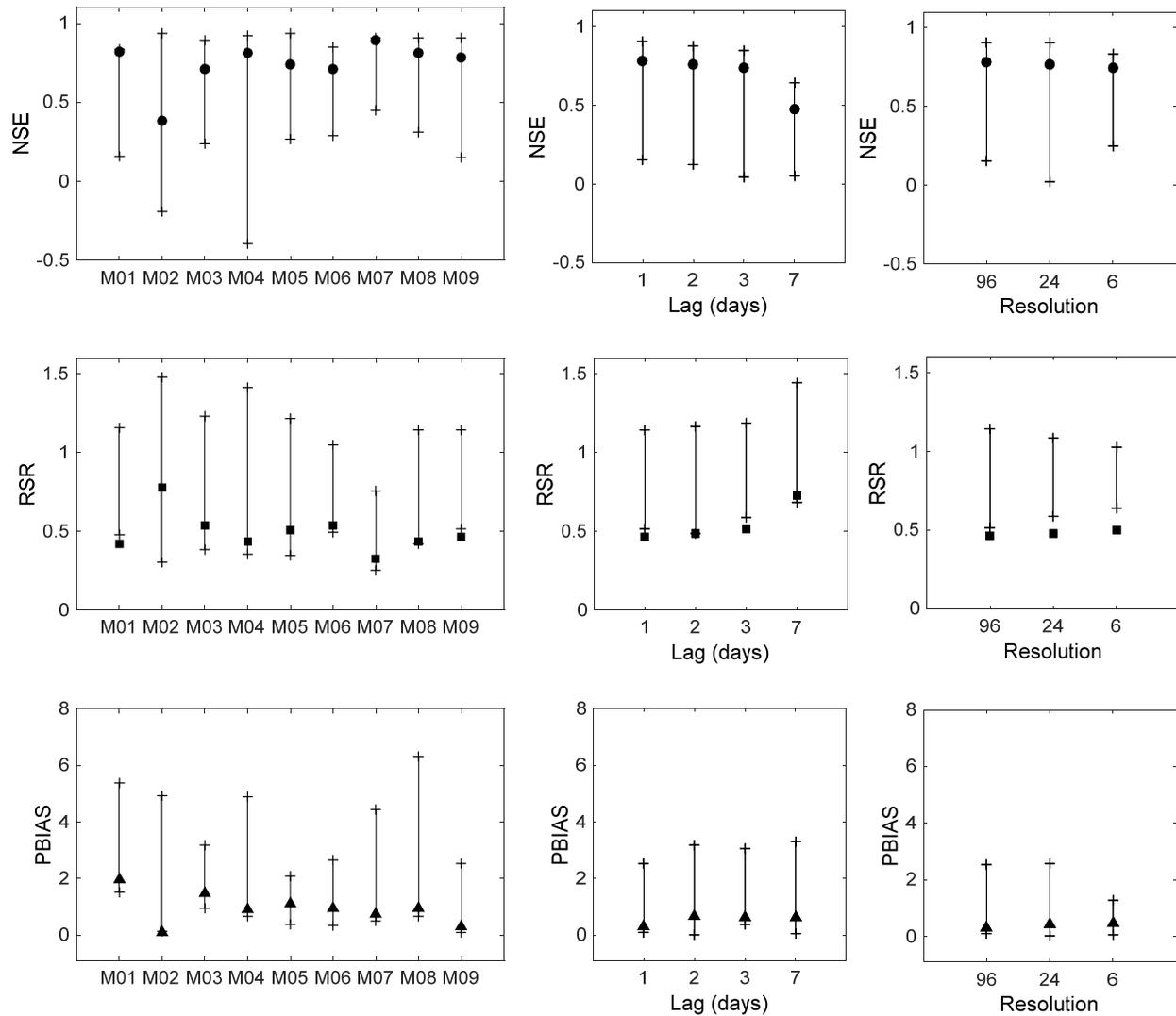


Figure 6. A summary of the three performance metrics for each model; the marker represents the mean BLR results, and the lines represent the FLR interval results.

Lastly, Figure 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 change 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 6 are only shown for one case – results for all other model cases follow the same pattern and are included in the Supplementary Material.

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.

4.2. 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 research) 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 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 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 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 7d): 1.12 for FLR compared to 4.52 for BLR.

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 $P(\overline{DO} < \min(DO_{obs}))$, 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 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).

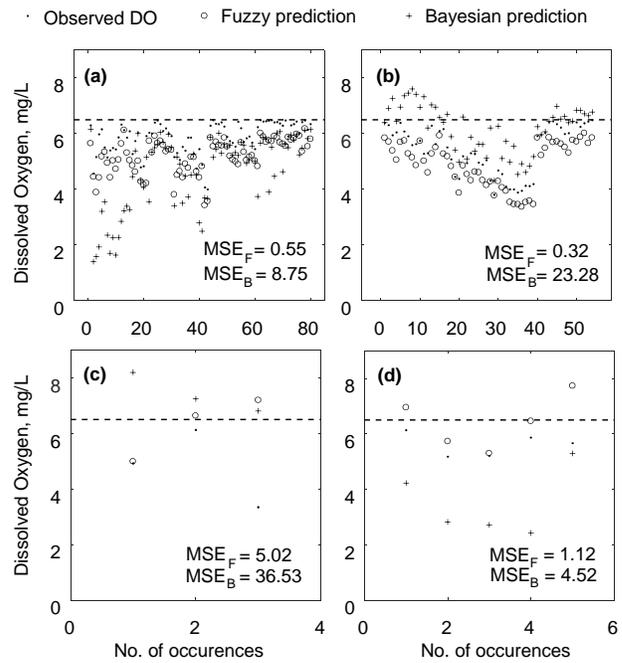


Figure 7. Observed minimum DO compared BLR and FLR (at $\mu = 0$) results; (a) when both methods capture the minimum value, (b) when only the FLR interval captures the minimum; (c) when neither captures the minimum; (d) when only the BLR interval captures the minimum.

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: $P(\overline{DO} < 6.5 \text{ mg/L})$. 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,

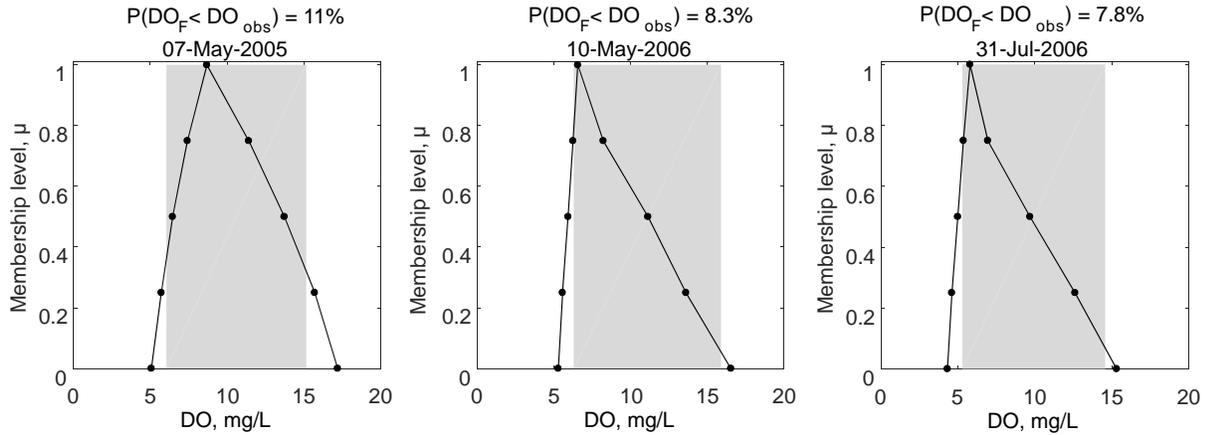


Figure 8. Results of the inverse transformation showing the risk of low DO for the observed (grey), BLR (dashed) and FLR (solid) results.

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. 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%.

5. Conclusions

In this research, an autoregressive 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 relative-

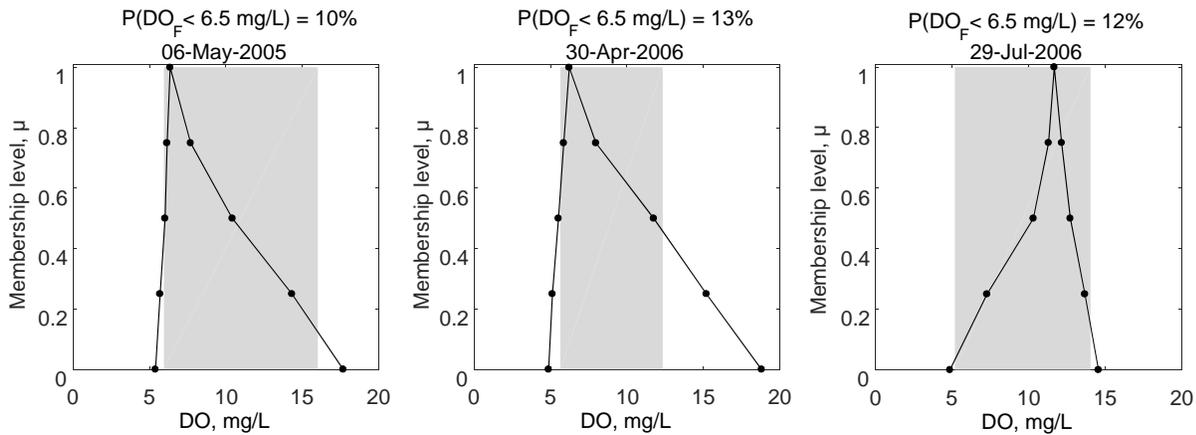


Figure 9. Results of the inverse transformation showing the risk of low DO for the observed (grey), BLR (dashed) and FLR (solid with dots) results, along with the warning level at 6.5 mg/L (solid).

vely 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.

Acknowledgement. The authors would like to thank the City of Calgary – Water Resources for the data used for this research, as well as NSERC, the Ministry of Higher Education (BC), and the University of Victoria for their financial support of this research.

References

- Adams, K.A., Barth, J.A., and Chan, F. (2013). Temporal variability of near-bottom dissolved oxygen during upwelling off central Oregon. *J. Geophys. Res. Oceans*, 118(10), 4839-4854. <http://dx.doi.org/10.1002/jgrc.20361>
- Antanasijević, D., Pocaajt, V., Perić-Grujić, A., and Ristić, M. (2014). Modelling of dissolved oxygen in the Danube River using artificial neural networks and Monte Carlo Simulation uncertainty analysis. *J. Hydrol.*, 519, 1895-1907. <http://dx.doi.org/10.1016/j.jhydrol.2014.10.009>
- Ay, M. and Kisi, O. (2011). Modeling of dissolved oxygen concentration using different neural network techniques in Foundation Creek, El Paso County, Colorado. *J. Environ. Eng.*, 138(6), 654-662. [http://dx.doi.org/10.1061/\(ASCE\)EE.1943-7870.0000511](http://dx.doi.org/10.1061/(ASCE)EE.1943-7870.0000511)
- Bárdossy, A. (1990). Note on fuzzy regression. *Fuzzy Sets Syst.*, 37(1), 65-75. [http://dx.doi.org/10.1016/0165-0114\(90\)90064-D](http://dx.doi.org/10.1016/0165-0114(90)90064-D)
- Birkes, D. and Dodge, Y. (1993). *Alternative Methods of Regression*, John Wiley & Sons, New York, NY, USA. <http://dx.doi.org/10.1002/9781118150238>
- BRBC (2010). Profile of the Bow River Basin: Hydrology. http://wsow.brbc.ab.ca/index.php?option=com_content&view=article&id=259&Itemid=83.
- Canadian Council of Ministers of the Environment (1999). Canadian water quality guidelines for the protection of aquatic life: Dissolved oxygen (freshwater). In: *Canadian environmental quality guidelines*, Canadian Council of Ministers of the Environment, Winnipeg, MB, Canada.
- Chang, Y.H.O. and Ayyub, B.M. (2001). Fuzzy regression methods-a comparative assessment. *Fuzzy sets syst.*, 119(2), 187-203. [http://dx.doi.org/10.1016/S0165-0114\(99\)00091-3](http://dx.doi.org/10.1016/S0165-0114(99)00091-3)
- Chang, F.J., Tsai, Y.H., Chen, P.A., Coynel, A., and Vachaud, G. (2015). Modeling water quality in an urban river using hydrological factors-Data driven approaches. *J. Environ. Manage.*, 151, 87-96. <http://dx.doi.org/10.1016/j.jenvman.2014.12.014>
- Chen, T. and Martin, E. (2009). Bayesian linear regression and variable selection for spectroscopic calibration. *Anal. Chim. Acta*, 631(1), 13-21. <http://dx.doi.org/10.1016/j.aca.2008.10.014>
- Cox, B.A. (2003). A review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers. *Sci. Total Environ.*, 314, 335-337. [http://dx.doi.org/10.1016/S0048-9697\(03\)00063-9](http://dx.doi.org/10.1016/S0048-9697(03)00063-9)
- Dorfman, R. and Jacoby, H.D. (1972). An illustrative model of a river basin pollution control. In Dorfman, R. Jacoby, H.D., & Thomas Jr., H.A. (Eds.), *Models for Managing Regional Water quality*, pp. 84-141. Harvard University Press, Cambridge, MA, USA. <http://dx.doi.org/10.4159/harvard.9780674419216.c3>
- Dubois, D. and Prade, H. (1997). Bayesian conditioning in possibility theory. *Fuzzy Sets Syst.*, 92(2), 223-240. [http://dx.doi.org/10.1016/S0165-0114\(97\)00172-3](http://dx.doi.org/10.1016/S0165-0114(97)00172-3)
- Dubois, D., Prade, H., and Sandri, S. (1993). On possibility/probability transformations. In Lowen, R. & Roubens, M. (Eds.), *Fuzzy Logic*, pp. 103-112. Kluwer Academic Publishers, Dordrecht, Netherlands. http://dx.doi.org/10.1007/978-94-011-2014-2_10
- Dubois, D., Foulloy, L., Mauris, G., and Prade, H. (2004). Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliab. Comput.*, 10(4), 273-297. <http://dx.doi.org/10.1023/B:REOM.0000032115.22510.b5>
- Dubois, D. and Prade, H. (2015). Possibility Theory and its Applications: Where do we stand? In Kacprzyk, J. & Pedrycz, W. (Eds.), *Springer Handbook of Computational Intelligence*, Springer-Verlag Berlin Heidelberg, 31-60. http://dx.doi.org/10.1007/978-3-662-43505-2_3
- El-Baroudy, I. and Simonovic, S.P. (2006). Application of the fuzzy performance measures to the City of London water supply system. *Can. J. Civ. Eng.*, 33(3), 255-265. <http://dx.doi.org/10.1139/05-11-3>
- Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D.P. (2010). Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-Part 1: Concepts and methodology. *Hydrol. Earth Syst. Sci.*, 14(10), 1931-1941. <http://dx.doi.org/10.5194/hess-14-1931-2010>
- Environment Canada (2014). Wateroffice: Hydrometric Station Meta Data. https://wateroffice.ec.gc.ca/station_metadata/stationList_e.html
- Freni, G. and Mannina, G. (2010). Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution. *J. Hydrol.*, 392(1), 31-39. <http://dx.doi.org/10.1016/j.jhydrol.2010.07.043>
- Fuller, W.A. (1987). *Measurement error models*, John Wiley & Sons. New York, NY, USA. <http://dx.doi.org/10.1002/9780470316665>
- Gelman, A., Carlin, J.B., Stern, H.S., et al. (2014). *Bayesian Data Analysis*, (3rd ed.). CRC Press, Boca Raton, FL, USA.
- Golder (2004). *Bow River Impact Study-phase I: Model development and application*, Golder Associates Ltd., Calgary, Alta.
- Greenberg, E. (2008). *Introduction to Bayesian econometrics*, Cambridge University Press. New York, NY, USA. <http://dx.doi.org/10.1017/CBO9781139058414>
- Guyonnet, D., Bourguin, B., Dubois, D., et al. (2003). Hybrid approach for addressing uncertainty in risk assessments. *J. Environ. Eng.*, 129(1), 68-78. [http://dx.doi.org/10.1061/\(ASCE\)E07339372\(2003\)129:1\(68\)](http://dx.doi.org/10.1061/(ASCE)E07339372(2003)129:1(68))
- Hall, M.J. (1984). *Urban Hydrology*, Essex, England: Elsevier Applied Science.
- Hauer, F.R. and Hill, W.R. (2007). Temperature, light and oxygen. In Hauer, F.R. & Lamberti, G.A. (Eds.), *Methods in stream ecology*, pp. 103-117. Academic Press, San Diego, CA, USA. <http://dx.doi.org/>

- 0.1016/B978-012332908-0.50007-3
- He, J. and Valeo, C. (2009). Comparative study of ANNs versus parametric methods in rainfall frequency analysis. *J. Hydrol. Eng.*, 14(2), 172-184. [http://dx.doi.org/10.1061/\(ASCE\)10840699\(2009\)14:2\(172\)](http://dx.doi.org/10.1061/(ASCE)10840699(2009)14:2(172))
- He, J., Chu, A., Ryan, M.C., Valeo, C., and Zaitlin, B. (2011). Abiotic influences on dissolved oxygen in a riverine environment. *Ecol. Eng.*, 37(11), 1804-1814. <http://dx.doi.org/10.1016/j.ecoleng.2011.06.022>
- Heddam, S. (2014). Generalized regression neural network-based approach for modelling hourly dissolved oxygen concentration in the Upper Klamath River, Oregon, USA. *Environ. Technol.*, 35(13), 1650-1657. <http://dx.doi.org/10.1080/09593330.2013.878396>
- Hu, Q., Huang, G.H., Cai, Y.P., and Sun, W. (2014). Planning of electric power generation systems under multiple uncertainties and constraint-violation levels. *J. Environ. Inf.*, 23(1), 55-64. <http://dx.doi.org/10.3808/jei.201400257>
- Huang, Y., Chen, X., Li, Y.P., Huang, G.H., and Liu, T. (2010). A fuzzy-based simulation method for modelling hydrological processes under uncertainty. *Hydrol. Process.*, 24(25), 3718-3732. <http://dx.doi.org/10.1002/hyp.7790>
- Kahraman, C., Beşkese, A., and Bozbura, F.T. (2006). Fuzzy regression approaches and applications. In Kahraman, C. (Ed.), *Fuzzy Applications in Industrial Engineering*, pp. 589-615. Springer Berlin Heidelberg, Berlin, Germany. <http://dx.doi.org/10.1007/3-54033517-X>
- Khan, U.T. and Valeo, C. (2014a). A new fuzzy linear regression approach for dissolved oxygen prediction. *Hydrol. Sci. J.*, 60(6), 1096-1119. <http://dx.doi.org/10.1080/02626667.2014.900558>
- Khan, U.T. and Valeo, C. (2014b). Predicting Dissolved Oxygen Concentration in Urban Watersheds: A Comparison of Fuzzy Number Based and Bayesian Data-Driven Approaches. *Proceedings of The International Conference on Marine and Freshwater Environments (iMFE 2014)* (Vol. 1, No. 1).
- Khan, U.T., Valeo, C., and He, J. (2013). Non-linear fuzzy-set based uncertainty propagation for improved DO prediction using multiple-linear regression. *Stochastic Environ. Res. Risk Assess.*, 27(3), 599-616. <http://dx.doi.org/10.1007/s00477-012-0626-5>
- Kim, K.J., Moskowitz, H., and Koksalan, M. (1996). Fuzzy versus statistical linear regression. *Eur. J. Oper. Res.*, 92(2), 417-434. [http://dx.doi.org/10.1016/0377-2217\(94\)00352-1](http://dx.doi.org/10.1016/0377-2217(94)00352-1)
- Kingston, G.B., Lambert, M.F., and Maier, H.R. (2005). Bayesian training of artificial neural networks used for water resources modeling. *Water Resour. Res.*, 41(12), 2179-2187. <http://dx.doi.org/10.1029/2005WR004152>
- Koop, G. (2003). *Bayesian Econometrics*, John Wiley & Sons. West Sussex, England, UK.
- Li, Y.P., Huang, G.H., and Chen, X. (2009). Multistage scenario-based interval-stochastic programming for planning water resources allocation. *Stochastic Environ. Res. Risk Assess.*, 23 (6), 781-792. <http://dx.doi.org/10.1007/s00477-008-0258-y>
- Li, W., Huang, G.H., Dong, C., and Liu, Y. (2013). An inexact fuzzy programming approach for power coal blending. *J. Environ. Inf.*, 21(2), 112-118. <http://dx.doi.org/10.3808/jei.201300238>
- Maddala, G.S. (1988). *Introduction to econometrics (1st ed.)*, Macmillan Publishing Company, New York, NY, USA.
- Mauris, G. (2013). A review of relationships between possibility and probability representations of uncertainty in measurement. *IEEE Trans. Instrum. Meas.*, 62(3), 622-632. <http://dx.doi.org/10.1109/TIM.2012.2218057>
- Moriasi, D.N., Arnold, J.G., van Liew, M.W., et al. (2007). Model evaluation guidelines for systematic quantification of accuracy of watershed simulations. *Trans. Am. Soc. Agric. Biol. Eng.*, 50(3), 88-90. <http://dx.doi.org/10.13031/2013.23153>
- Niemczynowicz, J. (1999). Urban hydrology and water management-present and future challenges. *Urban water*, 1(1), 1-14. [http://dx.doi.org/10.1016/S1462-0758\(99\)00009-6](http://dx.doi.org/10.1016/S1462-0758(99)00009-6)
- Ordaz, M., Singh, S.K., and Arciniega, A. (1994). Bayesian attenuation regressions: an application to Mexico City. *Geophys. J. Int.*, 117(2), 335-344. <http://dx.doi.org/10.1111/j.1365-246X.1994.tb03936.x>
- Ozbek, M.M. and Pinder, G.F. (2006). Non-probabilistic uncertainty in subsurface hydrology and its applications: an overview. *Water, Air, Soil Pollut. Focus*, 6(1-2), 35-46. <http://dx.doi.org/10.1007/s11267-005-9011-4>
- Oussalah, M. (2000). On the probability/possibility transformations: a comparative analysis. *Int. J. Gen. Syst.*, 29(5), 671-718. <http://dx.doi.org/10.1080/03081070008960969>
- Peters, G. (1994). Fuzzy linear regression with fuzzy intervals. *Fuzzy Sets Syst.*, 63(1), 45-55. [http://dx.doi.org/10.1016/0165-0114\(94\)90144-9](http://dx.doi.org/10.1016/0165-0114(94)90144-9)
- Pogue, T.R. and Anderson, C.W. (1995). *Factors controlling dissolved oxygen and pH in the upper Willamette River and major tributaries, Oregon, 1994*. In U. S. Geological Survey Water Resources Investigations Report 95-4205. Portland, OR, USA: US Geological Survey.
- Porter, D.W., Gibbs, B.P., Jones, W.F., et al. (2000). Data fusion modeling for groundwater systems. *J. Contam. Hydrol.*, 42(2), 303-335. [http://dx.doi.org/10.1016/S0169-7722\(99\)00081-9](http://dx.doi.org/10.1016/S0169-7722(99)00081-9)
- Radwan, M., Willems, P., El-Sadek, A., and Berlamont, J. (2003). Modelling of dissolved oxygen and biochemical oxygen demand in river water using a detailed and a simplified model. *Int. J. River Basin Manage.*, 1(2), 97-103. <http://dx.doi.org/10.1080/15715124.2003.9635196>
- Robinson, K.L., Valeo, C., Ryan, M.C., Chu, A., and Iwanyszyn, M. (2009). Modelling aquatic vegetation and dissolved oxygen after a flood event in the Bow River, Alberta, Canada. *Can. J. Civil Eng.*, 36(3), 492-503. <http://dx.doi.org/10.1139/L08-126>
- Shrestha, D.L. and Solomatine, D.P. (2008). Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. *Int. J. River Basin Manag.*, 6(2), 109-122. <http://dx.doi.org/10.1080/15715124.2008.9635341>
- Shrestha, R.R. and Nestmann, F. (2009). Physically based and data-driven models and propagation of input uncertainties in river flood prediction. *J. Hydrol. Eng.*, 14(12), 1309-1319. [http://dx.doi.org/10.1061/\(ASCE\)HE.1943-5584.0000123](http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000123)
- Shrestha, R.R. and Simonovic, S.P. (2010). Fuzzy set theory based methodology for the analysis of measurement uncertainties in river discharge and stage. *Can. J. Civil Eng.*, 37(3), 429-440. <http://dx.doi.org/10.1139/109-151>
- Singh, K.P., Basant, A., Malik, A., and Jain, G. (2009). Artificial neural network modelling of the river water quality-a case study. *Ecol. Model.*, 220(6), 888-895. <http://dx.doi.org/10.1016/j.ecolmodel.2009.01.004>
- Solomatine, D.P. and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *J. Hydroinf.*, 10 (1), 3-22. <http://dx.doi.org/10.2166/hydro.2008.015>
- Solomatine, D.P., Maskey, M., and Shrestha, D.L. (2008a). Instance-based learning compared to other data-driven methods in hydrological forecasting. *Hydrol. Process.*, 22(2), 275-287. <http://dx.doi.org/10.1002/hyp.6592>
- Solomatine, D.P., See, L.M., and Abraham, R.J. (2008b). Data-driven modelling: concepts, approaches and experiences. In Abraham, R.J., See, L.M., Solomatine, D.P. (Eds.), *Practical hydroinformatics*, pp. 17-30. Springer Berlin Heidelberg, Berlin, Germany. http://dx.doi.org/10.1007/978-3-540-79881-1_2
- Struve, D. and Zhou, M. (2010). Field Measurements. In Li, Y. & Migliaccio, K.W. (Eds.), *Water Quality Concepts, Sampling, and Analyses*, pp. 113-136. Taylor & Francis Group, Boca Raton, FL, USA. <http://dx.doi.org/10.1201/b10157-9>

- Suo, M.Q., Li, Y.P., Huang, G.H., Deng, D.L., and Li, Y.F. (2013). Electric power system planning under uncertainty using inexact inventory nonlinear programming method. *J. Environ. Inf.*, 22(1), 49-67. <http://dx.doi.org/10.3808/jei.201300245>
- Tan, Q., Huang, G.H., and Cai, Y.P. (2015). A fuzzy evacuation management model oriented toward the mitigation of vehicular emissions. *J. Environ. Inf.*, 25(2), 117-125. <http://dx.doi.org/10.3808/jei.201400272>
- Tetra Tech (2013). *Bow River Phosphorus Management Plan: Modelling Support for Phosphorus Reduction/Management, Task 2-Model Recalibration and Extension*, Submitted to Alberta Environment and Sustainable Resource Development.
- Thiemann, M., Trosset, M., Gupta, H., and Sorooshian, S. (2001). Bayesian recursive parameter estimation for hydrologic models. *Water Resour. Res.*, 37(10), 2521-2535. <http://dx.doi.org/10.1029/2000WR900405>
- Tyralis, H. and Koutsoyiannis, D. (2014). A Bayesian statistical model for deriving the predictive distribution of hydroclimatic variables. *Climate Dyn.*, 42(11-12), 2867-2883. <http://dx.doi.org/10.1007/s00382-013-1804-y>
- Vicens, G.J., Rodriguez-Iturbe, I., and Schaake, J.C. (1975). A Bayesian framework for the use of regional information in hydrology. *Water Resour. Res.*, 11(3), 405-414. <http://dx.doi.org/10.1029/WR011i003p00405>
- Vrugt, J.A., ter Braak, C.J.F., Gupta, H.V., and Robinson, B.A. (2009). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modelling? *Stochastic Environ. Res. Risk Assess.*, 23(7), 1011-1026. <http://dx.doi.org/10.1007/s00477-0080274-y>
- Wakefield, J. (2013). *Bayesian and frequentist regression methods*, Springer Science & Business Media, New York, NY, USA. <http://dx.doi.org/10.1007/978-1-4419-0925-1>
- Wen, X., Fang, J., Diao, M., and Zhang, C. (2013). Artificial neural network modelling of dissolved oxygen in the Heihe River, Northwestern China. *Environ. Monit. Assess.*, 185(5), 4361-4371. <http://dx.doi.org/10.1007/s10661-012-2874-8>
- West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)*, 46(3), 431-439. <http://www.jstor.org/stable/2345685>
- Xu, Y., Huang, G.H., and Xu, T.Y. (2012). Inexact management modeling for urban water supply systems. *J. Environ. Inf.*, 20(1) 34-43. <http://dx.doi.org/10.3808/jei.201200218>
- Xu, Y., Huang, G.H., Cheng, G.H., Liu, Y., and Li, Y.F. (2014). A two-stage fuzzy chance-constrained model for solid waste allocation planning. *J. Environ. Inf.*, 24(2), 101-110. <http://dx.doi.org/10.3808/jei.201400261>
- YSI Inc. (2013). YSI 5200A Multiparameter Monitor & Control: Specifications. <https://www.ysi.com/media/pdfs/W45-01-5200A.pdf>
- Zadeh, L.A. (1965). Fuzzy sets. *Inf. control*, 8(3), 338-353. [http://dx.doi.org/10.1016/S0019-9958\(65\)90241-X](http://dx.doi.org/10.1016/S0019-9958(65)90241-X)
- Zadeh, L.A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets syst.*, 1(1), 3-28. [http://dx.doi.org/10.1016/0165-0114\(78\)90029-5](http://dx.doi.org/10.1016/0165-0114(78)90029-5)
- Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*, John Wiley & Sons, New York, NY, USA.
- Zhang, K., Li, H., and Achari, G. (2009). Fuzzy-stochastic characterization of site uncertainty and variability in groundwater flow and contaminant transport through a heterogeneous aquifer. *J. Contam. Hydrol.*, 106(1), 73-82. <http://dx.doi.org/10.1016/j.jconhyd.2009.01.003>
- Zhang, K. and Achari, G. (2010). Correlations between uncertainty theories and their applications in uncertainty propagation. In Furuta, H., Frangopol, D.M., & Shinozuka, M. (Eds.), *Safety, reliability and risk of structures, infrastructures and engineering systems*, London, UK: Taylor & Francis Group, 1337-1344. <http://dx.doi.org/10.1201/9781439847657-c20>