20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Long Term Water Demand Forecasting: Use of Monte Carlo Cross Validation for the Best Model Selection

Md Mahmudul Haque [a], Khaled Haddad[a], Ataur Rahman [a], Mohammed Hossain[a], Dharma Hagare [a] and Golam Kibria [b]

[a] School of Computing, Engineering and Mathematics, University of Western Sydney, Australia
[b] Sydney Catchment Authority, Penrith, NSW, Australia
E-mail: a.rahman@uws.edu.au

Abstract: Selection and validation of any statistical models are very crucial in modelling and forecasting problems. In multiple regression analysis of forecasting long term water demand, various models are developed with a variety of predictor variables. Moreover, multiple regression models can take different forms such as linear, semi-log and log-log. In this paper, an effective but simple procedure named Monte Carlo cross validation (MCCV) is applied and compared to the most widely used leave-one-out validation (LOO) to select the best multiple regression model to forecast water demand. Unlike LOO validation, MCCV leaves out a major part of the sample during validation. Both methods are also used for estimating the prediction ability of the selected model on future samples. The advantage of MCCV is that it can reduce the risk of over fitting the model by avoiding an unnecessary large model. In this paper, MCCV and LOO are applied to the water demand data set for the Blue Mountains, NSW in Australia for single dwelling residential sector. The results show that MCCV has the ability to select an appropriate water demand forecasting model. It is also found that, MCCV assesses the prediction ability of the selected model with a higher degree of accuracy. Furthermore, the model selected by MCCV provides less uncertainty when forecasting long term water demand.

Keywords: *Multiple regression analysis, Monte Carlo cross validation, Water demand, Forecasting, Blue Mountains*

1. INTRODUCTION

Water supply and demand has become a major concern in many countries around the world due to decreasing water resources as a result of growing population, economic development and changing climate (McFarlane et al., 2012). The scarcity problem is more severe in regions experiencing reduced rainfall and increased temperature due to anthropogenic global warming. This has led to a need for better planning and designing of water supply systems, implementing system expansion, developing and managing water resources. Various water authorities have also considered alternative water resources such as rainwater tanks (Eroksuz and Rahman, 2010; Imteaz et al., 2011) and desalination plants (Ghaffour et al., 2013) as a means of enhancing water security. A vital component in planning, development and management of water supply systems including desalination plants is the accurate prediction of long term water demand (Jain et al., 2001). Accurately forecasted long term water demand is required to efficiently allocate water supplies among competing water users. Moreover, long term forecasting is helpful in assessing the effect of various conservation measures and taking suitable decisions on the development of policies and strategies for demand management.

Probabilistic approaches are the most widely used models and can be adopted to quantify the uncertainties in water demand prediction due to the stochastic nature of predictor variables. The number of predictor variables associated with this analysis are often large and can be correlated (i.e. multicollinearity), which cannot be accounted for explicitly by simple regression analysis. As such in forecasting long term water demand it still needs to be resolved (i) which set of the predictor variables are the best suited or the most optimal for inclusion in the final regression equation without over fitting the model; and (ii) which of the many candidate models is the most parsimonious one for making the most reliable prediction for the future samples, as addition of unnecessary predictor variables often leads to weaker models (e.g. producing greater uncertainty). There are so many different methodologies for model development and validation, during the last twenty years, the application of different validation methods has been widely examined in different fields of sciences such as Chemometrics (Faber and Kowalski, 1997 and Song Xu et al., 2005), Econometrics (Racine, 2000) and Hydrology (Haddad et al., 2013). In this study, a Monte Carlo cross validation (MCCV) technique is adopted for model development and validation. The analysis is carried out as follows: (1) Comparison of the MCCV with the most commonly applied leave-one-out (LOO) validation for selecting the most parsimonious regression model to be applied to estimate future samples, (2) Demonstration of the application of the MCCV method in long term water demand forecasting.

2. MATHEMATICAL FORMULATION

Let us assume a dataset of n data points with p potential predictor variables xi_1, xi_2,…, xi_p (such as rainfall, mean maximum temperature) and a response variable y_i (i = 1, 2,…, n) which can be the monthly per dwelling water use. The relationship between the response and predictor variables is often assumed to be linear. There are a few assumptions made on the data in long term water demand forecasting regression; for instance, the dataset are representative of the regression relationship to be developed and the random errors are homoscedastic. The ordinary least squares based regression (OLSR) model assumes that the quantity of interest y_i at a given point in time i can be described by a linear function of predictor variables (or a transformation there of, such as log-linear) with an additive error. In matrix notation, the model can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y} = (y_1, y_2,\ldots,y_n)^T$ is the response vector of the statistic of interest (the superscript 'T' denotes the transpose), $\mathbf{X}$ is a (n × p) matrix of predictor variables augmented by a column of ones, $\boldsymbol{\beta}$ is a (p × 1) vector of regression parameters that must be estimated and $\boldsymbol{\varepsilon}$ is an (n × 1) vector of random errors for each of the n data points used in the regression analysis, which is assumed to be normally distributed with zero mean and the covariance matrix of the form:

$$\mathbf{E}(\varepsilon\varepsilon^T)\sigma^2\mathbf{I} \tag{2}$$

where σ^2 is the model error variance and $\mathbf{I}$ is equal to the identity matrix. In the regression problem, the true $\boldsymbol{\beta}$ values (i.e. the regression coefficients) are unknown. To be able to determine the best possible model, it is necessary to decide which of the different $\boldsymbol{\beta}$s' should be included in the model. In typical ordinary stepwise regression this is equivalent to selecting the best set of predictor variables for a regression model. Considering the case above (see Eq. (1)), where a more parsimonious model may be true such that:

$$\mathbf{y} = \mathbf{X}_\phi \beta_\phi + \boldsymbol{\varepsilon} \quad (3)$$

where ϕ is a subset of {1, 2,…, p}, $\mathbf{X}_\phi$ indicates the matrix whose columns are the ones in $\mathbf{X}$ that are indexed by the integers in ϕ and $\boldsymbol{\beta}_\phi$ indicates the vector whose components are the ones in $\boldsymbol{\beta}$ that are also indexed by the integers in ϕ. Hence there are in total 2^p-1 possible different models of the form represented by Eq. (3). For the model of the form of Eq. (1), if ϕ is selected, the model is fitted based on Eq. (3):

$$\hat{\boldsymbol{\beta}}_\phi = (\mathbf{X}_\phi^T \mathbf{X}_\phi)^{-1} \mathbf{X}_\phi^T \mathbf{y} \quad (4)$$

After determining the optimal model for use in regression analysis the overall performance of the model is then evaluated according to its prediction ability, e.g. how well a model can predict future samples. In most regression applications, the mean squared error of prediction (MSEP) of a model represents its prediction ability. In practice the lower the MSEP, the better is the prediction ability of the model.

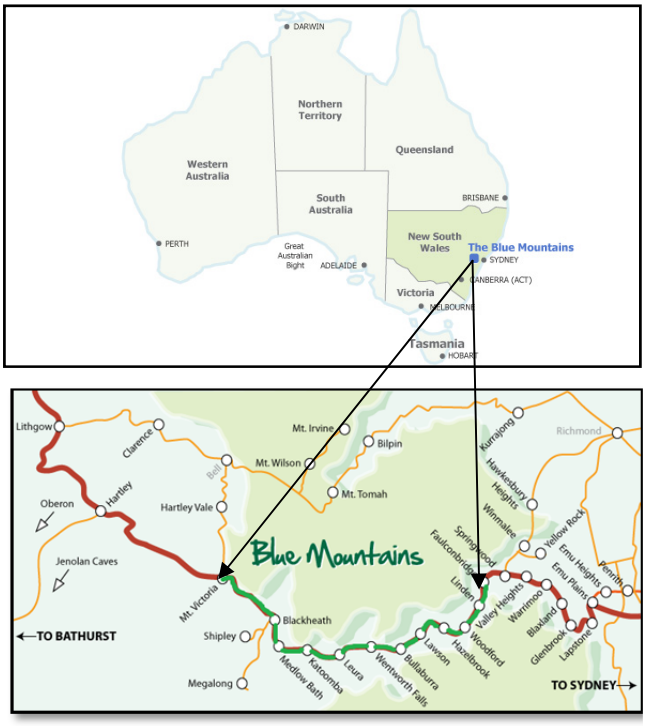

Figure 1. Blue Mountains region in Australia and Cascade and Greaves creeks water supply area (Bluemountainsaustralia.com n.d.).

2.1. Model selection by Monte Carlo Cross Validation

In general, validation attempts to select a model based on its prediction ability (Burman, 1989). For general validation, when ϕ is selected, the n data points (denoted by S) are split into two parts. The first part (calibration set), denoted by S_c (with corresponding submatrix $\mathbf{X}_{\phi Sc}$ and subvector $\mathbf{y}_{Sc}$), contains n_c data points for fitting the model. The second part (validation set), denoted by S_v (with corresponding submatrix $\mathbf{X}_{\phi Sv}$ and subvector $\mathbf{y}_{Sv}$), contains $n_v = n - n_c$ data points for validating the model. There are in total ${}^nC_{nv}$ different forms of split samples. For each of the split samples, the model is fitted by the n_c data points of the first part of S_c (Eq. (4)) to obtain $\hat{\boldsymbol{\beta}}_{\phi_c}$. The data ponts in the validation set are treated as if they are future samples. To assess the candidate models in validation the MSEP is usually used. Further details on mathematical formulation for estimating MSEP based on LOO and MCCV can be seen in Haddad et al. (2013).

3. STUDY AREA AND DATA

The Blue Mountains region (Figure 1) of New South Wales, Australia is selected as the study area. The Blue Mountains Water Supply System provides water to about 48,000 people residing between Faulconbridge and Mount Victoria, which are considered as Upper and Middle Blue Mountains area (Sydney Catchment Authority, 2009). Monthly metered water consumption data were collected from Sydney Water for the period of Jan 1997 to Sep 2011 for Cascades and Greaves Creek delivery systems. These two systems together make up the Blue Mountains Water Supply system which provides water to the twelve reservoir zones, namely, Mount Victoria, Blackheath, Catalina, Katoomba, Yosemite, Wentworth Falls, Bodington, Bullaburra, Lawson, Woodford, Linden, and Faulconbridge.

In this study, the deterministic water demand model was developed by multiple linear regression technique using a number of predictor variables (see below) to forecast per dwelling monthly water demand for the single dwelling residential sector. The regression coefficients were estimated by adopting the ordinary least squares regression approach. Modelling was done by the log-linear form of multiple regression techniques. The functional form of the log-linear model is given in eq. (5).

$$\log_{10} y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \quad (5)$$

where β_0 is the model intercept, $\beta_{1...n}$ are the regression coefficients, and p is the number of predictor variables.

Data for y (monthly per dwelling water use in kL) and X_3 (water price in AUD/kL) were collected from Sydney Water for the period of Jan 1997 to Sep 2011. The climatic data X_1 (monthly total rainfall) and X_2 (monthly maximum temperature) were collected from Sydney Catchment Authority for the study area. Data on approximate average yearly water savings for each of the water conservation programs implemented in the study area during the study period were collected from Sydney Water. These average yearly savings were converted into monthly savings by dividing it with 12. Data on the number of households that had participated in the programs were also collected in the monthly steps from Sydney Water. Then total monthly water savings from conservations programs were estimated by multiplying the average monthly savings with monthly participated household number. These monthly total savings were divided by the total number of households in that month to get the 'per dwelling saving' (X_4) from all of the conservation programs. Water restriction savings (WRS (X_5)) during drought periods (2003-2009) were calculated by deducting monthly per dwelling water conservation savings from monthly per dwelling total water savings, which can be expressed by the Eq. (6). Total per dwelling water savings were estimated by deducting observed water consumption for any month from the base water consumption of that month. In this study, the period 1997-2002 was chosen as the base consumption period as during these periods no water restriction was imposed in the study area. It should be noted that our modeling assumes that people's water use behavior does not change with time.

Table 1. Correlation between predictor variables used in this study

	X_1	X_2	X_3	X_4	X_5
X_1	1				
X_2	0.223	1			
X_3	0.109	-0.018	1		
X_4	0.074	-0.008	**0.92**	1	
X_5	0.102	-0.12	**0.74**	**0.88**	1

$$(WRS)_{ij} = (WS_T)_{ij} - (WCS)_{ij} \quad (6)$$

Where,

WRS = Per dwelling monthly water restrictions savings (kL/month/dwelling);

WCS = Per dwelling monthly water conservation savings arising from use of water-efficient appliances (kL/month/dwelling);

WS_T = Total water savings (kL/month/dwelling);

i = drought year (2003, 2004, …, 2009); and

j = month (Jan, ..., Dec).

Table 1 presents the correlation between the predictor variables. It can be seen from Table 1 that there is significant multicollinearity between (X_3, X_5), (X_3, X_4) and (X_4, X_5). The predictors X_1, X_2 and X_3 have no collinearity between them. These three variables are taken as the base case model and is referred to as model '1' in this analysis. Model 1 is then compared to three other candidate models using MCCV and LOO being:

- X_1, X_2, X_3 and X_4 (referred to as model '2' from now on);
- X_1, X_2, X_3 and X_5 (referred to as model '3' from now on); and
- X_1, X_2, X_3, X_4 and X_5 (referred to as model '4' from now on).

4. RESULTS AND DISCUSSIONS

4.1. Application to observed data

Given the 5 predictor variables and the 4 different regression models considered in this study (see above), some of the predictors may have minor effects on the estimation of monthly per dwelling water use, thus making some of the candidate models redundant. In order to select the best set of predictor variables for the regression models, LOO and MCCV in the OLSR framework were initially applied for the calibration data set (100 data points were selected randomly out of the 177 as the calibration data set). The results are listed in Tables 2a and 2b. Figure 2 illustrates the overall results of the modelling. For both the LOO and MCCV, they both tend to lean more towards the models with a greater number of predictors (i.e. model 4). One important aspect of the LOO as shown by Figure 2 is that it tends to underestimate the MSEP for all the candidate

models on the calibration data set as compared to the higher n_v. From this initial observation it is evident that LOO can over fit a selected regression model as compared to MCCV.

The optimal LOO selects 4 and 5 predictor variables (i.e. models 3 and 4). The obtained models along with some summary statistics are provided in Table 2b. In the MCCV (considering n_v = 15, 25, 35 and 50 data points during the validation and undertaking 1,000 simulations), the optimal MCCV also selects four and five predictor variables (i.e. model 3 and 4) as shown in Table 2b.

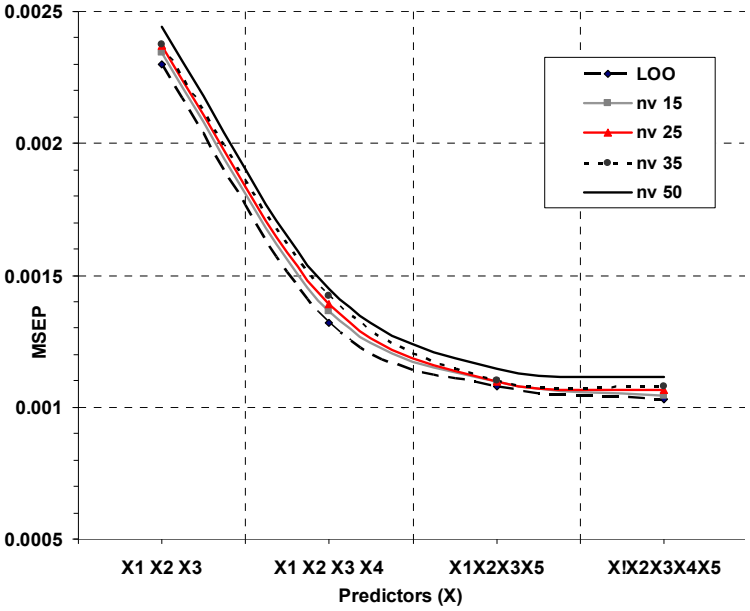

Figure 2. MSEP values associated with LOO and MCCV for different model combinations.

From a goodness-of-fit perspective, it can be seen that there is no notable difference between the models represented in Table 2b as the regression coefficients of the models are very similar. Also from the performance statistics the MSEPs' and R^2 values all fall in the similar range, except that the LOO model has the slightly higher R^2 value (72%). It can be seen from Tables 2a and 2b that the LOO and MCCV pick the same predictor variables, but both LOO and MCCV also recognise that model 3 as a potential model as the MSEPs' are quite similar. This suggests that both LOO and MCCV in this analysis are not adversely affected by multicollinearity (see Table 1). It can be seen from Table 2a that LOO gives the smaller MSEP for the calibration dataset. However, when comparing the different models from Tables 2a and 2b on the prediction data set the differences can be clearly illustrated.

Table 2a. MSEP values for calibration and validation data sets

	OLSR	MSEP on calibration set		MSEP on validation set	
n_v	Model	LOO	MCCV	Model by LOO	Model by MCCV
1	3, 4	0.00108, 0.00103		0.042, 0.041	
15	3, 4		0.0011, 0.00105		0.0415, 0.0403
25	3,4		0.0011, 0.00107		0.0416 0.0405
35	3,4		0.0011, 0.00109		0.0418, 0.0402
50	3,4		0.00114 0.00111		0.0415, 0.0403

The four regression equations (Table 2b) were finally to make prediction on the 77 validation data points. Figure 3 shows the graphical results from this validation (shown only for LOO and MCCV model 3). It is observed that the prediction performance of MCCV is slightly better than the LOO even though they both contain the same variables and a slightly smaller R^2 value. It was also found that model 3 was preferred to model 4 as model 3 showed the lower MSEP in validation over the 77 data points. Here the prediction performance is slightly better for the MCCV as compared to LOO. This shows the typical manifestation of over fitting often caused by the LOO validation approach. As such, the results may look good for the LOO calibration dataset; however, when wanting to predict future samples, the model obtained from MCCV should be used. What is noteworthy, when estimating the prediction ability of a model, the LOO on the calibration data set seems to underestimate the MSEP on the validation data set. This is illustrated in Table 2a, where the MSEP on the calibration dataset are notably smaller than that of validation dataset. This result points out that the MCCV most often will report a better measure (or more realistic value) of MSEP for the selected model as compared to LOO.

4.2. Application to forecasting

In this paper, the future water demand was estimated for the 2021-2040 time periods for the single dwelling residential sectors using the model developed in this paper i.e. MCCV model 3 and comparing it with the

model 4. Using 1 (one) future climatic scenario (A1B) and 1 (one) possible water restriction condition (Level 1), the future water demand were forecasted. Forecasted demand is presented by developing a 90% confidence band from the generated regression coefficients from the MCCV model for model 3 and model 4.

Table 2b. LOO and MCCV optimal models shown along with summary statistics

Validation	Regression Eqn.	R^2	MSEP
LOO (Model 4)	$1.14 + -9.9e^{-5}(X_1) + 0.005(X_2) + 0.016(X_3) - 0.06(X_4) - 0.024(X_5)$	70%	0.042
LOO (Model 3)	$1.17 + -8.4e^{-5}(X_1) + 0.004(X_2) - 0.026(X_3) - 0.031(X_5)$	72%	0.041
MCCV (Model 4)	$1.14 + -1.0e^{-4}(X_1) + 0.005(X_2) + 0.016(X_3) - 0.057(X_4) - 0.024(X_5)$	70%	0.0418
MCCV (Model 3)	$1.17 + -8.5e^{-5}(X_1) + 0.0044(X_2) - 0.026(X_3) - 0.030(X_5)$	71%	0.0402

The values of 5th and 95th percentiles were taken from the 1,000 simulated coefficients of the model. The 90% confidence bands may be interpreted that 90% of all the possible forecasts would fall within this band for any forecast year. The best possible (expected value) water demand for the forecasted periods is also reported here. It should be noted that the uncertainty/prediction limits reported here do not account for the model structure. Furthermore, the data length here is only 15 years, which seems to be inadequate to estimate the uncertainty with a reasonable amount of accuracy.

The 90% confidence limits of the forecasted total yearly water demand for the climatic scenario forecasted is presented in Figure 4. It can be clearly seen here that the model 4 provides the wider limits which suggests that 4 predictors is more than sufficient for forecasting future water demand. This clearly demonstrates that more predictors do not necessarily add anymore meaningful information, but on the contrary may over fit the data. As shown in this study the best model for forecasting can be obtained through proper validation techniques, lack of robust validation may give imprecise results which could undermine any meaningful long term future demand forecasting.

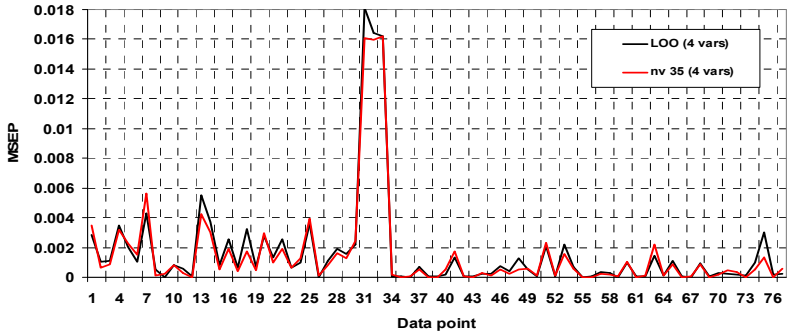

Figure 3. MSEP plot for 77 data points validation set using LOO and MCCV.

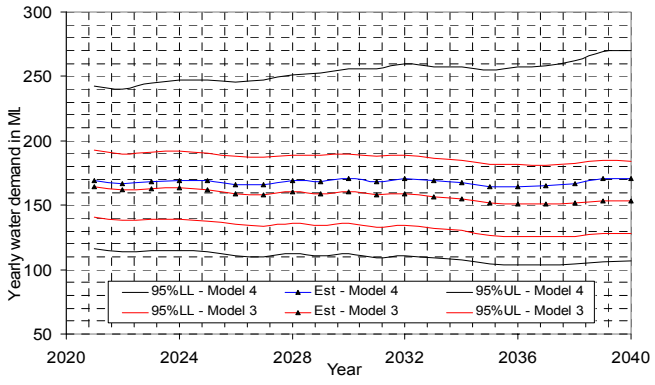

Figure 4. 90% confidence limits for total yearly water demand from 2021 to 2040 using model 4 and model 3.

5. CONCLUSIONS

Selection of the right regression model and estimation of its predictive ability are important steps in any regression analysis, this is usually undertaken by some kind of validation. This study compared the performances of the most commonly adopted leave-one-out (LOO) validation with the Monte Carlo cross validation (MCCV) procedures. This study uses observed water demand data from the Blue Mountains region in New South Wales State in Australia. It has been found that when developing long term water demand regression forecast models, application of MCCV is likely to result in more parsimonious model than the LOO case. The findings of this shed some light on the way that is usually adopted in regression analysis to estimate regression coefficients, which solely rely on the statistical significances of the regression coefficients in selecting an appropriate regression model. Finally the best model selected by MCCV was used to forecast water demand into the future while also examining the uncertainty in the water demand for an example case. Overall the study may help to provide water authorities with useful information on uncertainty estimates and an indication of future residential water demand.

ACKNOWLEDGMENTS

Water consumption data were collected from Sydney Water in 4 May 2012. The best available data at the time of study has been used, which may be updated in near future. The authors express their sincere thanks to Pei Tillman and Frank Spaninks of Sydney Water for their assistance in collating and providing the data. Further, the authors are very grateful to Lucinda Maunsell and Peter Cox of Sydney Water and Mahes Maheswaran of Sydney Catchment Authority for their cooperation and assistance during the study.

REFERENCES

Burman, P.A. (1989). A comparative study of ordinary cross validation, v-fold cross-validation and repeated learning-tested methods. *Biometrika*, 76, 503-514.

Bluemountainsaustralia.com (n.d.). Location and maps, viewed 10 February 2013, http://www.bluemts.com.au/info/about/maps/

Eroksuz, E., and Rahman, A. (2010). Rainwater tanks in multi-unit buildings: A case study for three Australian cities. *Resources, Conservation and Recycling*, 54(12), 1449-1452.

Faber, K., and Kowalski, B.R. (1997). Propagation of measurement errors for the validation of prediction obtained by principal component regression and partial least squares. *Journal of Chemometrics*, 11, 181-238.

Ghaffour, N., Missimer, T.M., and Amy, G.L. (2013). Technical review and evaluation of the economics of water desalination: Current and future challenges for better water supply sustainability. *Desalination*, 309, 197-207.

Haddad, K., Rahman, A., Zaman, M., and Shrestha, S. (2013). Applicability of Monte Carlo cross validation technique for model development and validation using generalised least squares regression. *Journal of Hydrology*, 482, 119-128.

Imteaz, M.A., Ahsan, A., Naser, J., and Rahman, A. (2011) Reliability analysis of rainwater tanks in Melbourne using daily water balance model. *Resources Conservation and Recycling*, 56, 80-86.

Jain, A., Kumar Varshney, A., and Chandra Joshi, U. (2001). Short-term water demand forecast modelling at IIT Kanpur using artificial neural networks. *Water Resources Management*, 15, 299-321.

McFarlane, D., Stone, R., Martens, S., Thomas, J., Silberstein, R., Ali, R., and Hodgson, G. (2012). Climate change impacts on water yields and demands in south-western Australia. *Journal of Hydrology*, 475, 488-498.

Racine, J. (2000). Consistent cross-validitory method for dependant data: hv-block cross validation. *Journal of Econometrics*, 99, 39-61.

Song Xu, Q., Zeng Liang, Y., and Ping Du, Y. (2005). Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics*, 18, 112-120.

Sydney Catchment Authority (2009). Blue Mountains water supply system: Strategic review. Sydney Catchment Authority, Penrith, Australia.