Efficient Selection of Inputs for Artificial Neural Network Models

Fernando, T.M.K.G., H.R. Maier, G.C. Dandy and R. May

Centre for Applied Modelling in Water Engineering (CAMWE), School of Civil and Environmental Engineering, The University of Adelaide, Adelaide, Australia, E-Mail: hmaier@civeng.adelaide.edu.au

Keywords: *Artificial neural networks; input identification; mutual information; outliers; robust statistics*

EXTENDED ABSTRACT

The selection of an appropriate subset of variables from a set of measured potential input variables for inclusion as inputs to model the system under investigation is a vital step in model development. This is particularly important in data driven techniques, such as artificial neural networks (ANNs) and fuzzy systems, as the performance of the final model is heavily dependent on the input variables used to develop the model. Selection of the best set of input variables is essential to being able to model the system under consideration reliably. When the available data set is high dimensional, it is necessary to select a subset of the potential input variables to reduce the number of free parameters in the model in order to obtain good generalization with finite data. The correct choice of model inputs is also important for improving computational efficiency. However, the topic of input selection is a difficult one. Real systems are generally complex and mostly associated with nonlinear processes. Consequently, the dependencies between output and input variables, as well as conditional dependencies between variables, are difficult to measure.

Mutual information (MI) has been used successfully to measure the dependence between output and input variables. In contrast to the linear correlation coefficient, which often forms the basis of empirical input variable selection approaches, mutual information is capable of measuring dependencies based on both linear and nonlinear relationships, making it well suited for use with complex nonlinear systems. Partial mutual information (PMI) has been proposed in recent years as a means of measuring conditional dependencies between output and input variables (Sharma, 2000). It is a robust technique for selecting input variables for multivariate, nonlinear, complex natural systems, such as hydrological processes. The PMI approach is a stepwise input variable selection algorithm. Consequently, it is necessary to have a reliable technique to indicate whether a selected candidate variable is significant or not. The original algorithm proposed by Sharma (2000) used the bootstrap method with 100 bootstraps to obtain the 95th percentile confidence limit for the PMI. However, as pointed out by Chernick (1999), about 5,000 bootstraps are needed for simple problems and about 10,000 bootstraps for more complicated problems in order to estimate the required confidence intervals reliably. Use of such a large number of bootstraps as the stopping criterion for the PMI algorithm would decrease the computational efficiency of the algorithm significantly, probably to the point of impracticality for most realistic problems.

The focus of this study is to introduce an alternative stopping criterion for PMI algorithm implementation, which is both robust and computationally efficient. As part of the proposed method, significant PMI scores are treated as outliers in the computed PMI scores. A robust outlier detection technique, the Hampel identifier (Davies and Gather, 1993), is used to evaluate the significance of selected candidate inputs. The reliability of the new technique is first investigated using two nonlinear data series where dependencies of attributes were known *a priori*. The new technique consistently selects the correct inputs, while being computationally efficient.

The modified PMI algorithm is then applied to select inputs to forecast salinity in the River Murray at Murray Bridge, South Australia, which are used to develop an ANN model. The results obtained in this study are compared with those obtained in three previous studies which developed ANN models for the same case study. The proposed PMI algorithm identifies only 11 inputs as significant from 1323 candidate inputs. The resulting ANN model has the smallest number of inputs when compared with the models developed in previous studies for this case study, with no loss in predictive performance.

1. INTRODUCTION

In recent years, there has been a significant amount of research on measuring the dependence between two random variables. In contrast, there has been very little work on multivariate or conditional measures of dependence among sets of random variables. The coefficient of linear correlation is the most widely used measure for computing dependence between bivariate data, whereas the partial linear correlation coefficient can be used to measure the conditional dependence between sets of variables. Autocorrelation is widely used to determine the number of lagged variables in modeling time series data. A shortcoming of all of the above measures is that they are based on covariance and therefore only account for linear dependence. However, most real world systems are governed by complex, nonlinear processes and as a result, the dependencies between variables are nonlinear. Mutual information possesses properties that make it well suited to measuring statistical dependence for both linear and nonlinear data.

1.1. Mutual Information

For a set of N bivariate measurements, $z_i = (x_i, y_i)$, i = 1,...,N which are assumed to be independent, identically distributed realizations of a random variable $Z = (X,Y)$, with joint probability density $f_{x,y}(x,y)$, mutual information is defined as

$$I(X,Y) = \iint dxdyf_{x,y}(x,y)\log\frac{f_{x,y}(x,y)}{f_x(x)f_y(y)} \quad (1)$$

where $f_x(x)$ and $f_y(y)$ are the marginal probability density functions of X and Y, respectively and $f_{x,y}(x,y)$ is the joint probability density function of X and Y.

1.2. Estimation of Mutual Information

Calculation of mutual information is not always easy, as it requires estimation of the marginal probability density functions (pdfs) of x and y and estimation of the joint pdf of x and y. The most widely used approach for estimating MI is based on histograms. However, this approach can be unreliable, which has resulted in the use of kernel based MI estimators (Moon et al., 1995).

When using kernel based methods, the mutual information score in (1) can be approximated as

$$I \approx \frac{1}{N}\sum_{i=1}^{N}\log_e\left[\frac{f_{x,y}(x_i,y_i)}{f_x(x_i)f_y(y_i)}\right] \quad (2)$$

where $f_x(x_i)$, $f_y(y_i)$ and $f_{x,y}(x_i, y_i)$ are the respective univariate and joint densities estimated at the sample data point (Sharma, 2000). This simplifies the integral in (1) to a summation and can be used to obtain kernel density based mutual information efficiently. The Gaussian kernel function (Scott, 1992), together with the Gaussian reference bandwidth (Scott, 1992, Silverman, 1986), are adopted in this study to estimate MI using (2).

1.3. Partial Mutual Information

The MI criterion can be used to identify dependence between bivariate data. However, as is the case with the linear correlation coefficient, it cannot be used to compute the dependence for multivariate data. To avoid this problem, Sharma (2000) proposed a new stepwise input selection algorithm, the partial mutual information (PMI). PMI provides a measure of the conditional dependence between a new candidate input and already selected inputs and the output. The PMI between the output y and the input x, for a set of already selected inputs z, is given by

$$I' = \iint f_{x',y'}(x',y')\log_e\left[\frac{f_{x',y'}(x',y')}{f_{x'}(x')f_{y'}(y')}\right]dx'dy' \quad (3)$$

where

$$x' = x - E[x\,|\,z] \;\; ; \quad y' = y - E[y\,|\,z] \quad (4)$$

where $E[\cdot]$ denotes the expectation operation. The variables x' and y' only contain the residual information in variables x and y after considering the effect of the already selected inputs z .

The discrete version of (3) can be used to approximate the sample PMI and is given as:

$$I' = \frac{1}{N}\sum_{i=1}^{N}\log_e\left[\frac{f_{x',y'}(x'_i,y'_i)}{f_{x'}(x'_i)f_{y'}(y'_i)}\right] \quad (5)$$

where x'_i and y'_i are the ith residuals in the sample data set of size N and $f_{x'}(x'_i)$, $f_{y'}(y'_i)$ and $f_{x',y'}(x'_i, y'_i)$ are the respective marginal and joint probability densities.

In this study, the general regression neural network (Specht, 1991) is used to estimate the conditional expectations in (4), as they are non-linear and only require estimation of a single parameter. This is in agreement with the approach taken by Bowden et al. (2005).

Implementation of the PMI algorithm requires a criterion to decide when to stop the addition of new candidate inputs to the already selected inputs. Sharma (2000) suggested the 95th percentile confidence limit of the sample PMI for this purpose. Sharma's original algorithm used the bootstrap method with 100 bootstraps to obtain the 95th percentile confidence limit. However, as pointed out by Chernick (1999), approximately 5,000 bootstraps are needed for simple problems and about 10,000 bootstraps for more complicated problems in order to estimate the required confidence intervals reliably. Use of such a large number of bootstraps as the stopping criterion for the PMI algorithm would decrease the computational efficiency of the algorithm significantly, probably to the point of impracticality for most realistic problems. Consequently, the focus of this study is on the development of a reliable and efficient alternative technique to the bootstrap method.

2. PROPOSED METHOD

The proposed approach is to treat significant PMI scores as outliers in the computed PMI scores. If the highest PMI score is found to be an outlier among the computed PMI scores for a particular step, then the input corresponding to the highest PMI score is a significant input. The rationale behind this approach is that if there are no outliers in the PMI scores, then all PMI scores would have the same level of significance (or in this case, no significance). This would only occur after all significant variables have been selected. Consequently, the presence of outliers indicates the presence of significant PMI scores. It should be noted that this approach does not work if the candidate inputs do not contain non-significant inputs. However, this scenario is unlikely to occur in practice, particularly when dealing with time series applications, where a number of lagged values are considered.

As part of the proposed approach, the presence of outliers in the PMI scores is detected using a robust outlier detection technique, the Hampel identifier (Davies and Gather, 1993). This is an improved, robust version of the commonly used "3σ edit rule" or "Z score" approach to outlier detection. For a normally distributed data set, the probability that Z score > 3 is only about 0.3% and is used to detect the outliers in the data set. However, this rule fails in the presence of multiple outliers due to an effect called "masking". The Hampel identifier replaces the outlier sensitive mean and standard deviation estimates with the outlier resistant median and median absolute deviation from the median (MAD), respectively. The Hampel distance, or modified Z score (MAD), for a data set $\{x_i\}$ is defined as:

$$\text{Hampel distance} = (x_i - x_{0.50}) / \text{S} \qquad (6)$$

where $x_{0.50}$ = median and S is the MAD scale estimate defined as (Pearson, 2001)

$$\text{S} = 1.4826 \text{ median } \{|\, x_i - x_{0.50} \,|\} \qquad (7)$$

The factor 1.4826 was chosen so that the expected value of S is equal to the standard deviation σ for normally distributed data.

2.1. Modified PMI Algorithm

The basic steps in the modified PMI input selection algorithm are as follows:

1. Identify the set of potential inputs that could be useful in modelling the system under investigation. Denote this input set as z_{in}. Denote the vector that will store the selected inputs as z.
2. Estimate the PMI between the output and each of the potential new inputs in z_{in}, conditional on the pre-existing input set z by using Equation (5).
3. Calculate the Hampel distance corresponding to the highest PMI score in step 2.
4. If the Hampel distance for the highest PMI value > 3, add the input variable corresponding to the highest PMI score to selected input set z and remove it from z_{in}. If the highest PMI is not an outlier, go to step 6.
5. Repeat steps 2 – 4 as many times as needed.
6. This step will be reached only when all significant inputs have been selected.

3. APPLICATION TO DATA SETS WITH KNOWN ATTRIBUTES

Two data sets with known dependence attributes were used to evaluate the reliability of the modified PMI algorithm. The two data series were the nonlinear threshold autoregressive models used by Sharma (2000), including:

TAR1 – Threshold Autoregressive order 1

$$x_t = \begin{cases} -0.9x_{t-3} + 0.1e_t & \text{if } x_{t-3} \le 0 \\ 0.4x_{t-3} + 0.1e_t & \text{if } x_{t-3} > 0 \end{cases} \qquad (8)$$

TAR2 – Threshold Autoregressive order 2

$$x_t = \begin{cases} -0.5x_{t-6} + 0.5x_{t-10} + 0.1e_t & \text{if } x_{t-6} \le 0 \\ 0.8x_{t-10} + 0.1e_t & \text{if } x_{t-6} > 0 \end{cases} \qquad (9)$$

where e_t was Gaussian random noise with zero mean and unit standard deviation for both models. One thousand and forty data points were generated for each of the models. The first 25 points were discarded and the first 15 lags were chosen as potential inputs. In addition to the modified Z score (MAD), the 95th percentile randomised sample PMI was also computed with 1000 bootstraps. The results obtained using the modified PMI algorithm applied to these data sets are shown in Table 1. It can be seen that the modified Z score (MAD) was able to correctly identify the inputs for both nonlinear threshold data sets. In addition, computation of the modified Z score (MAD) was also relatively computationally efficient. In contrast, even with 1000 bootstrap iterations, the 95th percentile randomized sample PMI did not correctly identify the inputs for the TAR2 model. The results obtained indicate that use of the modified Z score (MAD) shows promise as an efficient alternative stopping criterion for PMI algorithm implementation.

Table 1. PMI input selection algorithm results for TAR1 and TAR2

Model	Selected input	PMI	95th percentile randomised sample PMI	Z Score (MAD)
TAR1	**X_{t-3}**	**0.0662**	**0.0336**	**12.6**
	X_{t-13}	0.0318	0.0331	0.9
TAR2	**X_{t-10}**	**0.5104**	**0.0301**	**64.0**
	X_{t-6}	**0.0718**	**0.0335**	**22.5**
	X_{t-14}	*0.0330*	*0.0324*	1.7

4. CASE STUDY

The case study used to further test the effectiveness of the modified PMI algorithm was the forecasting of salinity in the River Murray at Murray Bridge, South Australia, with a lead time of 14 days. Maier and Dandy (1996) have previously developed ANN models for this case study. Bowden et al. (2002, 2005) further investigated different aspects of ANN model development using the same case study and tested the performance of the ANN models in a real-time forecasting simulation on an independent data set. Hence the River Murray salinity data provided a good benchmark for testing the performance of the modified PMI algorithm.

Maier and Dandy (1996) considered a total of 16 variables, including daily salinity, flow and river level data at different locations in the lower Murray River (Table 2) for the period 01-12-1986 to 30-06-1992 as potential model inputs. The same data set was used in this study to select model inputs using the modified PMI algorithm. In addition, more recent data for the period 01-07-1992 to 01-04-1998 were used to verify the real time forecasting capabilities of the developed model. This second validation data set was same as that used by Bowden et al. (2002, 2005).

Table 2. Available data and selected maximum lag for each variables

Location	Data type	Abbreviation	Maximum lag (days)
Murray Bridge	Salinity	MBS	74
Mannum	Salinity	MAS	81
Morgan	Salinity	MOS	104
Waikerie	Salinity	WAS	102
Loxton	Salinity	LOS	102
Lock 1 Lower	Flow	L1LF	116
Overland Corner	Flow	OCF	100
Downstream of Lock 7	Flow	L7F	103
Murray Bridge	Level	MBL	21
Mannum	Level	MAL	18
Lock 1 Lower	Level	L1LL	116
Lock 1 Upper	Level	L1UL	15
Morgan	Level	MOL	99
Waikerie	Level	WAL	69
Overland Corner	Level	OCL	101
Loxton	Level	LOL	102

5. INPUT VARIABLE SELECTION

The data set for the period 01-12-1986 to 30-06-1992 was used to obtain the input variables for the ANN model. Maier and Dandy (1996) used *a priori* knowledge of the system to select the initial time lags for each variable, whereas Bowden at al. (2005) used the first 60 lags of each variable as potential inputs. Fraser and Swinney (1986) proposed that the time lag corresponding to the first minimum of mutual information should be a better choice for the maximum time lag considered. This MI based criterion has since been used in various hydrological studies (e.g. Phoon et al., 2002). A slightly modified version of the same approach was used in this study to obtain the maximum number of lagged variables to consider as potential inputs. The MI between salinity at Murray Bridge with a 14 day lead time and each variable with different lags was computed and the optimal lag for each variable was selected, depending on the occurrence of first minimum in MI values. The selected potential lags for each variable are given in Table 2. This results in total of 1323 potential initial inputs with 2023 samples.

The results of the PMI algorithm applied to this data set are shown in Table 3. As can be seen, only 11 inputs were found to be significant. The 95th percentile confidence limits for the sample PMI obtained using 1000 bootstrap iterations were also calculated and it was found that the computed PMI values were greater than the 95th percentile confidence limit even after selecting more than 20 variables.

Table 3. PMI input selection algorithm results on the Murray River salinity data

Variable	MI/PMI	Zscore (MAD)
MAS (t-1)	1.170	4.204
WAS (t-1)	0.504	14.696
LOS (t-1)	0.147	4.629
MAS (t-2)	0.137	4.950
WAS (t-28)	0.115	3.171
L7F (t-1)	0.100	4.072
L7F (t-11)	0.139	7.027
L7F (t-2)	0.102	4.196
L7F (t-15)	0.104	3.761
MAS (t-3)	0.104	3.004
L7F (t-12)	0.098	3.116
MOS (t-9)	***0.099***	***2.836***

6. ARTIFICIAL NEURAL NETOWRK MODEL FORMULATION

A MLP neural network with a single hidden layer, trained with the backpropagation algorithm with a momentum term, was used in this study to model and predict salinity in the River Murray at Murray Bridge, South Australia, with a 14 day lead time and the input variables selected using the PMI algorithm.

The arbitrary data division method used by Maier and Dandy (1996) was adopted in this study, as this will yield a direct comparison of the results obtained with those obtained in previous studies. The data set consists of 1996 samples, after considering the lags for the selected inputs variables. The first 1597 samples (about 80% of the data) were used for calibration and 399 samples (about 20% of the data) were used for model validation. The calibration data set was further divided into 1278 training samples and 319 testing samples.

The hyperbolic tangent function was used as the activation function for both hidden and output layers. The input variables, as well as the target values, were scaled to lie in the range -0.8 to 0.8.

The number of nodes in the hidden layer affects the performance of the trained network and therefore should be optimised. In general, networks with fewer hidden nodes are preferable, as they usually have better generalisation capabilities, fewer over-fitting problems and are more computationally efficient. However, if the number of nodes is not large enough to capture the underlying behaviour of the data, the performance of the network might be impaired. In most cases, selection of the optimal number of nodes in the hidden layer is a trial and error procedure, with the help of some guidelines. A rule of thumb is that the number of samples in the training set should at least be greater than the number of synaptic weights. This gives the upper limit on the number of nodes. However, it is better to use fewer nodes in the hidden layer to avoid overfitting. In this study, a trial and error procedure for hidden node selection was used by gradually varying the number of nodes in the hidden layer from 10 to 60. The network with 20 hidden nodes gave the best results for the testing data set and was therefore selected for forecasting purposes.

The combination of a learning rate of 0.01 and a momentum term of 0.5 resulted in a smooth error reduction curve and was therefore used in this study. The synaptic weights of the networks were initialised with normally distributed random numbers in the range –1 to 1. The order in which the training samples were presented to the network was also randomised from iteration to iteration. The cross validation technique was used as the stopping criterion.

6.1. Comparison with Previous Studies

Maier and Dandy (1996) found that an ANN model with 51 inputs performed best for this case study. The optimal network obtained with 51 inputs had 30 hidden nodes. Bowden et al. (2002) used the same network structure with the same input variables, but further tested the generalisation ability of the model on a second validation set during a real-time forecasting simulation. Bowden et al. (2005) used the PMI algorithm, as well as a hybrid algorithm utilising a self organising map and a genetic algorithm coupled with a general regression neural network, (SOM-GAGRNN) to select the input variables for this case study. The PMI algorithm identified 13 inputs as significant and was implemented in two stages in that study. The SOM-GAGRNN identified 23 inputs as significant.

6.2. Results and Discussion

Time series plots of observed and predicted values of salinity at Murray Bridge with a lead time of 14-days obtained using the ANN with the 11 inputs selected using the modified PMI algorithm introduced in this paper are shown in Figures 1 and 2. It can be seen that the model performs extremely well for the training, testing and validation periods (Figure 1). Performance for the real-time forecasting simulation period was also very good, with the exception of the periods of uncharacteristic data identified by Bowden et al. (2002) (Figure 2).

The performance of the model developed in this study, as well as that of the models developed in previous studies, is shown in Table 4 in terms of root mean square error (RMSE). It should be noted that Bowden et al. (2005) used different techniques for data division, and hence direct comparison between the results for calibration and validation could not be conducted for the models with 13 and 21 inputs. It can be seen that the model with the 11 inputs identified using the modified PMI algorithm introduced in this study performs best overall, while also being the most parsimonious.

Table 4. Performance of ANN models with different inputs

	RMSE (EC units)			
ANN	Train	Test	Valid	Real-Time
51-30-1	38	52	59	86
13-32-1	-	-	-	95
21-33-1	-	-	-	113
11-20-1	36	35	50	87

7. SUMMARY AND CONCLUSIONS

In this paper, a modified version of the PMI algorithm for input identification introduced by Sharma (2000) and modified by Bowden et al. (2005) is introduced. In the proposed algorithm, use of the 95th percentile confidence level as the stopping criterion for input selection, which is generally obtained by bootstrapping, is replaced with an outlier detection statistic, the Hampel distance. The proposed stopping criterion is considered to be more robust and computationally efficient than bootstrapping. This is confirmed by the results obtained from two non-linear test functions, as well as a real-life application. However, the utility of the proposed approach requires further testing.

8. REFERENCES

Bowden, G.J., Maier, H.R. and Dandy, G.C. (2005), Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river, *Journal of Hydrology*, 301, 93- 107

Bowden, G.J., Maier, H.R. and Dandy, G.C. (2002), Optimal division of data fro neural network models in water resources applications, *Water Resources Research*, 38(2), 10.1029/2001 WR000266

Chernick, M.R. (1999), *Bootstrap Methods: A Practitioner's Guide*, Wiley, New York

Davies, L. and Gather, U. (1993), The identification of multiple outliers, *Journal of the American Statistical Association*, 88 (423), 782 - 792

Fraser, A.M. and Swinney, H.L. (1986), Independent coordinates for strange attractors for mutual information, *Physical Review A*, 33(2), 1134 – 1140

Pearson, R.K. (2001), Exploring process data, *Journal of Process Control*, 11, 179 – 194

Phoon, K.K., Islam, M.N., Liaw, C.Y. and Liong, S.Y. (2002) Practical Inverse Approach for Forecasting Nonlinear Hydrological Time Series, *Journal of Hydrological Engineering*, 7(2), 116 – 128

Maier, H.R. and Dandy, G.C. (1996), The use of artificial neural network for the prediction of water quality parameters, *Water Resources Research*, 32(4), 1013 – 1022

Moon, Y.I., Rajagopalan, B. and Lall, U. (1995), Estimation of mutual information using kernel density estimators, *Physical Review E*, 52(3), 2318 – 2321

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualisation*, Wiley, New York

Sharma, A. (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1- A strategy for system predictor identification, *Journal of Hydrology*, 239, 232 – 239

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York

Specht, D.F. (1991), A General Regression Neural Network, *IEEE Transactions on Neural Networks*, 2(6), 568 - 576

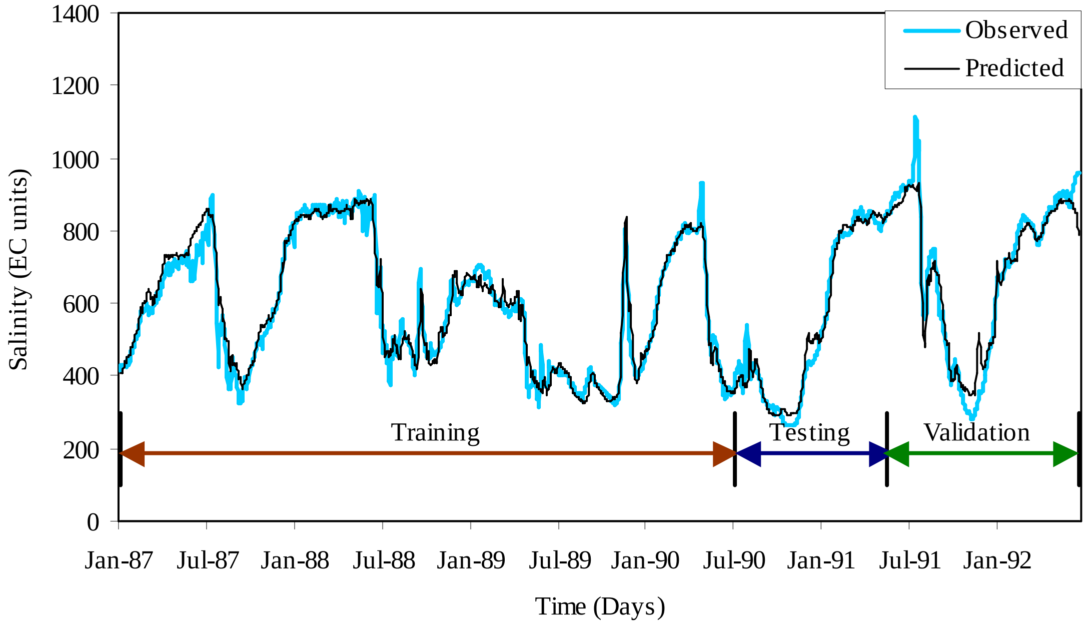

Figure 1. Observed and predicted values of salinity at Murray Bridge with 14 days lead time for calibration and validation data

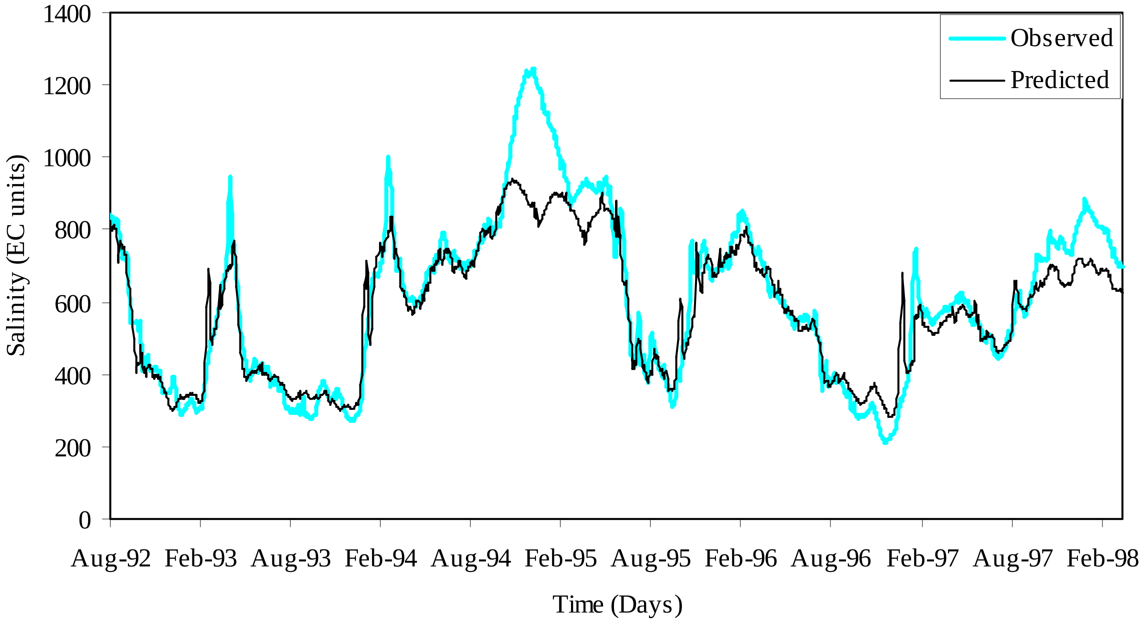

Figure 2. Observed and predicted values of salinity at Murray Bridge with 14 days lead time for real-time simulation data