



THE UNIVERSITY  
OF ADELAIDE  
AUSTRALIA



# Forecasting Water Resources Variables

## Using

# Artificial Neural Networks

By

Gavin James Bowden  
B.E. Civil/Env (Hons)

Thesis submitted for the degree of  
Doctor of Philosophy

The University of Adelaide  
School of Civil and Environmental Engineering  
Australia

February 2003

# Contents

<i>List of Figures</i> .....	<i>viii</i>
<i>List of Tables</i> .....	<i>xix</i>
<i>Abstract</i> .....	<i>xxv</i>
<i>List of Publications</i> .....	<i>xxvii</i>
<i>Statement of Originality</i> .....	<i>xxix</i>
<i>Acknowledgments</i> .....	<i>xxx</i>
<b>Chapter 1 Introduction</b> .....	<b>1</b>
<b>1.1 Reasons for this Research</b> .....	<b>1</b>
<b>1.2 Objectives and Scope</b> .....	<b>4</b>
<b>1.3 Layout and Contents of Thesis</b> .....	<b>5</b>
<b>Chapter 2 Review of Artificial Neural Networks</b> .....	<b>7</b>
<b>2.1 Background</b> .....	<b>7</b>
<b>2.2 Introduction to Artificial Neural Networks</b> .....	<b>9</b>
2.2.1 Analogy to the Brain.....	9
2.2.2 History of ANNs.....	10
<b>2.3 ANNs Compared With Statistical Models</b> .....	<b>14</b>
<b>2.4 Types of ANN Models</b> .....	<b>17</b>
2.4.1 Supervised, Feedforward ANNs.....	20
2.4.1.1 Multilayer Perceptron (MLP) .....	21
2.4.1.2 Evolutionary ANNs (EANNs).....	34
2.4.1.3 General Regression Neural Network (GRNN).....	37
2.4.2 Unsupervised ANNs.....	43

2.4.2.1 Self-Organizing Map (SOM) .....	43
<b>2.5 Important Steps in the Development of ANN Models for Time Series</b>	
<b>Forecasting .....</b>	<b>47</b>
2.5.1 Choice of Performance Criteria .....	48
2.5.2 Choice of Data Sets, Data Transformation and Preprocessing .....	50
2.5.3 Determination of Model Inputs.....	53
2.5.4 Choice of Model Architecture .....	54
2.5.4.1 Heuristic Approaches.....	56
2.5.5 Training (Optimisation) .....	57
2.5.5.1 Choice of Stopping Criteria.....	58
2.5.5.2 Internal Model Parameters.....	59
2.5.6 Validation .....	59
2.5.7 Model Deployment .....	59
<b>2.6 ANNs in Water Resources Applications .....</b>	<b>61</b>
<b>Chapter 3 ANN Methodology for Modelling Water Resources Variables.....</b>	<b>74</b>
<b>3.1 Introduction .....</b>	<b>74</b>
<b>3.2 Testing For Nonlinearity .....</b>	<b>75</b>
3.2.1 Introduction.....	75
3.2.2 Methods .....	78
3.2.2.1 The BDS test.....	79
3.2.2.2 Fuzzy Classification System (Kaboudan, 1999).....	80
3.2.3 Application of the Nonlinearity Tests to the Synthetic Test Data .....	84
3.2.3.1 BDS Test Results.....	85
3.2.3.2 Kaboudan's FCS Results.....	86
3.2.3.3 Nonlinearity Testing and ANN Modelling .....	87
3.2.4 Conclusions and Recommendations.....	90
<b>3.3 Division of Data for ANN Models .....</b>	<b>92</b>
3.3.1 Introduction.....	92
3.3.2 Methods .....	95
3.3.2.1 Data Division Using a Genetic Algorithm.....	95
3.3.2.2 Data Division Using a SOM .....	97
<b>3.4 Data Transformation.....</b>	<b>99</b>
3.4.1 Introduction.....	99
3.4.2 Methods .....	103
3.4.2.1 Linear Transformation .....	103
3.4.2.2 Logarithmic Transformation.....	103
3.4.2.3 Histogram Equalization (Looney, 1997).....	103

3.4.2.4 Kernel Transformation .....	104
3.4.2.5 Seasonal Standardisation .....	106
3.4.2.6 Transformation to Normality .....	107
<b>3.5 Determination of ANN Model Inputs .....</b>	<b>108</b>
3.5.1 Introduction .....	108
3.5.1.1 Review of Input Determination in Water Resources ANN Applications.. .....	110
3.5.2 Methods .....	115
3.5.2.1 Unsupervised Input Preprocessing .....	115
3.5.2.2 Supervised Input Determination.....	117
3.5.3 Application of the Input Determination Methods to Synthetic Data Sets	126
<b>3.6 Choice of Network Type and Architecture .....</b>	<b>133</b>
3.6.1 Introduction .....	133
3.6.2 Methods .....	135
3.6.2.1 Evolutionary Backpropagation Multilayer Perceptron (EBMLP).....	135
3.6.2.2 General Regression Neural Network (GRNN).....	139
<b>3.7 Training (Optimisation).....</b>	<b>145</b>
3.7.1 Choice of Performance Measure .....	145
3.7.1.1 Introduction .....	145
3.7.1.2 Review of Performance Measures Used in Water Resources Modelling.. .....	145
3.7.1.3 Method.....	152
3.7.2 Choice of Optimisation Methods.....	153
3.7.2.1 Introduction .....	153
3.7.2.2 Methods .....	156
<b>3.8 Model Deployment.....</b>	<b>164</b>
3.8.1 Introduction .....	164
3.8.2 Methods .....	166
3.8.2.1 Hybrid SOM-MLP Model .....	166
3.8.2.2 GRNN Outlier Detector.....	168
<b>Chapter 4 Case Study I: Salinity in the River Murray .....</b>	<b>171</b>
<b>4.1 Introduction .....</b>	<b>171</b>
<b>4.2 Background .....</b>	<b>174</b>
4.2.1 The Murray-Darling Basin and the River Murray .....	174
4.2.2 Salinity in the River Murray.....	178
4.2.3 Forecasting Salinity at Murray Bridge.....	182
<b>4.3 Available Data.....</b>	<b>184</b>

4.3.1	Salinity Data .....	185
4.3.2	Flow Data.....	196
4.3.3	River Level Data.....	199
<b>4.4</b>	<b>ANN Model Development .....</b>	<b>205</b>
4.4.1	Data Transformation.....	207
4.4.2	Model Inputs.....	207
4.4.3	Network Architecture .....	208
4.4.4	Internal Model Parameters.....	208
4.4.5	Stopping Criterion and Performance Measure.....	209
<b>4.5</b>	<b>Testing For Nonlinearity .....</b>	<b>210</b>
4.5.1	Results and Discussion .....	210
4.5.1.1	BDS Test.....	210
4.5.1.2	Kaboudan's FCS.....	210
4.5.2	Summary.....	211
<b>4.6</b>	<b>Division of Data for ANN Models .....</b>	<b>212</b>
4.6.1	Model Development .....	212
4.6.1.1	Data Division.....	212
4.6.2	Results and Discussion .....	220
<b>4.7</b>	<b>Data Transformation.....</b>	<b>226</b>
4.7.1	Model Development .....	226
4.7.1.1	Data Transformation.....	227
4.7.2	Results and Discussion .....	239
4.7.2.1	Real-time forecasting.....	253
<b>4.8</b>	<b>Determination of Model Inputs .....</b>	<b>256</b>
4.8.1	Model Development .....	256
4.8.1.1	Determination of Model Inputs.....	257
4.8.1.2	Determination of Network Architecture.....	270
4.8.2	Results and Discussion .....	271
4.8.2.1	Real-time forecasting.....	277
<b>4.9</b>	<b>Choice of Network Type and Architecture .....</b>	<b>283</b>
4.9.1	Model Development .....	283
4.9.1.1	Determination of Network Architecture .....	284
4.9.2	Results and Discussion .....	286
4.9.2.1	EBMLP .....	286
4.9.2.2	GRNN .....	289
4.9.2.3	Real-time forecasting.....	292
<b>4.10</b>	<b>Training (Optimisation).....</b>	<b>295</b>

4.10.1	Choice of Performance Measure .....	295
4.10.2	Choice of Optimisation Method .....	301
4.10.2.1	Feedforward MLP.....	301
4.10.2.2	Multiple-Sigma General Regression Neural Network.....	305
<b>4.11</b>	<b>Model Deployment.....</b>	<b>311</b>
4.11.1	Backpropagation MLP.....	311
4.11.2	GRNN.....	314
<b>4.12</b>	<b>Summary and Conclusions .....</b>	<b>319</b>
4.12.1	Testing for Nonlinearity .....	319
4.12.2	Data Division.....	319
4.12.3	Data Transformation.....	321
4.12.4	Determination of Model Inputs .....	321
4.12.5	Choice of Network Type and Architecture.....	323
4.12.6	Training (Optimisation).....	324
4.12.7	Model Deployment.....	325
<b>Chapter 5</b>	<b>Case Study II: Cyanobacteria in the River Murray .....</b>	<b>327</b>
<b>5.1</b>	<b>Introduction .....</b>	<b>327</b>
<b>5.2</b>	<b>Background.....</b>	<b>329</b>
5.2.1	Factors Affecting the Incidence of Cyanobacterial Blooms.....	330
5.2.1.1	'Bottom Up' Factors.....	331
5.2.1.2	'Top Down' Factors.....	344
5.2.2	Cyanobacterial Toxins, Their Production and Effects.....	345
5.2.2.1	Effects on Public Health.....	346
5.2.2.2	Effects on Animal Health.....	350
5.2.2.3	Ecological Effects.....	351
5.2.2.4	Effects on Recreation and Tourism .....	351
5.2.3	Toxic Cyanobacteria in the Murray-Darling Basin .....	352
5.2.3.1	History of Bloom Formation .....	353
5.2.3.2	Long Term Methods of Controlling Cyanobacteria .....	354
5.2.4	Modelling Cyanobacteria.....	356
5.2.4.1	Process-Based (Deterministic) Models .....	356
5.2.4.2	Statistically Based Models.....	357
5.2.4.3	Rule-based (Heuristic) Models.....	358
5.2.4.4	Artificial Neural Network Models.....	358
5.2.5	Forecasting <i>Anabaena</i> spp. at Morgan .....	359
<b>5.3</b>	<b>Available Data.....</b>	<b>360</b>
5.3.1	<i>Anabaena</i> spp. ....	364

5.3.2	Flow .....	370
5.3.3	River Level .....	370
5.3.4	Temperature .....	372
5.3.5	Colour .....	372
5.3.6	Turbidity .....	373
5.3.7	pH .....	374
5.3.8	Silica .....	375
5.3.9	TKN .....	376
5.3.10	Total Phosphorus .....	377
5.3.11	Soluble Phosphorus .....	379
5.3.12	Wind Run .....	379
<b>5.4</b>	<b>ANN Model Development .....</b>	<b>381</b>
<b>5.5</b>	<b>Testing For Nonlinearity .....</b>	<b>383</b>
5.5.1	Results and Discussion .....	383
5.5.1.1	BDS Test .....	383
5.5.1.2	Kaboudan's FCS .....	384
5.5.2	Summary .....	384
<b>5.6</b>	<b>Division of Data for ANN Models .....</b>	<b>385</b>
5.6.1	Results and Discussion .....	385
<b>5.7</b>	<b>Data Transformation .....</b>	<b>390</b>
5.7.1	Linear Transformation .....	395
5.7.2	Histogram Equalization .....	396
<b>5.8</b>	<b>Determination of Model Inputs .....</b>	<b>403</b>
5.8.1	Linear Transformation (Raw) Data .....	404
5.8.1.1	Stepwise PMI Algorithm .....	404
5.8.1.2	SOM-GAGRNN .....	406
5.8.2	Histogram Equalization Transformation Data .....	408
5.8.2.1	Stepwise PMI Algorithm .....	408
5.8.2.2	SOM-GAGRNN .....	411
<b>5.9</b>	<b>Choice of Network Type and Architecture .....</b>	<b>413</b>
5.9.1	Linear Transformation (Raw) Data .....	413
5.9.1.1	Stepwise PMI Algorithm .....	413
5.9.1.2	SOM-GAGRNN .....	417
5.9.2	Histogram Equalization Transformation Data .....	421
5.9.2.1	Stepwise PMI Algorithm .....	421
5.9.2.2	SOM-GAGRNN .....	424
5.9.3	Summary .....	428

<b>5.10 Training (Optimisation)</b> .....	<b>430</b>
5.10.1 Choice of Performance Measure .....	430
5.10.2 Choice of Optimisation Method .....	433
5.10.2.1 Number of strata .....	433
5.10.2.2 The minimum pheromone trail .....	434
5.10.2.3 Pheromone persistence coefficient .....	434
5.10.2.4 Number of ants in the population .....	435
5.10.2.5 Summary.....	436
<b>5.11 Model Deployment</b> .....	<b>438</b>
<b>5.12 Summary and Conclusions</b> .....	<b>448</b>
<b>Chapter 6 Conclusions and Recommendations</b> .....	<b>451</b>
<b>6.1 Contributions of the Research</b> .....	<b>451</b>
<b>6.2 General Conclusions</b> .....	<b>456</b>
<b>6.3 Specific Conclusions</b> .....	<b>460</b>
6.3.1 Case Study I: Salinity in the River Murray.....	460
6.3.2 Case Study II: Cyanobacteria in the River Murray.....	462
<b>6.4 Recommendations for Future Work</b> .....	<b>463</b>
<i>Appendix A Notation</i> .....	<b>465</b>
<i>Appendix B Abbreviations</i> .....	<b>471</b>
<i>Bibliography</i> .....	<b>475</b>



# Abstract

Artificial neural networks (ANNs) have shown themselves to be a viable alternative to many traditional water resources models, particularly in the field of forecasting hydrologic variables. However, it has been acknowledged by a number of researchers that there is currently no systematic methodology available for the development of ANN models. In this research, a methodology is formulated for the successful design and implementation of ANN models for water resources applications. Attention is paid to each of the steps that should be followed in order to develop an optimal ANN model. These steps include: determining when ANNs should be used in preference to more conventional statistical models; dividing the available data into subsets for modelling purposes; deciding on a suitable data transformation; determination of significant model inputs; choice of network type and architecture; selection of an appropriate performance measure; training (optimisation) of the network's weights; and deployment of the optimised ANN model in an operational environment.

The developed methodology is successfully applied to two water resources case studies, namely, the forecasting of salinity in the River Murray at Murray Bridge, South Australia and the forecasting of cyanobacteria (*Anabaena* spp.) in the River Murray at Morgan, South Australia. These case studies are used to test and illustrate the methods proposed for each step in the methodology. Guidelines for the use of these methods are presented.

Two state-of-the-art statistical tests for nonlinearity are developed as tools for determining when it is appropriate to proceed with an ANN model, in preference to a conventional linear time series model. Using these tests, the salinity and cyanobacteria time series data are both found to exhibit significant nonlinear serial dependence, thereby justifying the use of ANNs for these case studies.

The way in which the available data are divided into training, testing and validation sets has a significant influence on the performance of an ANN model. Therefore, two methods for dividing data into representative subsets are developed, namely, a genetic algorithm (GA) and a self-organizing map (SOM). Both data division methods are able to lead to the development of ANN models that are capable of producing consistent forecasting results.

A number of data transformations are investigated to determine their effect on ANN performance. Three new transformations are proposed, including, a transformation for removing seasonality from data, a transformation for producing normally distributed

data and a kernel transformation for producing uniformly distributed data. The results from this research reveal that the most suitable transformation is case study dependent.

Three analytical techniques are developed to determine significant model inputs for ANNs. The first technique utilises a partial measure of the mutual information criterion to characterise the dependence between a potential model input and the output variable. The remaining techniques utilise unsupervised approaches (i.e. principal component analysis (PCA) and a SOM) to reduce the input dimensionality, and a hybrid GA and general regression neural network (GRNN) to select inputs that have a significant influence on the model's forecast.

Two types of feedforward networks are considered in this research, namely, the multilayer perceptron (MLP) and the GRNN. An evolutionary method for selecting an appropriate architecture for a MLP is investigated and these models are compared with GRNN models for the salinity and cyanobacteria case studies. The GRNN is found to give superior performance when tested on data within the training domain, whereas the MLPs are more robust when uncharacteristic data are encountered. Two architectural modifications are investigated for the GRNN model and a stochastic optimisation technique inspired by the behaviour of ant colonies is developed as an approach for training GRNNs with multiple sigma weights.

A framework is developed for deploying operational ANN models in a real-world environment. Various retraining strategies are investigated and two new methods for updating ANN models in real-time are developed. These methods are able to provide significant improvement in model performance when compared to the scenario of not retraining the model.

The developed methodology results in ANN models that are successful in providing salinity forecasts, 14 days in advance. ANN models are also successfully developed for predicting the timing and relative magnitude of significant growth events of *Anabaena* spp., 4 weeks in advance, provided that representative data are available in the calibration set. For the cyanobacteria case study, the causal variables are found to interact with concentrations of *Anabaena* in a very complex manner and the most important input variables are temperature, flow, pH and previous values of *Anabaena*.