Chapter 7
Conclusions and Recommendations

7.1 Contributions of the Research

The methods presented in this thesis were divided into deterministic approaches (based on a single weight vector) and Bayesian approaches (based on a posterior distribution of weight vectors). The contributions to both deterministic and Bayesian ANN development made throughout this thesis are as follows:

7.1.1 Deterministic ANN Development

1. A step-by-step approach for the successful development and implementation of deterministic ANNs in the field of water resources modelling has been recommended, based on a review of the current state-of-the-art ANN development methods and on the results of the investigations undertaken using synthetic data sets. This includes recommended methods for:

   (a) dividing the data into training, testing and validation subsets;

   (b) preprocessing the data, given the assumption of independently, normally distributed errors with zero mean and constant variance;

   (c) determining ANN inputs;

   (d) selecting the optimum level of complexity of an ANN;

   (e) training an ANN to find the single best estimate of the weights; and

   (f) validating an ANN based on model fit and on how well the underlying data-generating function has been estimated.

2. The performance of the SCE-UA algorithm has been assessed as a method for training ANNs. This global search algorithm has not been used previously for this purpose, but has found success in optimising conceptual hydrological models. The algorithm was assessed on numerous ANNs of varying complexity when applied to 3 synthetic and 2 real-world data sets of varying length, nonlinearity, complexity, and levels of noise. Upper limits for the algorithm parameters have also been specified in order to reduce training times experienced with this algorithm.

3. The optimisation abilities of the backpropagation algorithm, a GA and the SCE-UA algorithm have been compared, in order to recommend the most suitable training algorithm for ANNs. The training performances of the algorithms were compared under a range of different model specification conditions (e.g. underparameterised, overparameterised) when applied to synthetic data sets with different nonlinearity and noise characteristics. For each of the models developed in the investigations, the algorithms were initialised with five different sets of weights in order to assess their robustness, given different initial conditions. It has been found that the SCE-UA algorithm is the most suitable training method for consistently obtaining good solutions for ANNs; however, given the time required for training with this algorithm, it has been recommended that backpropagation be used to obtain the weights needed to initialise the Bayesian training approach.

4. A variety of in-sample and out-of-sample model selection criteria have been compared and the in-sample BIC (calculated based on weights obtained when overfitting is prevented) and out-of-sample AIC criteria have been recommended for this purpose. The criteria were assessed for their ability to measure an ANN's generalisability when used in a trial-and-error model selection procedure. Given preliminary results obtained when the criteria were compared using synthetic data sets, a further assessment of the in-sample BIC criterion was carried out on a complex real-world case study. The in-sample BIC values obtained when the models were trained until convergence were compared to those obtained when training was stopped early to prevent overfitting, in order to investigate whether this criterion could adequately penalise model complexity and select the most appropriate model structure based on in-sample performance alone, without the use of a test data set. It has been demonstrated that prevention of overfitting is necessary when using the in-sample BIC criterion for model selection.

5. Four methods for quantifying the relative importance of ANN inputs, based on the

weights of a trained ANN, have been compared in order to recommend the best approach for interpreting the modelled relationship. Two of these approaches were new methods introduced in this research, based on modifications of the other two existing methods. The input importance measures were compared in terms of their accuracy in approximating the order and relative magnitude of input importance when applied to three synthetic data sets. It was not possible to determine the exact input contributions for two of these data sets directly; therefore, the estimates of relative input importance obtained using the various measures were compared to relative importance estimates obtained using the PMI criterion. This criterion has been recommended as a suitable model-free method for quantifying the relative importance of input variables when there is little, or no, *a priori* information. It is acknowledged, however, that this criterion also provides an approximation of input importance and is subject to errors. The modified Connection Weight Approach introduced in this thesis has been recommended as the most appropriate method for assessing the relationship modelled by an ANN using the optimised weights.

7.1.2 Bayesian ANN Development

1. The primary contribution of this research has been the development of a new Bayesian framework for ANNs, incorporating “Bayesian training and prediction” and “Bayesian model selection” components. This framework enables the uncertainty in ANN weight estimates to be explicitly accounted for, which in turn, enables the generation of probabilistic predictions. It also provides an objective method for selecting the optimal complexity of an ANN, which helps to prevent the use of overparameterised models, and in turn, results in improved generalisability and interpretability. The framework incorporates methods selected for their ability to produce accurate results, while maintaining minimal complexity. The simplicity of the framework was considered to be of utmost importance for the adoption of the framework in the field of ANN modelling of water resources.

2. A two-step iterative MCMC training procedure has been recommended, where the weights are sampled using the AM algorithm and the hyperparameters are updated using the Gibbs sampler. This results in a simple, yet relatively efficient, Bayesian training algorithm. The probabilistic predictions generated can be used to evaluate mean (expected) predictions and construct prediction intervals, which account for the uncertainty in the predictions, while other probabilistic outputs from the algorithm can be used to provide probabilistic estimates of model assessment criteria.

For example, it has been demonstrated how the posterior weight distribution can be used to express measures of relative input importance, obtained using the modified Connection Weight Approach, as probability distributions, which quantify the uncertainty in these estimates. Furthermore, the log likelihood values resulting from each of the generated weight vectors can be used to evaluate the -1/2BIC distribution, which can then be used for Bayesian model selection.

3. Three different forms of prior distribution, resulting in different levels of algorithm complexity, have been compared in terms of their ability to prevent overfitting and escape poor local modes on two synthetic data sets. The effects of simulated annealing and fixing the hyperparameters for a short initial period have also been investigated in terms of increasing the MCMC algorithm's ability to escape local modes. The hierarchical prior distribution with initially fixed hyperparameters has been recommended.

4. A trial-and-error model selection approach has also been developed, which involves the use of the -1/2BIC evidence estimator to evaluate Bayes factors and rank competing models in order of posterior probability. The procedure also involves the inspection of marginal posterior hidden-output weight distributions of the highest ranked models to check the Bayes factor results and determine the exact number of hidden nodes necessary. The $-1/2\mathrm{BIC}$ evidence estimator was recommended for ranking competing models based on a comparison of this estimator with the G-D and C-J evidence estimators.

5. The proposed deterministic and Bayesian ANN development approaches have been compared in their abilities to develop useful forecasting models when applied to two real-world problems, including:

   (a) forecasting salinity concentrations in the River Murray at Murray Bridge, 14 days in advance, for which it was known that the models would be required to extrapolate; and

   (b) forecasting cyanobacteria (*Anabaena* spp.) concentrations in the River Murray at Morgan, 4 weeks in advance, where an additional aim was to determine whether it was possible to develop a model that could be used for hypothesis testing of management scenarios.

7.2 Conclusions

7.2.1 General

The overall objective of this thesis, as stated in the introduction (Chapter 1), was to

> ...use Bayesian methods to help overcome some of the limitations that prevent ANNs from becoming more widely accepted and reaching their full potential as reliable water resources models, namely the lack of consideration of prediction uncertainty, the difficulty in estimating appropriate parameter values, the difficulty in selecting the optimum complexity and the lack of an objective method to properly validate the model and interpret the relationship modelled.

Through the research presented, it was shown that the Bayesian framework developed in this thesis could be used to effectively address these issues. In comparison to the state-of-the-art ANN development approach recommended, the Bayesian development framework resulted in predictive models with significantly greater usability than the best deterministic models developed for the same purpose, when applied to both synthetic and real-world data sets.

Using the proposed Bayesian framework, the mean performance the ANN models developed was found to be similar to, if not better than, the performance of the ANN models developed using the deterministic approach in an interpolative context. For the models developed on the synthetic data sets, the mean Bayesian predictions always provided a better fit to the “true” data, indicating that the underlying data-generating functions had been approximated more accurately using Bayesian methods, while, in each case, the 95% prediction limits accounted for at least 95% of the “measured” data. The importance of accounting for the uncertainty associated with ANN predictions from a water resources management point of view was highlighted in the real-world salinity and cyanobacteria forecasting case studies. By accounting for the entire range of plausible weight vectors in the salinity case study, it was found that the mean performance of the Bayesian ANN developed was a significant improvement over the performance of the deterministic ANN in a real-time forecasting scenario, indicating that the model was more robust to the presence of uncharacteristic data (data outside the range of those used for training). However, of greater importance was the generation of prediction limits, which indicated the quality of the forecasts and signified the increased level of uncertainty in the forecasts when the model was required to extrapolate. For the cyanobacteria forecasting case study, both the deterministic and Bayesian ANNs performed poorly due to the poor quality of the available data and the possible exclusion of important input variables. Neither the deterministic forecasts, nor the mean Bayesian forecasts, were able to predict the occurrence of significant growth events, at which point management actions need to be taken. However, the majority of significant growth events were accounted for within the 95% prediction limits generated using the Bayesian ANN and, as a result, the usability of this model was considered to be much greater than that of the deterministic model, purely due to the provision of these confidence bounds.

Overall, the main advantages of the proposed Bayesian framework over the deterministic ANN development approach were found to be:

1. the explicit handling of uncertainty in both the weights and the predictions, which enabled the generation of prediction limits;

2. the ability to objectively compare ANNs of varying complexity in order to select the optimal number of hidden nodes required;

3. the automatic incorporation of weight regularisation, which prevents overfitting;

4. the ability to escape local modes and provide a more complete picture of the solution surface;

5. the lack of need for a testing data set, which maximises the data and, hence, information available for training (although this was not taken advantage of in this research as a fair comparison with deterministic methods was required); and

6. the ability to evaluate probabilistic measures with which the ANN models may be assessed (e.g. RI and -1/2BIC distributions).

On the other hand, a major challenge facing the application of any MCMC ANN training technique is that, due to the complexity of ANNs and the strong correlations between the weights, it is difficult to effectively and efficiently explore the weight space and achieve convergence to the posterior distribution within a reasonable time frame. In trying to maintain the simplicity of the Bayesian training approach developed in this thesis, limited focus was placed on achieving optimal efficiency of the MCMC algorithm. Therefore, the proposed approach may be more computationally intensive than the complex MCMC algorithms previously developed for ANN training, requiring a larger number of iterations to converge. Nevertheless, due to the increasing power of modern computers, the efficiency of MCMC algorithms is becoming less of a concern than it once was. Furthermore, while emphasis was placed on assessing convergence of the MCMC training algorithm in this research, it is acknowledged that true convergence can never be guaranteed for multimodal problems like ANNs, as there may still be undiscovered modes which cannot be diagnosed. While this may lead to sub-optimal results, statistical optimality has never been the main concern of ANN practitioners and, as the results presented in this thesis have demonstrated, it is still better to approximate what may be a ‘local’ posterior distribution around a ‘good’ mode (high likelihood) than to rely on a single set of deterministic weight estimates.

7.2.2 Deterministic ANN Development

When used for ANN training, the SCE-UA algorithm is more robust to different initial conditions than backpropagation or a GA, particularly on smaller networks with more complicated error surfaces. However, a major limitation of this algorithm is the time required for training. Even with modifications to the SCE-UA parameters, the resulting training times can still be excessively long when a moderate number of inputs or hidden nodes are required for a given problem. If the best deterministic results are required and time is not an important consideration, then this algorithm shows promise as an ANN training method. However, to obtain the initial weights for the MCMC algorithm, a simpler, gradient-based algorithm, such as backpropagation, is considered more suitable. Nevertheless, the best estimate of the global minimum SSE value should be sought by initialising the algorithm with several different sets of random weights, in order to maximise the performance of the MCMC algorithm.

Both the in-sample BIC and out-of-sample AIC criteria are suitable for ANN model selection in a trial-and-error procedure. However, in complex real-world problems with noisy data, it is important to ensure that the models are not overtrained, in order to correctly select the optimal ANN structure using the in-sample BIC criterion. Therefore, unless overfitting can be prevented through weight regularisation or the use of long training data sets, it is likely that a test data set will still be required to stop training before overfitting occurs. When assessing generalisability using out-of-sample performance, it is apparently beneficial to penalise model complexity to some extent, as this reduces the sensitivity of a criterion to factors such as the testing data used or the weights obtained during training. It is apparent that this is why the AIC criterion was found to be the best out-of-sample measure of generalisability.

The modified Connection Weight Approach, which accounts for squashing of the weights by nonlinear hidden layer activation functions, is more accurate in terms of approximating the order and relative magnitude of ANN input importance than the currently available Garson's method and original Connection Weight Approach. It is acknowledged that that this method is unable to perfectly summarise the information contained in the weights of a trained ANN and there may be significant variation in the magnitudes of the estimated RI values, depending on the weights obtained during training. However, it is concluded that this method currently provides the best interpretation of the function modelled by an ANN and the RI values estimated by this approach may be used to validate the model in terms of its physical plausibility, and once validated, may be used to gain additional information about the physical system.

7.2.3 Bayesian ANN Development

The best results can be obtained with the MCMC training algorithm when a hierarchical prior distribution is assumed and the hyperparameters of the prior and likelihood are fixed for a short initial period, such that the likelihood function is flattened slightly and the MCMC chains can move more freely around the search space during this period. This configuration of the MCMC algorithm reduces the bias introduced by the initial weights and prevents, or corrects, overfitting, while maintaining the algorithm's relative simplicity. However, a good weight initialisation is still of utmost importance in achieving optimal results with the algorithm, which means that the best estimate of the maximum likelihood weights, given a rigorous search of the weight space, should be used to initialise the algorithm. This provides confidence that the algorithm will converge to at least a 'good' local mode of the posterior distribution, even if it cannot always be guaranteed that the algorithm will converge to the true posterior. Simulated annealing has been suggested in the past as a method to improve convergence to the true posterior. However, based on the investigations carried out in this research, it is considered that this method increases the complexity of the MCMC algorithm without significantly improving the results.

The mean -1/2BIC evidence estimator is the most consistent estimator for comparing the evidence values of competing ANNs in order to select the optimal level of complexity. In comparison to the G-D and C-J estimators, the -1/2BIC estimator is the least sensitive to inexact convergence of the MCMC algorithm, as it depends only on the log likelihood values generated, and not on the prior and proposal distributions, as the G-D and C-J estimators do. However, while the mean -1/2BIC estimator provides a good initial guide for selecting the appropriate number of hidden nodes in an ANN, this is not an entirely accurate estimate of a model's evidence and is known to contain an error that does not decrease with data set size. Therefore, the final step in the proposed BMS framework, involving inspection of marginal posterior hidden-output weight distributions, is extremely important in selecting the appropriate model. The fact that this check is available is one of the greatest advantages of the proposed approach over deterministic model selection methods, which do not provide such a test and rely on a single estimate of a given selection criterion, which may be sensitive to the weights obtained during training.

By applying the modified Connection Weight Approach in a probabilistic context, using the posterior weight distribution, more information is revealed about the relationship modelled by a given ANN. Significant inputs, and those of relatively minor importance, become more obvious to the modeller, as the distance that the RI distributions are from zero gives an indication of the relative strengths of the input-output relationships. Furthermore, the degree of variation in the RI distributions can be used to gauge how much confidence should be placed in the information gained from the weights.

7.3 Recommendations for Future Work

1. The results presented in this thesis show the proposed Bayesian framework to be a superior ANN development approach to best practice deterministic approaches. However, it is recommended that the advantages of this framework be further assessed under a range of different conditions; for example, different amounts of data, different signal-to-noise ratios and/or various forecasting horizons.

2. The modified Connection Weight Approach introduced and recommended in this thesis for assessing the relative importance of ANN inputs was only evaluated using the hyperbolic tangent activation function on the hidden layer nodes. It would, therefore, be worthwhile to assess whether or not this approach is accurate if it is extended to include a range of different hidden layer activations, and possibly, ANNs with multiple hidden layers.

3. The results of the cyanobacteria forecasting case study presented in this thesis were limited by the quality and quantity of the available data. It is recommended that the deterministic and Bayesian ANN development approaches be applied to a water resources case study where the data are not limiting, in order to determine whether or not it is possible to develop an ANN model using these methods that can be used successfully for hypothesis testing and to assess the relative advantages, if any, of the Bayesian model when used in this context.

4. An advantage of the Bayesian training approach not investigated in this research arises due to the sequential nature of Bayes’ theorem, which allows information

about a set of parameters to be updated as more observations are taken. Therefore, in terms of ANN training, the posterior weight distribution may be updated once a second set of data $\mathbf{y}_2$ has been observed by using the posterior distribution estimated with the first set of data $\mathbf{y}_1$ as the prior distribution and using information contained in $\mathbf{y}_2$ to update this prior, as follows:

$$p(\mathbf{w}|\mathbf{y}_2, \mathbf{y}_1) \propto p(\mathbf{w}|\mathbf{y}_1)L(\mathbf{w}|\mathbf{y}_2) \tag{7.1}$$

Consequently, it is considered worthwhile to investigate the best method for taking advantage of this property, which would primarily involve finding the most suitable method for describing the multidimensional posterior weight distribution approximated using the first set of data.

5. Rather than using evidence estimates for selecting a single "optimal" model structure, these values may be used to combine the predictions of several different models, in order to account for uncertainty in the predictions as a result of the uncertainty in the model structure. To do this, a weighted average of the models' results may be evaluated, where the evidence estimates are used to weight the predictions of different models. This is another advantage of the Bayesian modelling paradigm that was not investigated in this research. However, it is considered that extending the framework in the future to include this source of uncertainty is something that should be investigated. Under this approach, not only should ANN models with different numbers of hidden nodes be considered, but also models with different sets of inputs, which would account for the uncertainty in the predictions associated with using a defined set of predictor variables.

6. In this research, the adoption of a Gaussian likelihood function was based on standard Bayesian regression assumptions, where uncertainty in the input variables is ignored and response error and model error are lumped into a single white noise term. However, as stated in Section 4.2.3.1 of this thesis, the Gaussian residual model may be inappropriate for many practical problems. Recently, *Kavetski et al.* (2002) introduced a Bayesian total error analysis (BATEA) approach, where a more realistic error model is specified to enable correct representation of the way errors enter and propagate through environmental models. It is recommended that the likelihood function used in the proposed Bayesian training approach be reconsidered in the future to incorporate any uncertainty in the input variables, similar to the BATEA approach of *Kavetski et al.* (2002).