Chapter 6
Case Study 2 - Cyanobacteria Forecasting

6.1 INTRODUCTION

The case study considered in this chapter is that of forecasting concentrations of the cyanobacterium *Anabaena* spp. in the River Murray at Morgan, South Australia, 4 weeks in advance. Similar to the salinity problem considered in Chapter 5, cyanobacterial blooms are a major water quality problem in the lower River Murray and, like the salinity case study, there has been substantial research conducted previously into developing ANNs for modelling this problem (*Maier*, 1995; *Maier and Dandy*, 1997; *Maier et al.*, 1998, 2000, 2001; *Bowden*, 2003). However, it was noted by *Bowden* (2003) that the data available for this case study possess a high degree of uncertainty, resulting from sampling and counting errors, and consequently, the predictive performances of the models developed have been limited. In this chapter, ANN models are developed using both the state-of-the-art deterministic and Bayesian approaches proposed in Chapters 3 and 4, respectively, and the results are compared to determine whether a superior predictive model can be developed using the Bayesian ANN framework in the face of this data uncertainty. A secondary aim of this case study is to determine whether or not it is possible, given the poor quality of data, to develop an ANN model that is representative of the underlying physical mechanisms that drive the development of cyanobacterial blooms, such that the model can be used for hypothesis testing of different management scenarios.

6.2 BACKGROUND

6.2.1 Cyanobacteria in the River Murray

Cyanobacteria, which are more commonly known as blue-green algae, are naturally occurring in aquatic environments. In a balanced and healthy river system, they provide a major sink for carbon, nitrogen and phosphorus and produce much of the world's atmospheric oxygen (*Reynolds*, 1984). However, when the natural balance of the system is upset, the production of cyanobacteria can become excessive, resulting in toxic blooms that have a destructive effect on the water body and inhibit the use of the water for irrigation, drinking, livestock and recreational purposes. Decaying cyanobacteria cells deplete the water body of oxygen, which causes stress to, or even death of, other aquatic organisms, while the effects of the toxins produced by cyanobacteria on human and animal health can range from skin and eye irritations to liver damage, tumour promotion and death (*Crabb*, 1997). Furthermore, water supply operations may be disrupted, as large numbers of cyanobacteria cells increase the suspended solids load in the water, thereby blocking filters and reducing the effectiveness of disinfection.

The problem of cyanobacterial blooms is particularly significant in Australia due to the generally arid climate, combined with the current land and water management practices that create conditions in which the cyanobacteria thrive. While such blooms are not a new occurrence in Australian rivers, there has been increased attention on the effects of cyanobacteria and the management of this problem since the wide publicity of the 1991 Darling River bloom, which was the largest cyanobacterial bloom ever recorded anywhere in the world (extending for over 1000 km) (*Blue-Green Algae Task Force*, 1992). This bloom resulted in the death of an estimated 10,000 livestock, required emergency water supplies for a number of towns and, consequently, caused the New South Wales government to declare a state of emergency (*MDBMC*, 1994). Although not as extensive as the Darling River bloom, significant cyanobacterial blooms have also occurred periodically in the lower River Murray. Since European settlement, the River Murray has been heavily regulated, primarily to aid navigation and facilitate diversion of water. The number of weirs in the lower River Murray downstream of the South Australian border (Figure 6.1) is such that this section of the river is now a series of continuous weir pools and does not have the characteristics of a flowing river except under high flows (*Thoms et al.*, 2000). In summer, these weir pools often stratify creating conditions conducive to the development of cyanobacterial blooms. This poses a significant threat to water supply in South Australia since, as discussed in Section 5.2, the River Murray is essential for water supply purposes in South Australia and a number of water supply offtakes are located

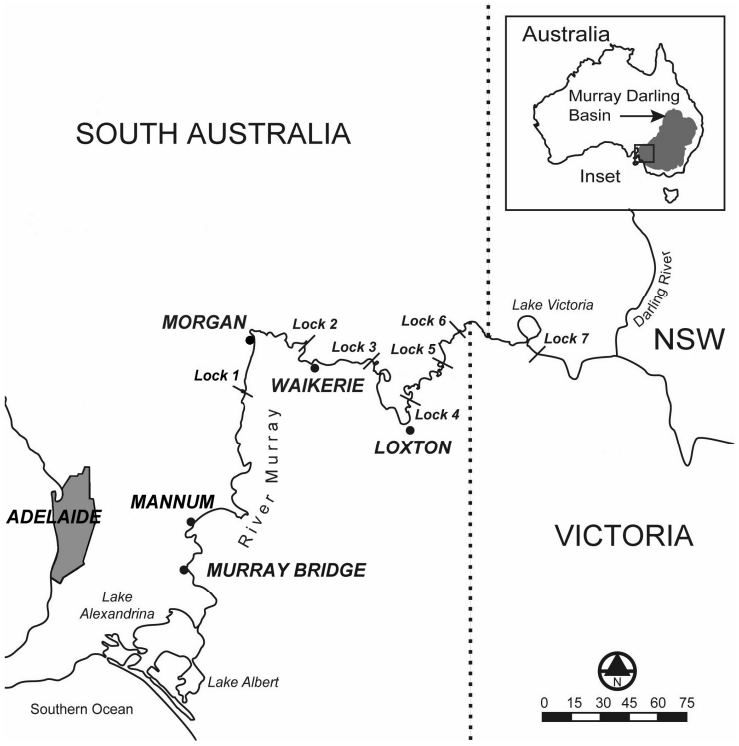

Figure 6.1 Regulatory structures in the lower River Murray, South Australia.

in this stretch of the river (see Figure 5.2).

Several water authorities throughout Australia have adopted an Alert Levels framework for managerial responses to outbreaks of cyanobacteria in drinking water supplies (*Jones et al.*, 2003). This framework is summarised in Table 6.1, where it can be seen that advanced treatment processes (in addition to conventional treatment) are required for cell concentrations exceeding 2,000 cells/mL, and concentrations above 15,000 cells/mL may prohibit the use of the water for supply purposes. Advanced treatment processes, such as oxidation or activated carbon filtration, can be very expensive (*MDBMC*, 1994); therefore, management strategies that reduce the frequency and intensity of cyanobacterial blooms may be a preferred option. However, in order to successfully implement preventative management strategies, a better understanding of the factors that trigger cyanobacterial blooms is required. Excessive cyanobacterial growth relies on a combination of calm and stable water conditions, nutrient enrichment (in particular nitrogen and phosphorus), warm temperatures and poor light attenuation. However, while a basic understanding of the links between these factors and cyanobacterial growth is well documented, it is extremely difficult to attribute cyanobacterial growth to any specific set of factors, due to the continual and complex interactions occurring between environmental variables (*MDBMC*,

Table 6.1 National cyanobacterial Alert Level framework for managerial response (Source (*Jones et al.*, 2003)).

NOTE: This table is included on page 258 of the print copy of the thesis held in the University of Adelaide Library.

1994), as illustrated in Figure 6.2. This is accentuated in rivers where the effects of flow alter each of the above mentioned environmental conditions that are conducive to cyanobacterial growth. In fact, flow is the only variable that is consistently (inversely) correlated with concentrations of cyanobacteria (*MDBMC*, 1994).

6.2.2 Modelling Cyanobacteria Concentrations with ANNs

As it is either impossible or infeasible to empirically investigate the direct response of a river system to changes in environmental variables, predictive models can be useful for determining which factors have an inhibitory effect on cyanobacterial bloom development and how best to exploit these factors, as a means of preventative control. Furthermore, as cyanobacterial bloom control is only effective if the management strategy is applied preventatively or in the very early stages of bloom development, a predictive model can be useful for providing advanced warnings of bloom occurrences (*Recknagel*, 1997). In the 1990s, *French and Recknagel* (1994); *Recknagel* (1997) and *Recknagel et al.* (1997) showed that, unlike a number of process-based and traditional statistical approaches, ANNs were a promising tool for the predicting the timing and magnitude of the incidence of cyanobacteria.

The first ANN models for predicting cyanobacteria concentrations in the lower River Murray were developed by *Maier* (1995); *Maier and Dandy* (1997) and *Maier et al.* (1998). These models were developed to provide 2- and 4-week forecasts of a species

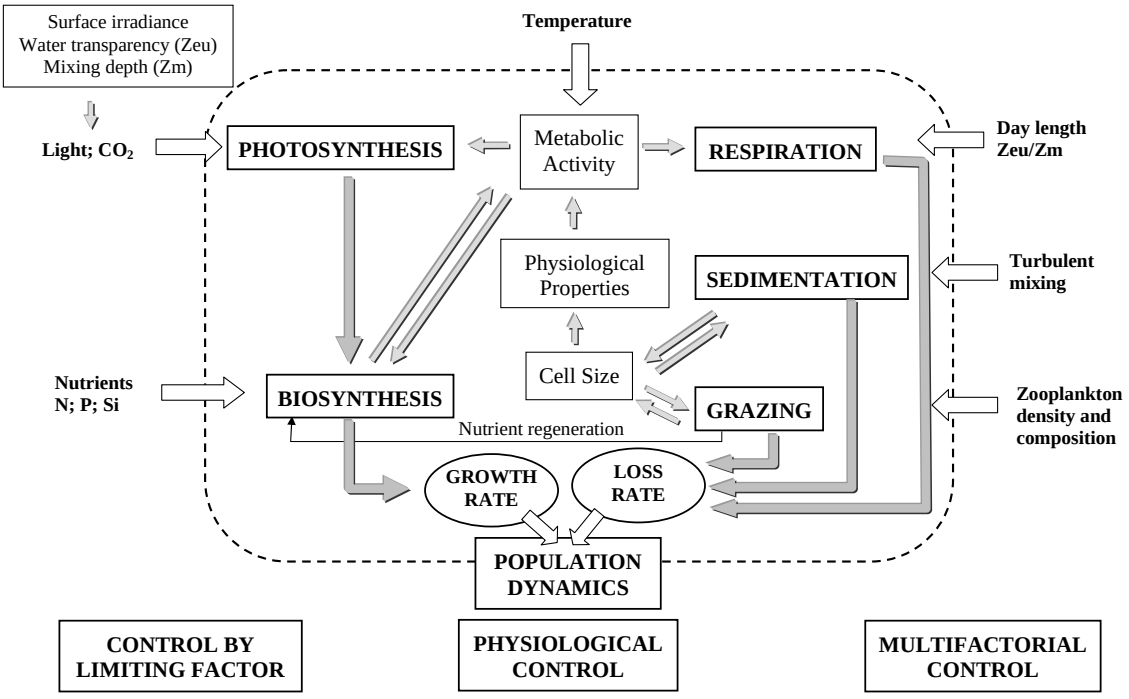

Figure 6.2 Algal dynamics (adapted from *Capblancq and Catalan* (1994)).

group of the cyanobacterium *Anabaena* spp. at Morgan, South Australia, since *Anabaena* are the most prolific species of cyanobacteria in the lower River Murray and a major water supply offtake and water filtration plant are located at Morgan (see Figure 5.2). As well as providing forecasts of the onset, peak and duration of blooms of *Anabaena*, one of the main aims of the study conducted by *Maier et al.* (1998) was to investigate the strengths of the relationships between the model inputs and outputs, by means of sensitivity analyses, in order to establish the environmental factors which result in high incidences of cyanobacteria. It was found that flow, temperature and colour (representing light availability) were the most important input variables for modelling the incidence of cyanobacteria at Morgan. Similar studies were conducted by *Maier et al.* (2000, 2001), where a neurofuzzy approach was used to forecast cyanobacteria concentrations. The input-output relationships modelled by the ANNs developed in these studies could be interpreted by a set of fuzzy rules that were generated by the networks.

In each of these studies, it was assumed that once the predictive performances of the ANN models developed had been validated, information about the underlying process could then be gained from the models, either by fuzzy rules or sensitivity analyses. However, validation of the ANN models in these studies was subjective, where predic-

tive performance was assessed by visually inspecting plots of actual versus predicted cell densities in terms of the models' ability to forecast onset, peak and duration of growth events, as it was considered that this assessment could describe model performance better than a composite error measure (e.g. RMSE), which does not account for the timing of the predictions (*Maier et al.*, 1998). It was also known that the available data were limited and the precision of the *Anabaena* cell count data was $\pm 20\%$ or more (*Maier et al.*, 1998). Therefore, it is considered optimistic to have assumed the "validated" models could provide useful information about the general underlying relationship. This was noted in *Maier et al.* (2001). In the present research, ANN models were developed for providing 4-week forecasts of *Anabaena* spp. at Morgan, using both deterministic and Bayesian methods. The RI values (or distributions) were then calculated for each of the model inputs and compared to *a priori* knowledge of the relationship underlying the incidence of *Anabaena* in the lower River Murray in order to validate the usefulness of the model as a means of examining the implications for preventative management options. The input RI values provide similar information to the sensitivity analyses carried out by *Maier et al.* (1998); however, sensitivity analyses involve manipulating the inputs one at a time to determine their relative impacts on the output variable, which is not only time consuming, but can not properly reflect the modelled relationship, nor the causality relations in the actual system, if the inputs are not independent, since interactions between the inputs are not considered.

6.3 AVAILABLE DATA AND MODEL INPUTS

The available data for this case study are summarised in Table 6.2. As seen, the data were supplied from a number of different sources, including the Australian Water Quality Centre (AWQC); the South Australian Department for Water Resources (DWR); and the Murray-Darling Basin Commission (MDBC). All data were available for the period 8 January 1980 to 20 November 1996 and were collected at Morgan as part of routine monitoring, with the exception of flow data, which were recorded at the border between South Australia and New South Wales. Daily flow and river level data were converted to weekly averages. It can be seen, in comparison with Figure 6.2, that the available data set did not include all variables that may be significant in regulating the growth and loss rates of cyanobacteria. For example, while colour and turbidity affect water transparency and light penetration, there is no data available for surface irradiance. Furthermore, turbulent mixing is a function of flow, river level and wind. The only wind data available for the lower River Murray were subject to measurement errors and, as a result, contained an

Table 6.2 Available data for forecasting *Anabaena* spp. in the River Murray at Morgan.

Variable	Units	Sampling Interval	Source
Anabaena spp.	cells/mL	weekly	AWQC
Flow	ML/day	daily	DWR
River level	m	daily	DWR
Temperature	°C	weekly	MDBC
Colour	Hazen Units (HU)	weekly	MDBC
Turbidity	Nephelometric Turbidity Units (NTU)	weekly	MDBC
pH	–	weekly	MDBC
Silica	mg/L	weekly	MDBC
Total Kjedahl Nitrogen (TKN)	mg/L	weekly	MDBC
Total Phosphorus (TP)	mg/L	weekly	MDBC
Soluble Phosphorus (SP)	mg/L	weekly	MDBC

uncharacteristic trend and heteroskedasticity (*Bowden*, 2003). Therefore, these data were not included in the study. Furthermore, the available data set excludes any information on grazing pressure, such as zooplankton numbers.

The *Anabaena* spp. data represent a species complex, which primarily includes counts of the species *Anabaena circinalis* and *Anabaena flos-aquae* (*Maier et al.*, 1998). A plot of the *Anabaena* spp. concentrations in the River Murray at Morgan between January 1980 and November 1986 is shown in Figure 6.3. During this time, pulsed growth events occurred in most years, predominantly in summer. It can be seen that there was an absence of *Anabaena* spp. in the river between 1983 and 1985, which corresponded to high and persistent turbidity levels, and hence, poor light attenuation in the water column. For the majority of the period for which data were available (88.5% of the time), the concentration of *Anabaena* spp. remained below 500 cells/mL, which corresponds to the lowest significant concentration in the Alert Levels framework (see Table 6.1). However, during this time, Alert Level 1 was triggered 77 times, Alert Level 2 was triggered 23 times and Alert Level 3 was triggered once, reaching a maximum concentration of 25,252 cells/mL.

The *Anabaena* spp. data used in this study were collected as grab samples at a fixed depth and location in the river (0.3 m below the water surface and 10 m from the river bank at a point where the river is 150 m wide). However, samples taken from a fixed depth and location are unlikely to provide a representative sample of the population of cyanobacteria in the river, as cyanobacteria tend to be patchy in both their horizontal and vertical distributions (*Jones et al.*, 2003). Furthermore, samples were taken early in

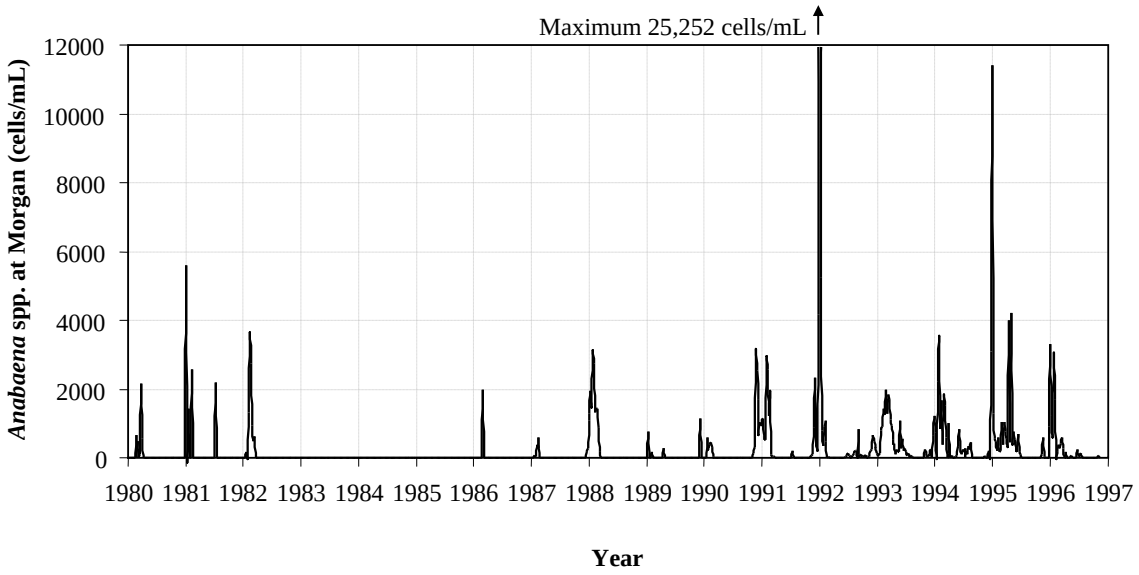

Figure 6.3 *Anabaena* spp. time series at Morgan.

the morning, which may have also affected how representative they were. Cyanobacterial species such as *Anabaena* are able to regulate their buoyancy in calm, stratified conditions. Over night, cells tend to accumulate at the water surface, but as winds increase throughout the day, they are dispersed and may be mixed back into the water column (*Jones et al.*, 2003). Therefore, by taking samples early in the morning, the concentration of *Anabaena* spp. in the river may have been overestimated, given that samples were taken from a fixed location near the water surface.

The concentration of *Anabaena* spp. in a sample is calculated by counting the number of colonies or trichomes of *Anabaena* under a microscope using a calibrated counting chamber and converting the count to cells/mL (*Laslett et al.*, 1997). Therefore, estimates in *Anabaena* spp. concentration are not only subject to sampling errors, but also counting errors, which may arise from bias in technique or random sources (*Jones et al.*, 2003). *Laslett et al.* (1997) derived the following formula for estimating counting precision:

$$\text{Counting error } (\pm\%) = 100\sqrt{2/n} \tag{6.1}$$

where n is the number of units (colonies or trichomes) counted. This gives an estimate of the variability about the observed mean value when repeated counts are made, but does not account for other sources of error, such as unrepresentative sampling or cell losses after sampling. Therefore, the overall error in estimated cell concentration is always greater than the estimated counting error (*Jones et al.*, 2003). During the period for which the

data used in this study were collected, the number of trichomes counted did not remain constant, and consequently the counting precision varied between approximately $\pm 20\%$ and $\pm 50\%$ (*Bowden*, 2003).

The inputs used in this research were those used by *Bowden* (2003) and were selected using the two step PMI selection procedure described in Sections 3.2.4 and 5.3. These inputs are given in Table 6.3, together with their PMI score, the PMI-based RI values and the corresponding rank order of importance. The maximum number of lags considered for each of the variables in Table 6.2 was 26 weeks, as it was assumed that *Anabaena* spp. concentrations would not be affected by conditions that occurred more than 6 months previously. This resulted in a total of 286 candidate inputs, which were then reduced to 8 significant inputs, as shown in Table 6.3.

Table 6.3 Inputs used in *Anabaena* forecasting ANN model, together with the PMI scores, PMI-based RI estimates and rank order of importance.

Variable	Lag (days)	PMI Score	RI (%)	Rank Importance
Anabaena	1	0.296	28.25	1
Anabaena	7	0.178	16.97	2
Anabaena	21	0.076	7.21	7
Silica	1	0.150	14.28	3
Temperature	26	0.120	11.41	4
Flow	2	0.079	7.49	6
Flow	18	0.094	8.93	5
pH	16	0.057	5.44	8

Time series plots of the significant input variables versus *Anabaena* spp. concentration at Morgan are shown in Figures 6.4 to 6.7. In Figures 6.4 and 6.5, it can be seen that there is an inverse relationship between silica and flow with concentrations of *Anabaena* spp., respectively. Silica is not used for *Anabaena* growth itself; however, it is important in determining phytoplankton succession (*Sullivan*, 1990). Silica is necessary for the growth of diatoms and when large populations of diatoms occur, silica becomes depleted from the water. After the silica supply has been exhausted, diatoms cease to grow, allowing cyanobacteria become dominant (*Harris*, 1994). This explains why recent silica concentration was found to be an important predictor of *Anabaena* spp. Furthermore, silica concentration is positively correlated with turbidity, so it is possible that this input not only provides information about phytoplankton succession, but also provides information about the relationship between light availability and *Anabaena* spp. Two flow inputs were found to be significant using the stepwise PMI selection procedure: a recent flow (at a lag

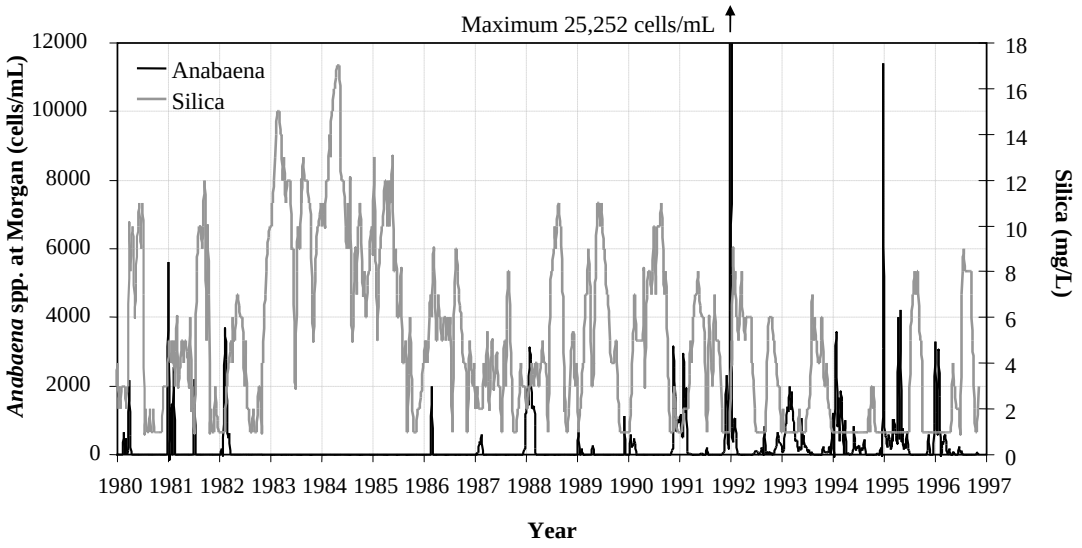

Figure 6.4 Silica versus *Anabaena* spp. at Morgan.

of 2 weeks), for which there was an inverse relationship with *Anabaena* growth, and an earlier flow (at a lag of 18 weeks), for which there was a positive relationship with *Anabaena* growth. This is consistent with the results obtained by *Maier et al.* (1998, 2001), who found that *Anabaena* spp. tend to occur during periods of low flow, following the recession of a flood. It is believed that this is due to the advection of *Anabaena* cells from hydraulically connected lagoons adjacent to the river channel during periods of high flow (*Baker et al.*, 2000).

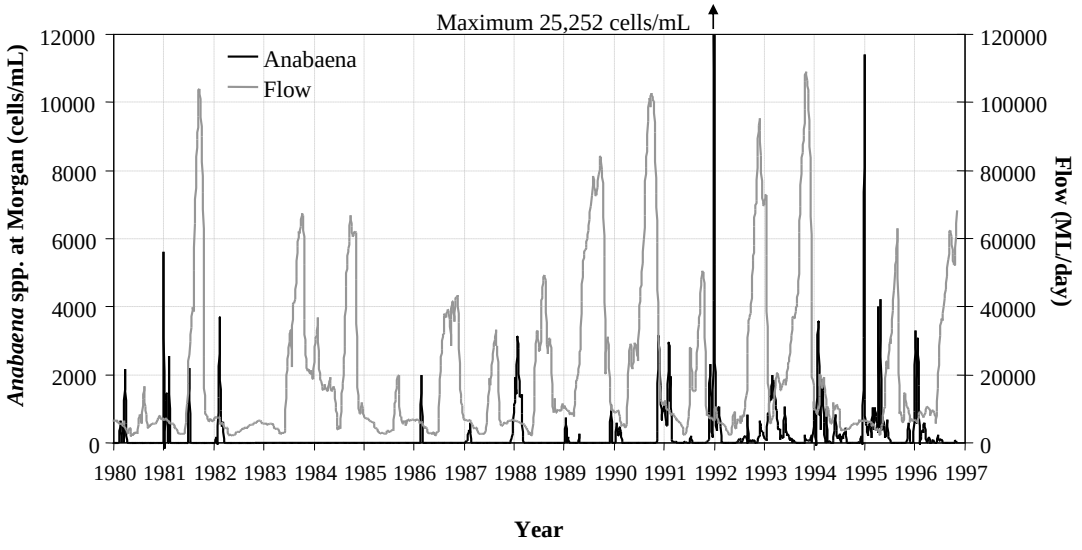

Figure 6.5 Flow into South Australia versus *Anabaena* spp. at Morgan.

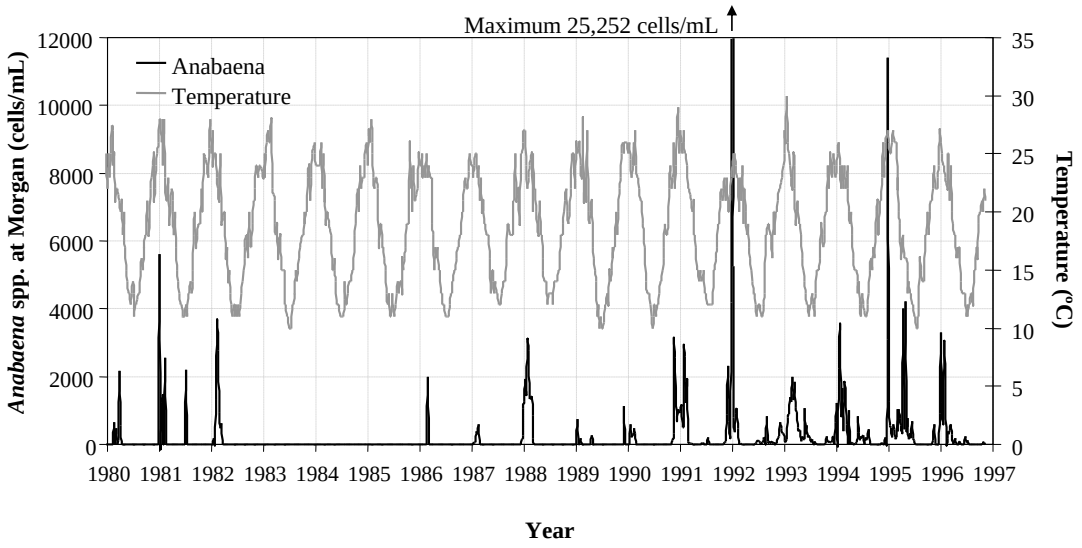

Figure 6.6 Temperature versus *Anabaena* spp. at Morgan.

Figures 6.6 and 6.7 display positive relationships between temperature and pH with *Anabaena* spp. concentration, respectively. In both cases, the PMI algorithm found large lags of these variables to be of more significance than recent values. High temperatures increase growth rates and affect the solubility of dissolved gases in water, but more importantly, when combined with calm conditions can lead to thermal stratification of the river, which has been shown to be a necessary condition for bloom development (*Sherman et al.*, 1998). As seen in Figure 6.6, the temperature time series displays strong seasonal

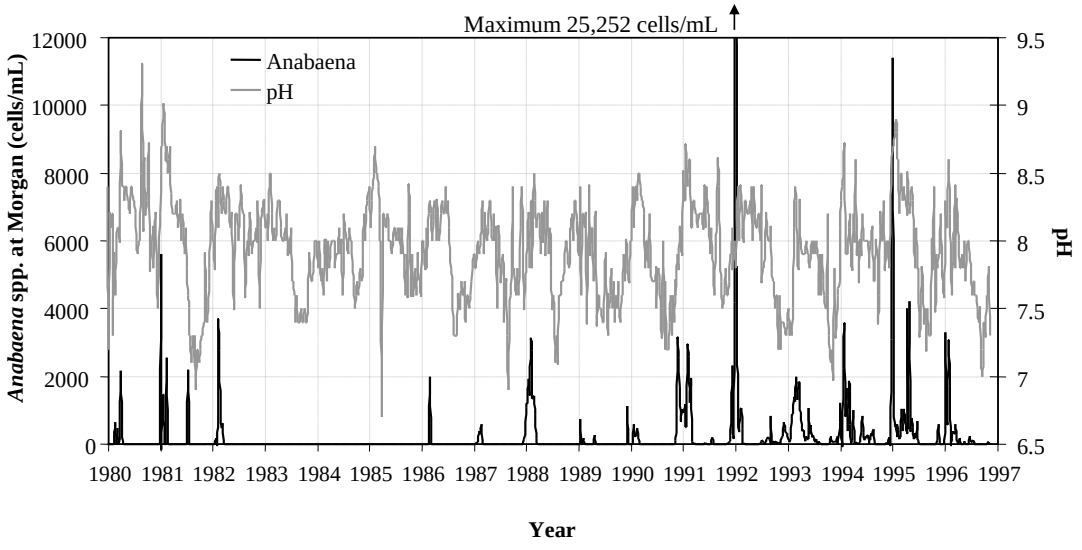

Figure 6.7 pH versus *Anabaena* spp. at Morgan.

variation; thus, there is an almost equally strong inverse relationship between *Anabaena* concentration and temperature at a lag of 26 weeks (6 months) as there is a positive relationship between *Anabaena* concentration and current temperature. Therefore, it is not surprising that the PMI algorithm found temperature at a lag of 26 weeks to be significant. It is considered that this is also the reason why pH at a lag of 16 weeks, which has an inverse relationship with *Anabaena* concentration, was selected as a significant input. During photosynthesis, CO_2 is consumed, which then raises the pH of the water. At high pH levels and low CO_2 concentration, cyanobacteria can utilise bicarbonate in the water and continue photosynthesising efficiently (*Blue-Green Algae Task Force*, 1992). Therefore, incidences of cyanobacteria are favoured by high pH.

Overall, the most important input selected using the PMI algorithm was the most recent *Anabaena* spp. concentration, indicating that this input yields significant information about concentrations of *Anabaena* spp. 4 weeks in advance. This was not surprising, as the concentration of *Anabaena* spp. depends not only on the growth rate, but also on the initial population size. However, due to the rapidly varying nature of *Anabaena*, it is considered unlikely that the concentrations of *Anabaena* spp. at lags of 7 and 21 weeks would significantly affect *Anabaena* spp. concentrations 4 weeks into the future. It is possible that these inputs were selected by the PMI procedure due to spurious correlations in the data or the inappropriate use of the Gaussian reference bandwidth in calculating the PMI score of non-normally distributed data (see Section 3.2.4).

The SOM data division method discussed in Section 3.2.2 was used to divide the available data into training, testing and validation subsets. After accounting for the appropriate lags of the input and output variables, there were 851 available data samples. A comparison of the average silhouette widths and discrepancy values (see Section 3.2.2.2) for various SOM grid sizes ranging from 1×2 to 15×15 showed that a grid size of 12×12 was optimal for clustering this data set. This resulted in 89 clusters containing at least 3 samples, 17 clusters containing 2 samples and 13 clusters containing only 1 sample (25 grid cells were empty). From these clusters, 545 (64%) samples were allocated to the training data subset, 136 (16%) samples were allocated to the testing subset and the remaining 170 (20%) samples were allocated to the validation subset. All clusters containing only 1 sample were allocated to the training set, while the first sample of each 2-sample cluster was also allocated to the training subset and the other was allocated to the testing subset. A histogram displaying the probability density of the *Anabaena* spp. data is shown in Figure 6.8. As it can be seen, the data are highly non-normal, positively skewed and appear to have several outliers. Logarithmic transformations are commonly

used in microbiological studies where the data are distributed in this manner and there is the potential for exponential growth and decay. Therefore, it was expected that a linear transformation would be insufficient for the SSE to be an appropriate error model. However, to check this, the input and output data were initially only linearly standardised, and the resulting residuals were inspected after the preliminary models had been developed.

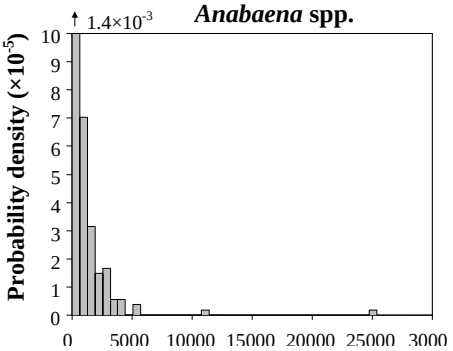

Figure 6.8 Probability density of the available *Anabaena* spp. data.

6.4 DETERMINISTIC ANN DEVELOPMENT

The state-of-the-art deterministic ANN methodology, described in Chapter 3, was first applied to develop an ANN model for forecasting *Anabaena* spp. in the River Murray at Morgan, 4 weeks in advance.

6.4.1 Methods

6.4.1.1 Model Selection

The trial-and-error approach discussed in Chapter 3 was used to determine the optimal model structure for this case study. Given the findings of Section 3.4 and Chapter 5, the in-sample BIC and out-of-sample AIC, calculated based on the weights obtained when training was stopped early to prevent overfitting, were used to assess the generalisability of the models developed. In *Bowden* (2003), only ANNs with two hidden layers were considered for this case study, given the inputs in Table 6.3. The best model developed contained 10 hidden nodes (7 in the first hidden layer and 3 in the second), which resulted in 91 weights. For a single hidden layer network, this corresponds to an ANN with 9 hidden nodes, which also has 91 weights. However, two hidden layer ANNs generally require fewer weights to obtain an equivalent solution to a single hidden layer network (*Bebis and Georgiopoulos*, 1994); therefore, the upper limit for the number of hidden

nodes considered in this research was 20. Similar to the salinity case study presented in Chapter 5, ANNs containing 5, 10, 15 and 20 hidden nodes were initially included in the trial-and-error search in order to narrow down the search to a 10 hidden node range (i.e. 1–10 or 10–20). Following selection of this range, the trial-and-error approach was repeated, with the number of hidden nodes increasing in increments of 2. Finally, the search was reduced to 1 hidden node increments, where ANNs containing one fewer and one more hidden node than the best network selected in the previous trial were tested.

6.4.1.2 Training

The SCE-UA algorithm was used to train the models, subject to the parameter constraints given in Table 5.3. For each network, the SCE-UA algorithm was initialised with three different sets of random weights and the best results obtained were used for further analysis. Training was run until the stopping criterion given by (3.33) was met or after 10 million error function evaluations, whichever occurred first. Cross-validation using the test data set was also employed during training and the weights resulting in the minimum test set error were saved, and the corresponding model outputs computed.

6.4.1.3 Validation

Similar to the studies carried out by *Maier et al.* (1998, 2000, 2001) and *Bowden* (2003), the optimal model selected was validated by inspecting plots of the predicted *Anabaena* spp. concentrations versus the observed concentrations, to assess the model's ability to forecast the onset, peak and duration of growth events. These are the three most important characteristics describing a cyanobacterial bloom; however, the error measures described in Section 3.2.1 are unable to properly describe model performance in terms of these factors (*Maier et al.*, 1998). For example, the RMSE is a good measure of general model performance and fit; however, two models may have very similar RMSE values, but may differ significantly in terms of the usefulness of their predictions. If the predictions of one model lead the actual event, while the other model's predictions lag it, the former is a more useful model. However, in order to compare the models developed in this research to those developed by *Bowden* (2003), the RMSE was also evaluated for the training, testing and validation subsets.

The modified Connection Weight Approach (see Section 3.4.4.3) was used to evaluate RI values for each of the model inputs, which were then compared to the PMI-based RI estimates given in Table 6.3. As mentioned in Section 3.4.4, absolute RI values are used in such an evaluation, as the PMI approach does not give directions of the input-output

relationships. On the other hand, the actual RI values calculated using the modified Connection Weight Approach do give directions of the input-output relationships and can be useful for comparing the modelled function to *a priori* knowledge of the underlying physical relationship. For this case study, *a priori* knowledge of the physical relationship is vague; however, as discussed in Section 6.3, there is a known inverse relationship between flow and silica with *Anabaena* spp. concentration and a positive relationship between temperature and pH with *Anabaena* spp. concentration. If it could be shown that the model had successfully approximated these relationships, there would be the potential to use the model for hypothesis testing of preventative management options, such as flow management.

6.4.2 Results

Histograms of the 5, 10, 15 and 20 hidden node ANN models' standardised residuals are shown in Figure 6.9, in comparison to the standard normal distribution. Similar to the *Anabaena* spp. data, the residuals were found to be highly non-normal and positively skewed when a linear transformation was applied to the data, confirming that a nonlinear transformation was required. A logarithmic (base 10) transformation was therefore applied to the observed *Anabaena* spp. data in order to compress the distribution and reduce the impact

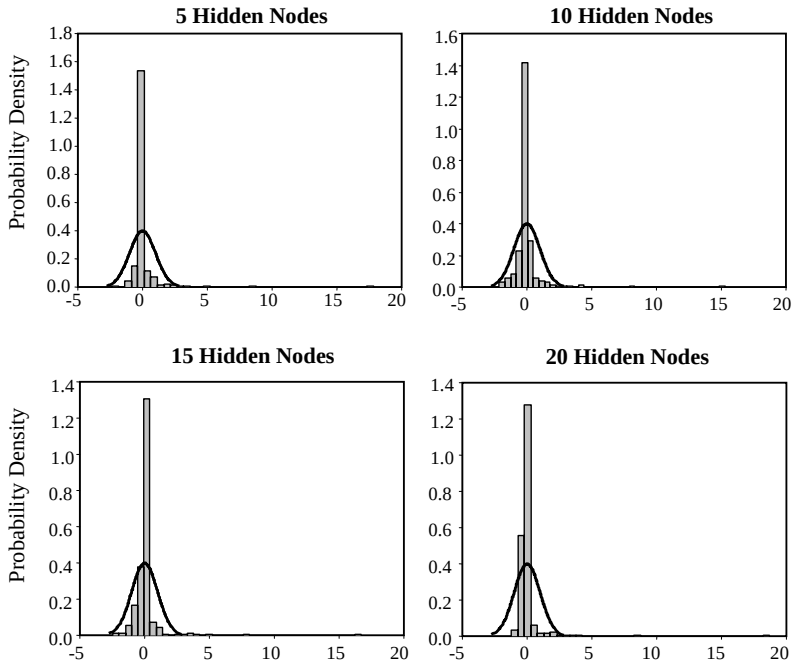

Figure 6.9 Residuals resulting from linearly scaled data.

of the large *Anabaena* counts on the model error. Histograms of the resulting standardised model residuals for the 5, 10, 15 and 20 hidden node ANNs, using log transformed output data, are shown in Figure 6.10. While still not normal, the residual distributions are much more compact, with all values lying within the $[-5, 5]$ range (unlike the distributions shown in Figure 6.9, which contained values as high as 18.8). Therefore, it was considered that this transformation resulted in more appropriate processing of the data.

The the out-of-sample AIC and in-sample BIC values, obtained when training stopped early according to the test set error, are plotted against one another in Figure 6.11 for the different network sizes. There is good agreement between the different criteria, with both indicating that the 5 hidden node ANN had the best generalisability when used to model the *Anabaena* spp. data. Therefore, the trial-and-error search for the optimal structure was reduced to the 1–10 hidden node range and the subsequent ANN models trained and tested for their generalisability contained $2, 4, \ldots, 8$ hidden nodes. The resulting in-sample BIC and out-of-sample AIC values for these networks are plotted in Figure 6.12. Once again, it can be seen that there is reasonable agreement between these criteria; however, the AIC shows a more definite preference for the 2 hidden node model, whereas the BIC indicated the 2 and 4 hidden node ANNs had similar generalisability. Consequently, ANN models containing 1 and 3 hidden nodes were also trained and tested for their generalisability and the resulting in-sample BIC and out-of-sample AIC values were compared to those

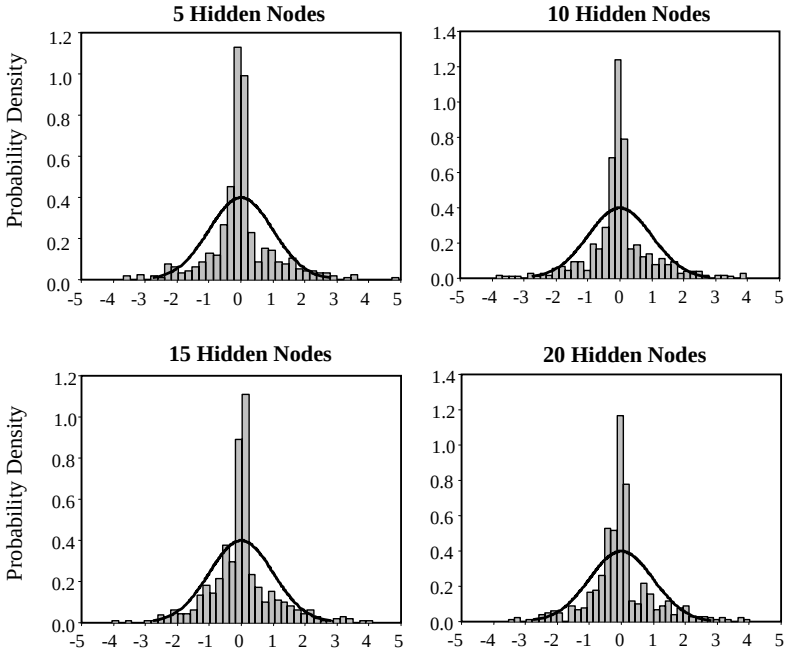

Figure 6.10 Residuals resulting from log transformed data.

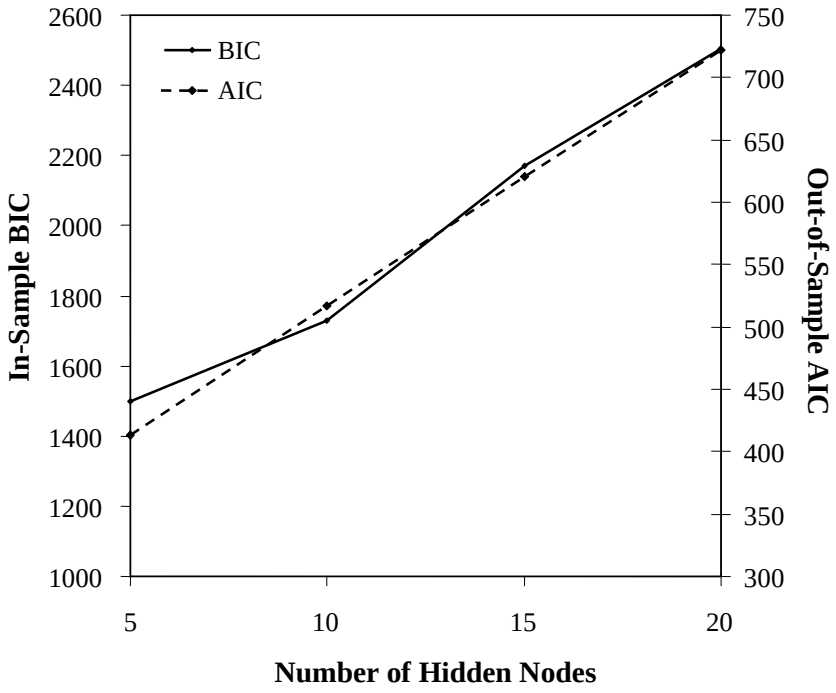

Figure 6.11 In-sample BIC and out-of-sample AIC values for 5, 10, 15 and 20 hidden node ANN models.

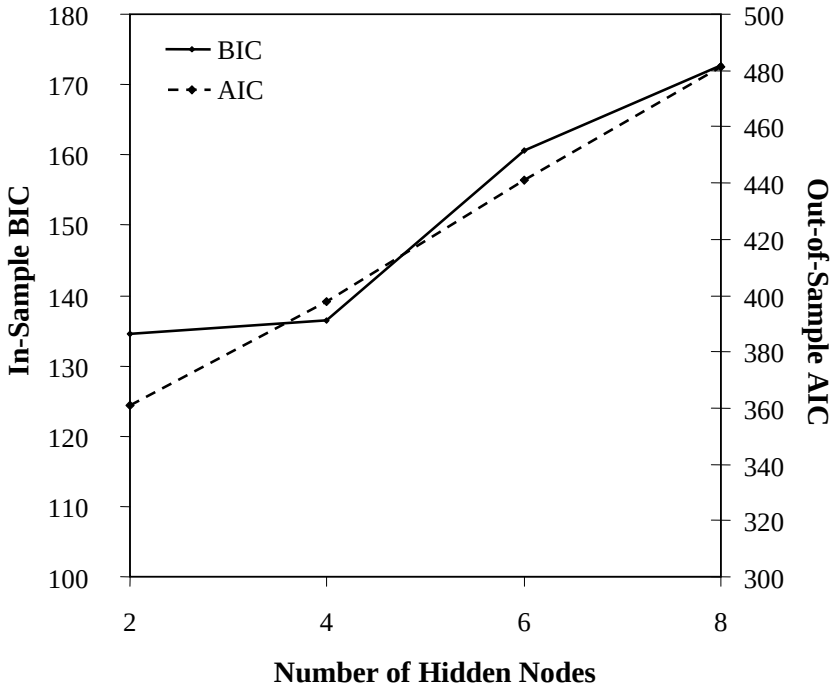

Figure 6.12 In-sample BIC and out-of-sample AIC values for 2, 4, 6 and 8 hidden node ANN models.

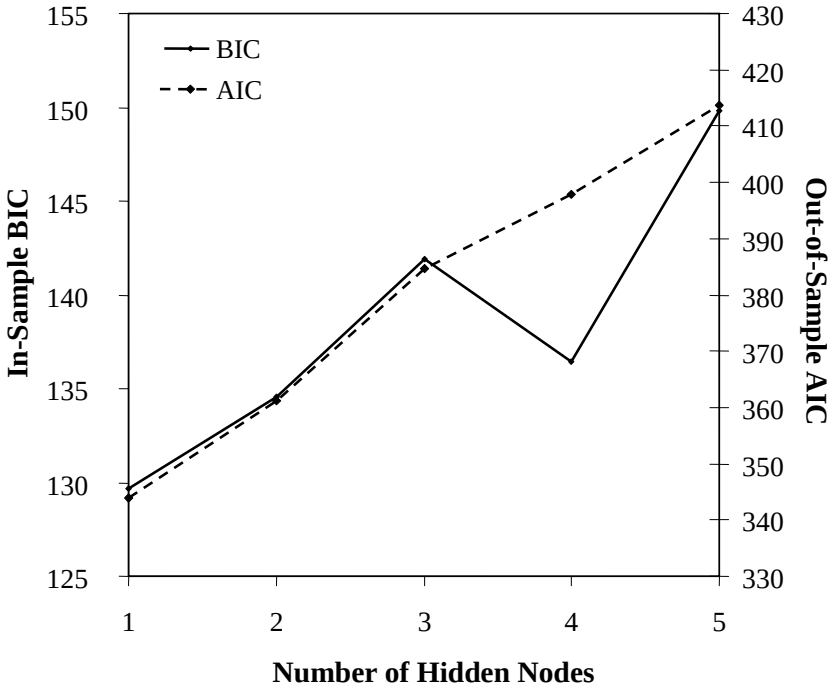

Figure 6.13 In-sample BIC and out-of-sample AIC values for 1, 2, 3, 4 and 5 hidden node ANN models.

obtained for the 2, 4 and 5 hidden node models, as shown in Figure 6.13. As can be seen, both criteria indicate that the 1 hidden node ANN had the best generalisability of the models tested; however, the out-of-sample AIC appears to be more consistent. This is likely due to the fact that the BIC values were calculated based on the training set error when training was stopped early, which may result in various degrees of overfitting or underfitting the training data. In order to identify the 1 hidden node ANN as optimal from the possible 20 network sizes considered, only 10 ANNs required training and testing.

The selected 1 hidden node model was then subjected to the independent validation data and the resulting RMSE values for the (log transformed) training, testing and validation subsets are shown in Table 6.4, in comparison to those obtained by *Bowden* (2003) with a two hidden layer ANN containing 10 hidden nodes. This was not the best performing ANN developed by *Bowden* (2003) for this case study; however, it was the best MLP developed using the inputs given in Table 6.3 (general regression neural networks (GRNNs) were also developed by *Bowden* (2003), which outperformed this MLP). The 1 hidden node ANN selected in this research contains 11 weights, compared to the 91 weights contained in the two hidden layer ANN developed by *Bowden* (2003). However, as it can be seen in Table 6.4, the models' performances, in terms of fit to the observed data, were comparable.

Table 6.4 RMSEs for the 1 hidden node ANN *Anabaena* spp. forecasting model developed using deterministic methods, in comparison to *Bowden* (2003) results.

Model	Train	Test	Validation
1 hidden node ANN	0.746	0.790	0.809
10 hidden node ANN *Bowden* (2003)	0.733	0.699	0.852

A time series plot of the 4-week forecasts obtained for the recombined training, testing and validation data is shown in Figure 6.14. It can be seen that the model was able to forecast the onset and duration of growth events of *Anabaena* spp. reasonably well; however, it tended to under-predict the peak concentrations. Not once was the model able to predict a significant growth event (i.e. concentration $>$ 500 cells/mL, shown in Figure 6.14 in log scale). The model performance in terms of forecasting peak, onset and duration of growth events, was also assessed based only on the independent validation data, as shown in Figure 6.15. These results show that, while the model was able to forecast the occurrence (and absence) of growth events, the predictions often slightly lagged the actual values.

The model-based RI values for each of the model inputs, calculated using the modified Connection Weight Approach, are given in Table 6.5, together with the resulting rank order of modelled input importance and, for comparison, the PMI-based RI estimates.

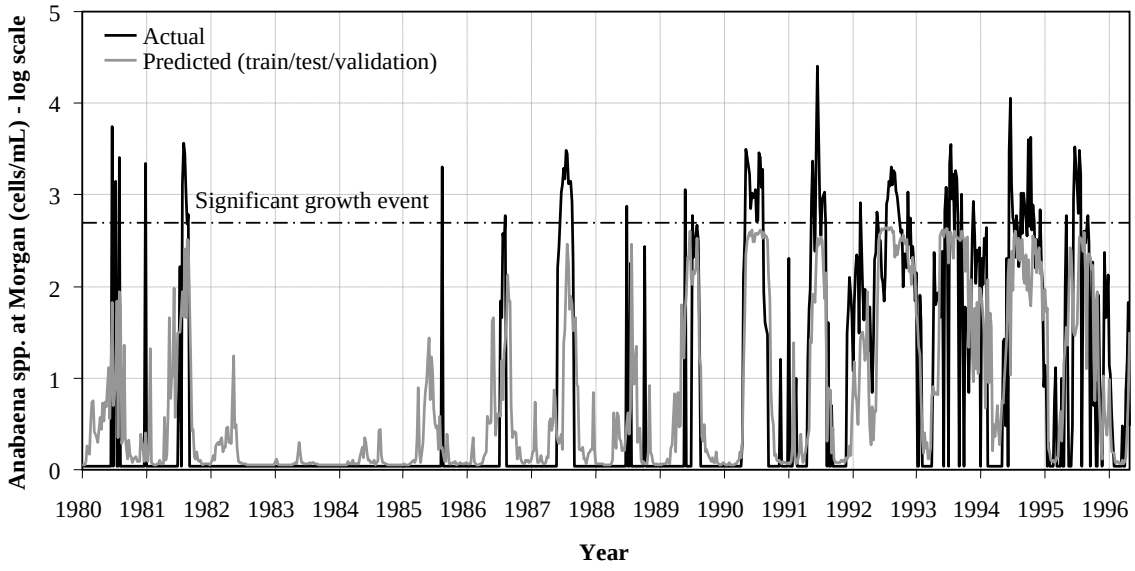

Figure 6.14 Time series plot of the 4-week *Anabaena* spp. forecasts obtained for the training/testing/validation data using the 1 hidden node ANN model.

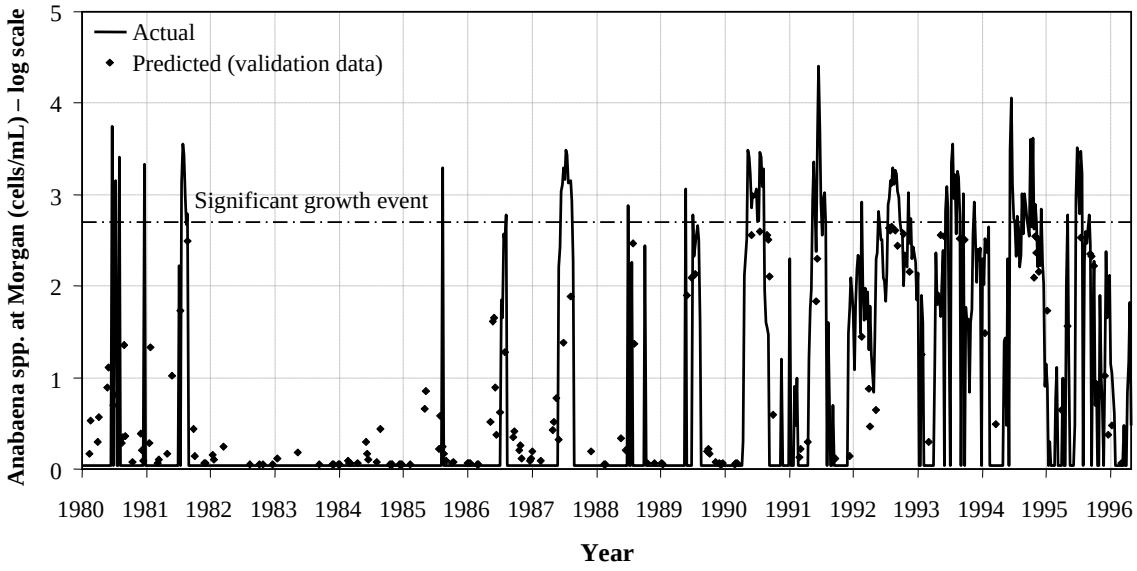

Figure 6.15 4-week *Anabaena* spp. forecasts obtained for the validation data using the 1 hidden node ANN model.

The magnitudes of the model-based RI estimates are relatively close to the magnitudes of the PMI-based estimates, indicating that the model was able to adequately capture the information contained in the data.

Given the inputs included in the model, it is acknowledged that, at best, a very simplified picture of the physical process could be obtained. The model-based RI values were used to determine whether this simplified picture would be sufficient for hypothesis testing of management strategies that could potentially be used to control the occurrence

Table 6.5 Model-based RI estimates and order of input importance in comparison to the PMI-based RI estimates.

	Rank	*RI* (%)	
Input	Importance	Model-based	PMI-based
$Anabaena_{t-1}$	5	13.59	28.25
$Anabaena_{t-7}$	8	3.95	16.97
$Anabaena_{t-21}$	7	7.62	7.21
Silica_{t-1}	1	-20.09	14.28
$\text{Temperature}_{t-26}$	2	-16.93	11.41
Flow_{t-2}	3	14.23	7.49
Flow_{t-18}	6	9.44	8.93
pH_{t-16}	4	14.15	5.44

of cyanobacterial blooms. As it can be seen in Table 6.5, the ANN correctly modelled the inverse relationship between silica_{t-1} and $Anabaena_{t+3}$ discussed in Section 6.3. The positive relationship between $Anabaena_{t-1}$, which provides information about the initial population size, and $Anabaena_{t+3}$ was also correctly modelled by the ANN. As discussed in Section 6.3, it was considered that longer lags of $Anabaena$ would not be important predictors of $Anabaena_{t+3}$ and may have been selected by the PMI procedure due to either spurious correlations in the data or the inappropriate use of the Gaussian reference bandwidth in calculating the PMI scores. As seen in Table 6.5, these inputs were given the least importance by the model. Although the inverse relationship between $\text{temperature}_{t-26}$ and $Anabaena_{t+3}$ was correctly estimated, it would have been more physically plausible to substitute temperature_{t-1} for $\text{temperature}_{t-26}$ and allow the model to derive the correct positive relationship between temperature and $Anabaena$ growth. Nevertheless, as temperature cannot be controlled, this was considered to be of relatively minor significance to the usefulness of the model as a means of hypothesis testing. Of greater importance was the fact that the ANN incorrectly modelled a positive relationship between recent flow conditions and the incidence of $Anabaena$, as seen by the positive RI value given to flow_{t-2}. Flow management strategies involve increasing flows to disrupt thermal stratification or flush blooms or potential blooms (*Senate Standing Committee*, 1993). However, the model indicates that if current flows are increased, the concentration of *Anabaena* spp. will also increase in 4 weeks time. Therefore, this model is considered not to be useful for investigating the direct response of *Anabaena* spp. to changes in the flow regime.

6.5 BAYESIAN ANN DEVELOPMENT

The Bayesian ANN framework, proposed in Chapter 4, was also applied to develop a probabilistic ANN model for forecasting Anabaena spp. at Morgan. The results obtained were then compared to those presented in Section 6.4.2, to determine whether a more useful model could be developed using the Bayesian ANN framework, given the uncertainty associated with the available data.

6.5.1 Methods

6.5.1.1 Model Selection

In the deterministic part of the case study it was found that a 1 hidden node ANN was the optimal model structure for forecasting *Anabaena* spp. concentrations. Therefore,

networks containing $2, 4, \ldots, 10$ hidden nodes were initially developed and their evidence values compared, in order to determine the most appropriate structure under the Bayesian framework. The -1/2BIC, G-D and C-J evidence estimators, discussed in Section 4.3.4, were used to evaluate the evidence values of each network. The hidden-output weight distributions for the ANN model with the maximum evidence value were then inspected to determine whether any of the hidden nodes could be pruned from the most probable model, as described in Section 4.3.4.2. These distributions were also inspected for the model containing two more hidden nodes than the ANN with the maximum evidence, in order to verify the results obtained for the highest evidence model.

6.5.1.2 Training

The MCMC training algorithm, described and developed in Chapter 4, was used to train each of the models developed. Four parallel chains were simulated using this algorithm, initialised using the weights obtained in the deterministic part of this study when training was stopped early. A hierarchical prior distribution in the form given by (2.15) was assumed for the network weights. For the first 500 iterations ($t_{\sigma_0^2} = 500$), the $\sigma_{\mathbf{y}}^2$ hyperparameter was fixed equal to 0.8, as the residual variance values $\hat{\sigma}_{\mathbf{y}}^2$ calculated at maximum log likelihood (based on scaled data) were between approximately 0.33 (10 hidden node ANN) and 0.41 (2 hidden node ANN), with corresponding log likelihood values between approximately -530 (2 hidden node ANN) and -470 (10 hidden node ANN). Fixing $\sigma_{\mathbf{y}}^2 = 0.8$ resulted in reductions in the log likelihood values between approximately 9% and 17%. The MCMC algorithm was initially run for 600,000 iterations and traces of the mean $\log p^*(\mathbf{w}|\mathbf{y})$, the mean $\log L(\mathbf{w})$ and the mean $\log p(\mathbf{w})$ densities, calculated by taking the average of the four parallel chains, were inspected to determine whether or not convergence had been achieved within this time and to determine the appropriate number of burn-in iterations t_b. Traces of the $\log p^*(\mathbf{w}|\mathbf{y})$, $\log L(\mathbf{w})$ and $\log p(\mathbf{w})$ values obtained from the individual chains were also inspected to assess convergence. Predictive distributions, from which mean predictions and 95% prediction limits were evaluated, were calculated based on 10,000 weight vectors, randomly sampled after approximate convergence had been achieved.

6.5.1.3 Validation

The optimal model structure selected using the BMS approach was validated by subjecting it to the independent testing and validation subsets, not used for training. Similar to the deterministic part of this study, plots of the mean predicted *Anabaena* spp. con-

centrations and the 95% prediction limits were inspected, to assess the model's ability to forecast the onset, peak and duration of growth events. The RMSE values were also evaluated based on the mean forecasts for the training, testing and validation subsets and compared to those obtained in the deterministic part of the study and those obtained by *Bowden* (2003).

The RI distributions for each input were determined by applying the modified Connection Weight Approach (without taking absolute values) to the weight vectors sampled from the posterior distribution. Similar to the deterministic part of this study, these distributions were compared to *a priori* knowledge of the underlying physical relationship, in order to assess the usefulness of the model as a means for investigating the response of *Anabaena* spp. to different management strategies.

6.5.2 Results

Shown in Figure 6.16 are traces of the mean $\log p^*(\mathbf{w}|\mathbf{y})$, $\log L(\mathbf{w})$ and $\log p(\mathbf{w})$ densities obtained for the $2, 4, \ldots, 10$ hidden node ANNs during training with the MCMC algorithm. It can be seen that, apart from the 10 hidden node ANN, the MCMC algorithm had reached approximate convergence within 300,000 iterations for each of the network sizes. For the larger models (containing 6–10 hidden nodes), there was greater variation

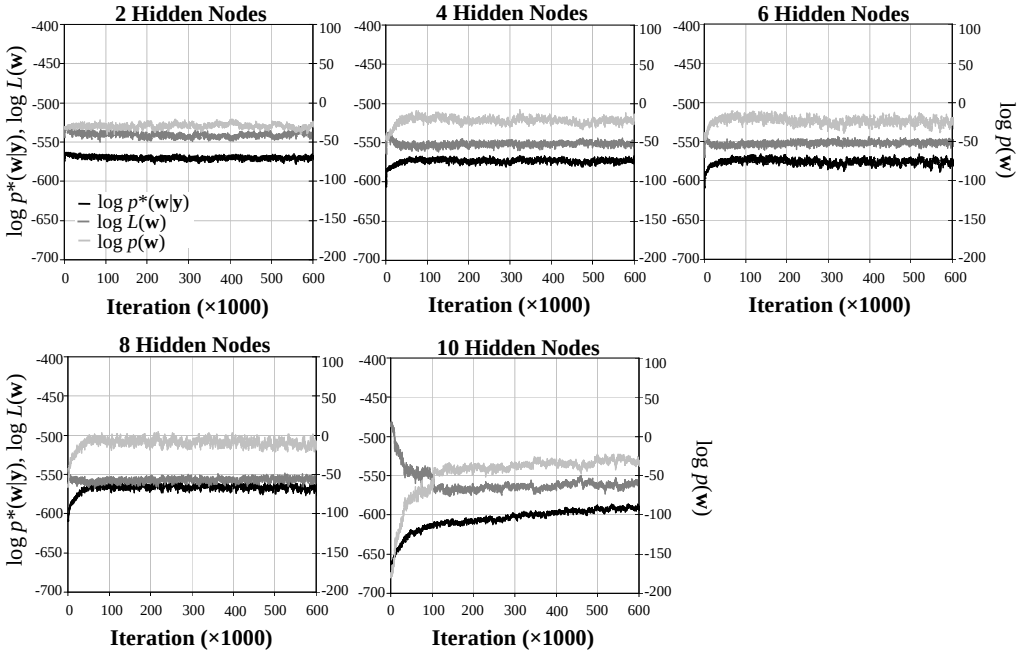

Figure 6.16 Mean $\log p^*(\mathbf{w}|\mathbf{y})$, $\log L(\mathbf{w})$ and $\log p(\mathbf{w})$ traces for the $2, 4, \ldots, 10$ hidden node ANNs.

in the $\log p^*(\mathbf{w}|\mathbf{y})$, which can be seen more clearly in Figure 6.17, which shows traces of the $\log p^*(\mathbf{w}|\mathbf{y})$ densities obtained from the individual MCMC chains for each network size. Also seen more clearly in this figure is the non-convergence of the MCMC algorithm when applied to train the 10 hidden node ANN. However, it is considered that this was primarily due to the increasing prior probabilities as a result of the still decreasing weight magnitudes, rather than non-convergence about the most appropriate likelihood. It can be seen in Figure 6.16 that the mean $\log L(\mathbf{w})$ densities had roughly converged to an approximate value of around -550 for all of the network sizes. As the -1/2BIC evidence estimator used in the proposed BMS approach does not depend on the values of $\log p^*(\mathbf{w}|\mathbf{y})$ or $\log p(\mathbf{w})$, but only on the value of $\log L(\mathbf{w})$, it was considered unnecessary to train the networks for longer in order to achieve accurate results when selecting the most appropriate model structure. Therefore, the number of burn-in iterations that were discarded for all of the ANNs was 300,000, while the remaining 300,000 were used for further analysis.

Shown in Figure 6.18 are plots of the evidence values estimated with the -1/2BIC, G-D and C-J estimators for the $2, 4, \ldots, 10$ hidden node ANNs (however, due to the similarity between the G-D and C-J evidence estimates, these are difficult to distinguish from

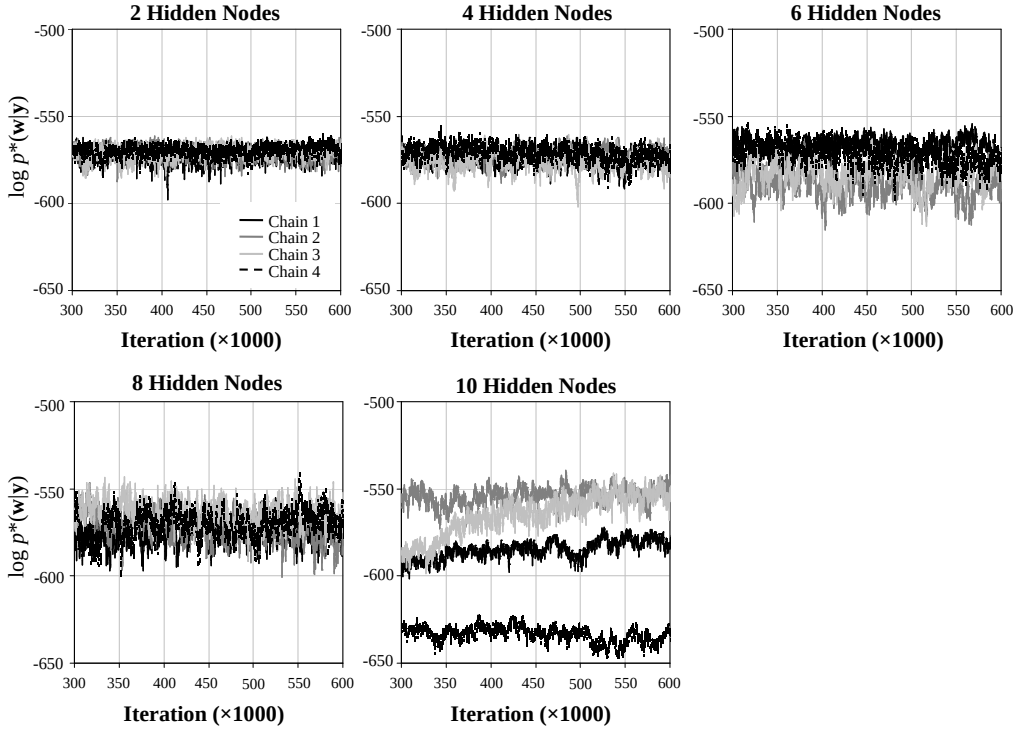

Figure 6.17 Log $p^*(\mathbf{w}|\mathbf{y})$ traces obtained from the 4 parallel MCMC chains for the $2, 4, \ldots, 10$ hidden node ANNs.

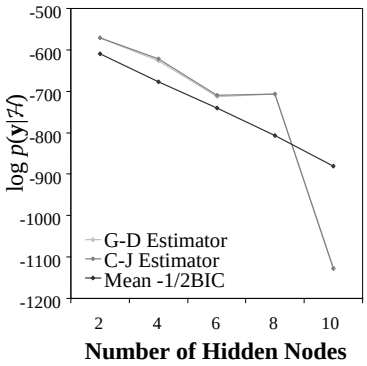

Figure 6.18 Evidence estimates for the $2, 4, \ldots, 10$ hidden node ANNs.

one another). As can be seen, there was a significant negative relationship between the estimated evidence values and the number of hidden nodes in the network. This was as expected, since approximately the same $\log L(\mathbf{w})$ values had been obtained by each of the networks. Also as expected was the fact that the -1/2BIC estimator is apparently unaffected by the inappropriate convergence of the 10 hidden node ANN. However, this would not have been the case if the non-convergence was due to an increasing or decreasing trend in the $\log L(\mathbf{w})$ values, rather than in the $\log p(\mathbf{w})$ values. It is also apparent that the G-D and C-J estimators were sensitive to the inappropriate convergence, which is understandable, as calculation of these values also depends on the prior density. Nevertheless, the 2 hidden node ANN was found to have the highest posterior probability according to each of the estimators. The $BF_{Rank1,i}$ results, calculated based on the -1/2BIC evidence values, and presented in Table 6.6, indicate that there is very strong evidence in favour of the 2 hidden node ANN over the other network sizes, according to the interpretive scale given in Table 4.1.

To check these results, the marginal posterior distributions of the hidden-output weights of the 2 hidden node ANN were inspected. These are shown in Figure 6.19, where it can

Table 6.6 Log Bayes Factors in favour of the highest ranked model.

Rank	Hidden Nodes	$\log_e BF$ in Favour of Rank 1 Model
1	2	–
2	4	68.232
3	6	131.543
4	8	197.078
5	10	270.676

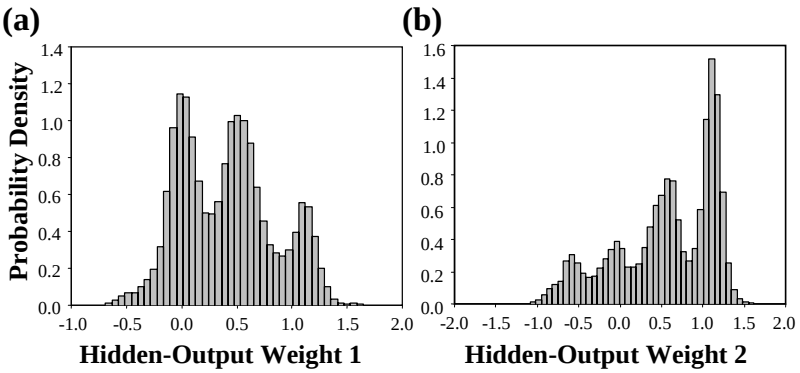

Figure 6.19 Marginal posterior hidden-output weight distributions for the 2 hidden node ANN.

be seen that the distributions of both hidden-output weights include zero within the 95% highest density regions, indicating that at least one of the hidden nodes was unnecessary. The scatter plot of the weights, shown in Figure 6.20, indicates that only one of the hidden nodes could be removed from the network, as the joint distribution of the weights does not pass through the origin. However, as discussed in Section 4.4.3.2, when there are only two hidden nodes in the network, the joint distribution of the hidden-output weights will never pass through the origin unless there is no relationship between the model inputs and outputs, since the inputs would then be disconnected from the output. The relationship between environmental variables and *Anabaena* spp. growth is known to be highly non-linear (even if there is insufficient data to properly describe this nonlinearity); therefore, a linear model, resulting from an ANN with no hidden layer, was not considered in this research. The strong correlation between the two hidden-output weights is also evident in Figure 6.20. This is a symptom of overparameterisation, which may be difficult to see in higher dimensional models. It was therefore concluded that a 1 hidden node network was

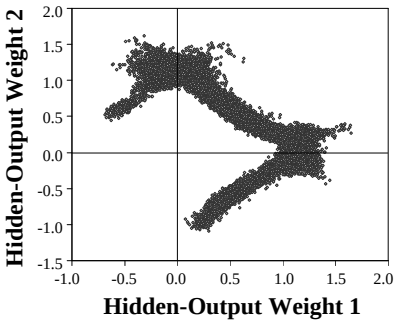

Figure 6.20 Scatter plot of hidden-output weight 1 versus hidden-output weight 2 for the 2 hidden node ANN.

the most appropriate structure for forecasting *Anabaena* spp. concentrations.

In order to verify this result, the marginal posterior distributions for the hidden-output weights of the 4 hidden node ANN were also inspected and are shown in Figure 6.21. It can be seen that all four of the distributions include zero within the 95% highest density regions. However, to determine how many hidden nodes could be pruned from the network, scatter plots of every possible hidden-output weight pair had to be inspected. This resulted in six scatter plots, as shown in Figure 6.22. It can be seen in this figure that all but one (subplot (c)) of the joint distributions passed through the origin, indicating that three of the hidden nodes were unnecessary in the model and could therefore be pruned. This would result in a 1 hidden node ANN; thus, confirming the results obtained above.This example also highlights why it is useful to first evaluate the evidence values of competing models in order to identify the model with the highest posterior probability. If inspection of hidden-output weight distributions alone was relied upon to find the most appropriate structure, the process could become very time consuming when considering larger models, since the joint distributions of all possible hidden-output weight pairs that include zero would need to be inspected.

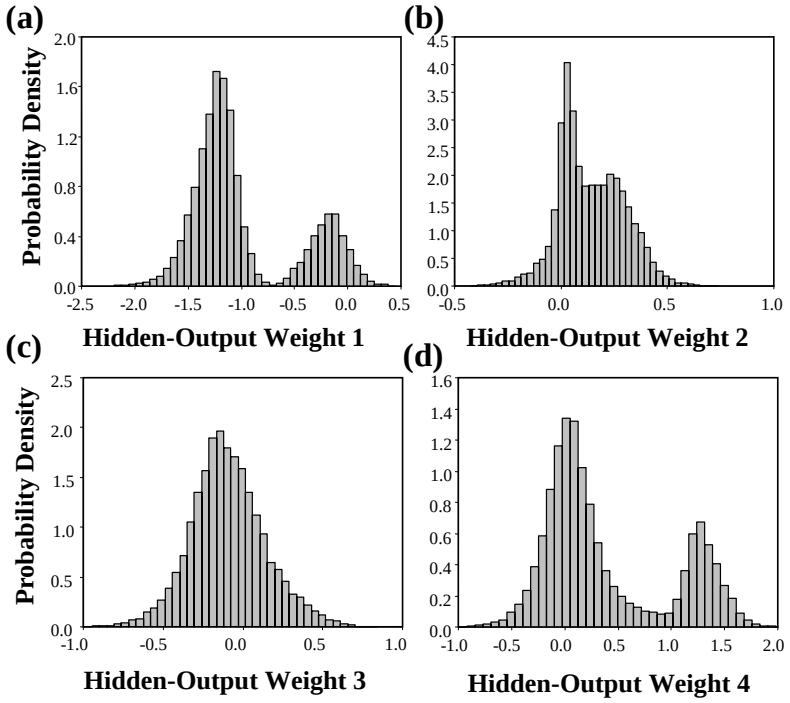

Figure 6.21 Marginal posterior hidden-output weight distributions for the 4 hidden node ANN.

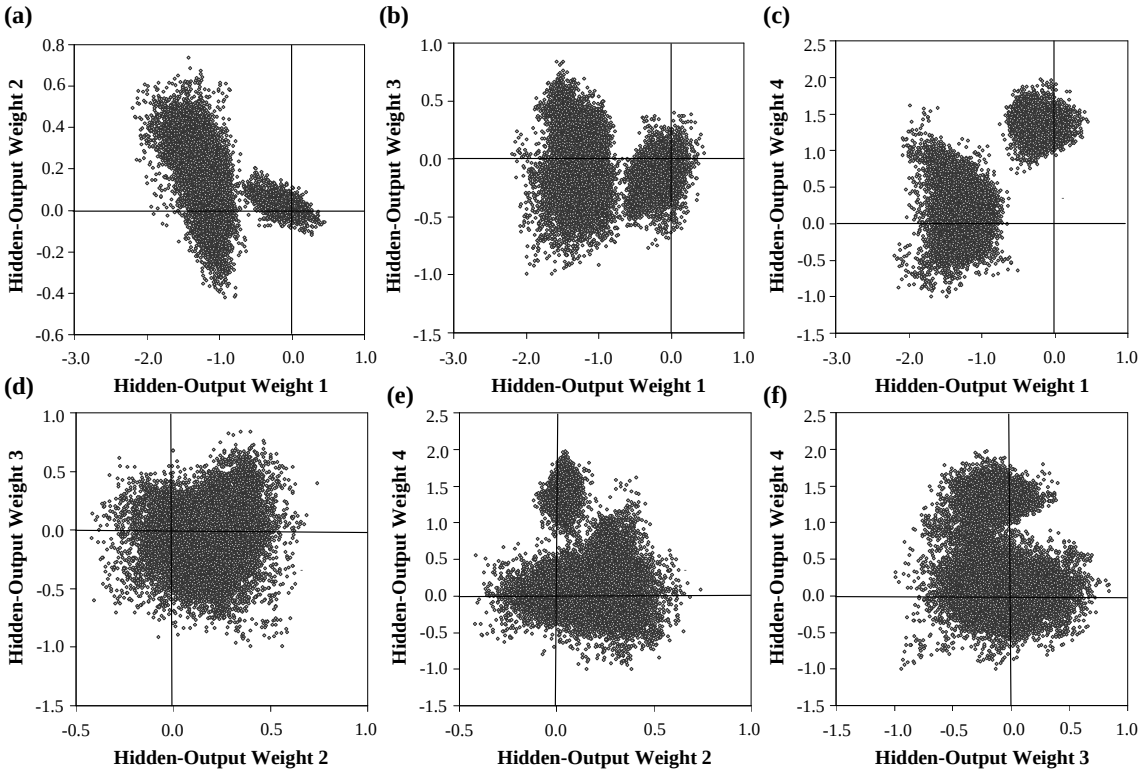

Figure 6.22 Scatter plots of the hidden-output weights for the 4 hidden node ANN.

Given the results of the BMS procedure, a 1 hidden node ANN was trained using the MCMC algorithm. It was found that convergence of the algorithm was easily achieved within 300,000 iterations; thus, this number of iterations was still appropriate for burn-in. As a final check that the 1 hidden node ANN was the most appropriate structure for this problem, the evidence values were estimated using the -1/2BIC, G-D and C-J estimators. These are plotted in Figure 6.23, in comparison to the evidence values estimated for the $2, 4, \ldots, 10$ hidden node ANNs, where it can be seen that the -1/2BIC evidence estimate confirmed that the 1 hidden node ANN was the most appropriate structure.

A time series plot of the mean 4-week *Anabaena* spp. forecasts and 95% prediction limits for the recombined training, testing and validation data is shown in Figure 6.24 (in log scale). It can be seen from the width of the prediction limits that there is significant uncertainty associated with the forecasts. However, it can also be seen that, unlike the deterministic model, the majority of peak concentrations have been accounted for within these limits (91.3% of the recombined data set was accounted for). Furthermore, the model was able to correctly forecast the occurrence, or absence, of significant growth

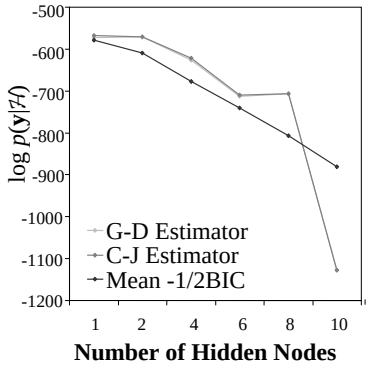

Figure 6.23 Estimated evidence values including that for the 1 hidden node ANN.

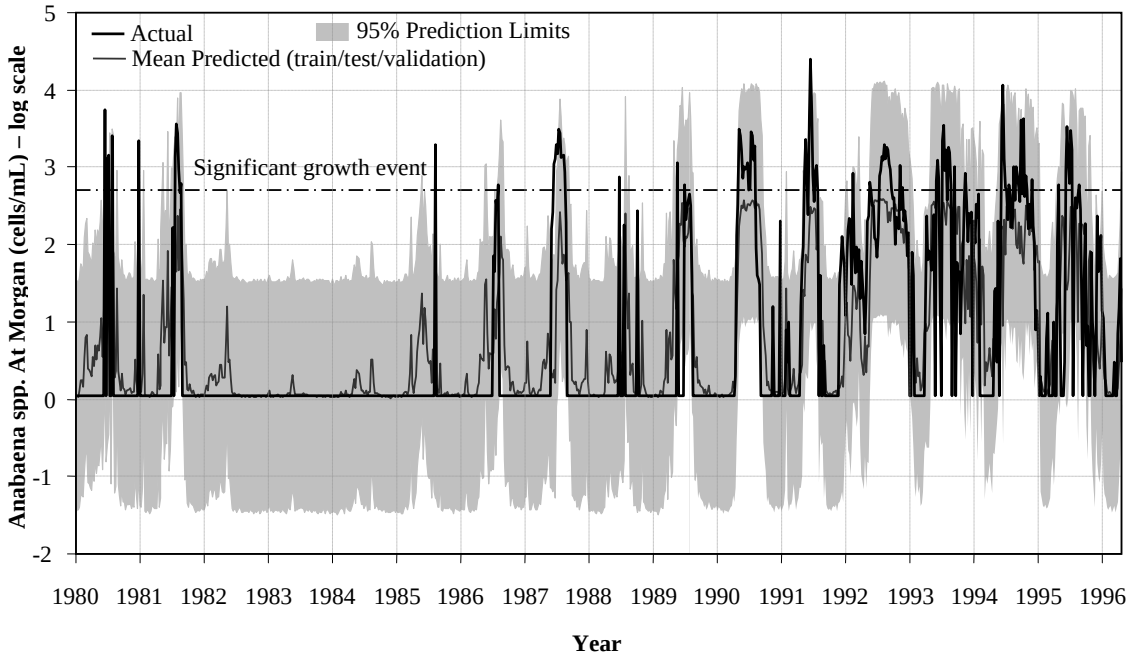

Figure 6.24 Time series plot of the mean 4-week *Anabaena* spp. forecasts and 95% prediction limits obtained for all data with the 1 hidden node ANN model.

events (i.e. concentration $>$ 500 cells/mL) in each year. However, while the duration of growth events were also forecast well, the predicted onset of these events sometimes slightly lagged the actual onset. Shown in Figure 6.25 are the mean forecasts and 95% prediction limits for the independent validation data only. This plot also shows that the model was able to forecast peak, onset and duration of growth events reasonably well, with predicted onset sometimes lagging actual onset.

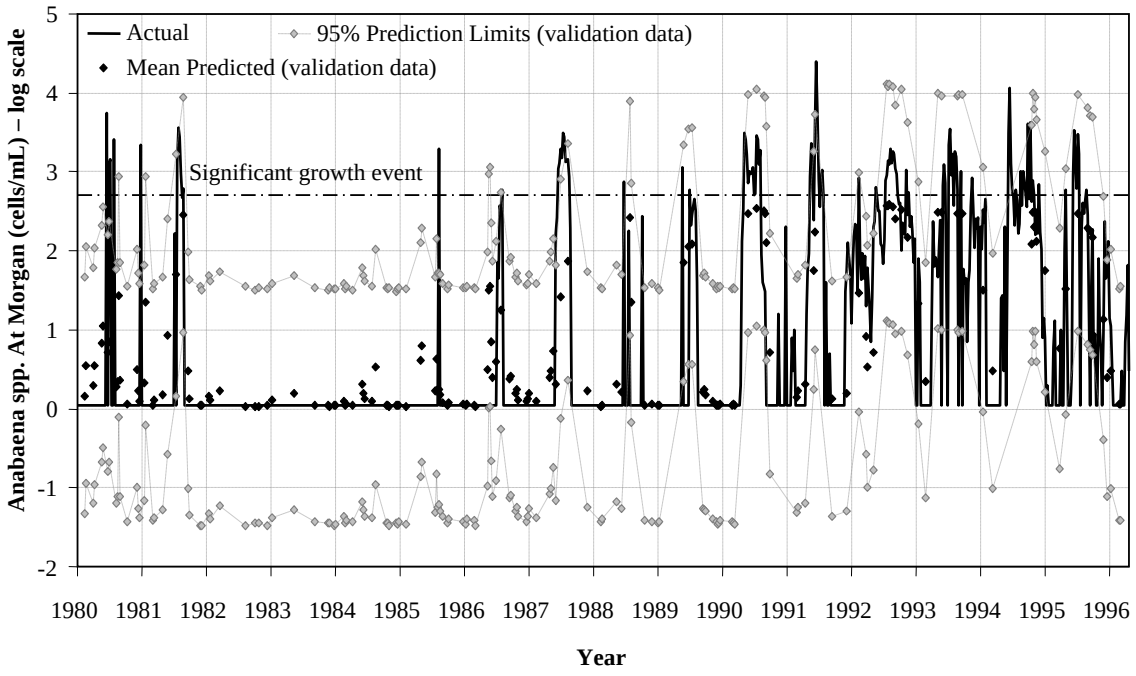

Figure 6.25 Time series plot of the mean 4-week *Anabaena* spp. forecasts and 95% prediction limits obtained for validation data with the 1 hidden node ANN model.

Presented in Table 6.7 are the RMSE values obtained based on the mean forecasts for the training, testing and validation subsets, in comparison to those obtained using the 1 hidden node deterministic ANN and the 10 hidden node ANN developed by *Bowden* (2003). It can be seen that the mean performance of the Bayesian ANN was slightly worse than that of the deterministic model developed in this research. This was not surprising given the width of the prediction limits and the fact that the mean forecasts account for this entire range. However, overall, the usability of the forecasts obtained using the Bayesian ANN is considerably greater than the deterministic forecasts, as the model was able to successfully indicate the occurrence (and non-occurrence) of all significant growth events. Furthermore, the width of the prediction limits provides information about the level of confidence that should be placed in the mean forecasts, and in this case, suggest that these

Table 6.7 RMSEs for the 1 hidden node deterministic and Bayesian ANN *Anabaena* spp. forecasting models, in comparison to *Bowden* (2003) results.

Model	Train	Test	Validation
1 hidden node ANN (deterministic)	0.746	0.790	0.809
1 hidden node ANN (Bayesian)	0.747	0.790	0.812
10 hidden node ANN *Bowden* (2003)	0.733	0.699	0.852

forecasts should be relied upon with caution.

The RI distributions for each input, calculated by applying the modified Connection Weight Approach to the weight vectors sampled from the posterior distribution, are shown in Figure 6.26. These distributions give very similar results to the single-valued RI estimates obtained from the weights of the deterministic ANN. By accounting for the entire range of plausible weights (those that provided a good fit to the data), a positive relationship between recent flow conditions and growth of *Anabaena* spp. was still modelled. Therefore, the discussion given in Section 6.5.2 regarding the usefulness of the ANN model as a means for investigating the response of *Anabaena* spp. to different preventative management strategies also applies to the Bayesian ANN model developed. Even though the usefulness of the Bayesian model was an improvement over that of the deterministic model in terms of providing forecasts of the occurrence of *Anabaena* spp. at Morgan, the model was still found to be insufficient as a hypothesis testing tool. In the attempt to develop an ANN model that can be used as such a tool, it is hypothesised that the silica input could be left out of the model, in which case, the inverse relationship between *Anabaena* spp. and current flow may be modelled correctly. While silica was found to be the most important predictor of *Anabaena* occurrence, it is also positively correlated with flow. Therefore, it is possible that the model is partially accounting for the inverse relationship between current flow conditions and *Anabaena* occurrence through the inverse relationship modelled between current silica concentration and *Anabaena* spp. However, it is also acknowledged that, by omitting the input variable found to provide the greatest amount of information about the incidence of *Anabaena* spp., the predictive performance of the resulting ANN would likely be diminished. Such an investigation is beyond the scope of this thesis, as its main purpose is the development and testing of an improved methodology for ANN development and implementation.

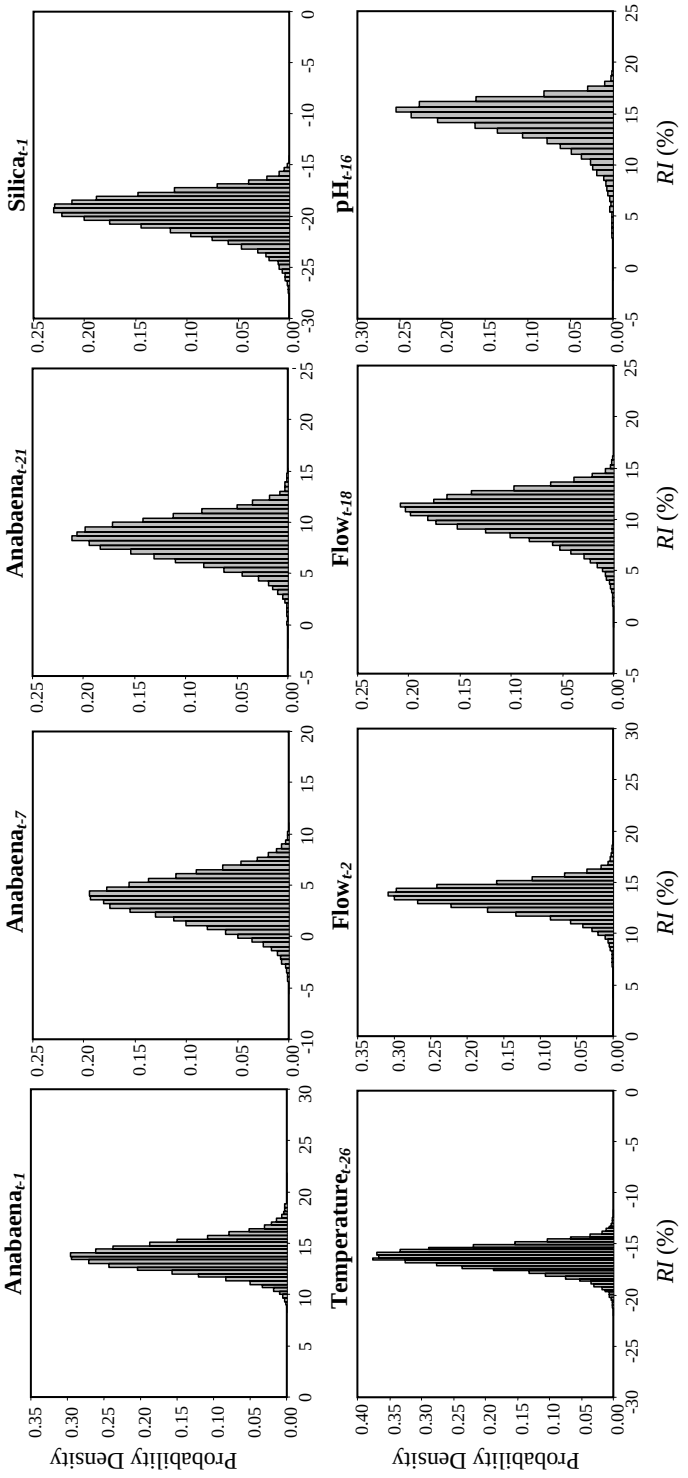

Figure 6.26 *RI* distributions for *Anabaena* spp. inputs.

6.6 CONCLUSIONS

There are generally two main goals associated with model development: (1) to provide predictions of a system response, and/or (2) to gain understanding about the system. However, these two goals place emphasis on different parts of the modelling procedure. If the goal is to predict, then emphasis is placed on the predictive accuracy of the model. On the other hand, if the goal is to gain understanding, emphasis is placed on finding the smallest model able to adequately describe the data and examine the relationship modelled (*Omlin and Reichert*, 1999). In this research, and previous research conducted using this case study, the aim was to develop an ANN model with both of these goals in mind. However, unlike previous studies, the usefulness of the models developed in this research for meeting both of these objectives was checked, rather than assumed. The models were developed using purely data-driven methods and, given that there was a significant amount of noise associated with the available data and that these data were not collected with the purposes of this study in mind, it was not surprising that the resulting models performed neither task well. However, using Bayesian methods, the usefulness of the resulting ANN model was considerably greater than that of the deterministic model, purely because it gave an indication of the high level of uncertainty associated with the model forecasts and was, therefore, able to forecast the possibility of significant growth events of *Anabaena* spp., whereas the deterministic model was not. It is hypothesised that the usefulness of the models developed, in terms of investigating the response of *Anabaena* spp. to different flow management strategies, may be increased if the silica input is omitted, which may result in the inverse relationship between current flow conditions and *Anabaena* spp. being modelled correctly. However, this would also likely result in a reduction in predictive performance. Therefore, it may be better to develop two different models, each with its own distinct goal, rather than attempting to achieve both goals at once and resulting in a model that does neither well.