19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Coupled Bayesian Networks and recursive partitioning method for wetland ecological modelling

B. Fu [a], C. A. Pollino [b], W. Merritt [a] and S. Capon [c]

[a] *Integrated Catchment Assessment and Management (iCAM) Centre, Fenner School of Environment and Society, Australian National University, ACT, Australia. Email: baihua.fu@anu.edu.au*
[b] *CSIRO Land and Water, Canberra, ACT, Australia*
[c] *Griffith University, Nathan, Queensland, Australia*

Abstract: Bayesian Networks (BNs) are increasingly recognised as a useful tool for ecological modelling due to their ability to incorporate a broad range of data types and explicit representation of uncertainty through the use of probabilities. They allow considerable flexibility with respect to the detail and focus of the models allowing conceptually simple habitat condition models or more complex mechanistic representation of ecological response. However, BN outputs are sensitive to the model structure, particularly the selection and linking of variables and how the states are defined for each variable. In this paper, we use recursive partitioning to inform the configuration of the structure of BNs, used for ecological response modelling.

The Narran Lakes Ecosystem is a nationally important floodplain wetland complex considered at risk. In 2005, a mesocosm experiment was conducted as part of the Narran Lake Ecosystem Project to investigate the seedbanks of ephemeral plant communities in response to a range of flood scenarios. From that data and existing literature, a conceptual model describing the major hydrological influences on the ephemeral herbfields was generated. BN ecological response models were then constructed using these data. The models were developed using information generated from recursive partitioning (regression tree and random forests) and BN learning approaches (Figure 1). The models have been incorporated into an environmental flow Decision Support System (DSS), IBIS DSS. The BN model is linked to a hydrological model of the Narran Lakes allowing the modelling of ecological response to flow series.

Recursive partitioning analyses of biophysical and ecological data informed the development of BNs in two ways. Firstly, random forests analyses were used to identify important predictor variables: those variables that, statistically, best explain the ecological responses. Secondly, thresholds were identified using decision tree analysis to reduce subjectivity in the discretisation of variables in the BNs.

Using the coupled BNs and recursive partitioning method improved the rigour and certainties associated with state discretisation, and allowed us to refine the choice of model variables, while maintaining the advantages associated with applying BNs within the DSS.

Keywords: *Bayesian Networks, recursive partitioning, ecological modelling*

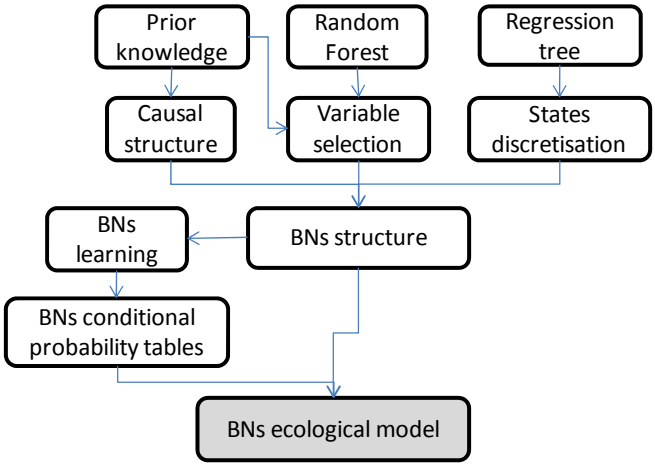

Figure 1. Framework for coupling Bayesian Networks and recursive partitioning (BN-RP).

1. INTRODUCTION

Healthy wetlands support important habitats for diverse fauna and flora. Many wetlands are increasingly under threat from climate change and human activities. Many rivers in Australia have provisions for environmental flows to supply water for maintaining river and wetland ecosystems. A decision support system (DSS) tool, IBIS DSS (Merritt et al., 2009), was developed for the NSW Office of Environment and Heritage (OEH) to assist environmental water managers on the duration, timing, and quantity of environmental flows. The Narran Lakes – being a nationally and internationally important wetland system considered at risk – was selected as one of the five key focus areas. A central part of the DSS is an ecological model to link hydrological regime with ecological response such as vegetation composition and biomass.

Bayesian Networks (BNs) are increasingly recognised as a useful approach for linking hydrological and ecological information, where uncertainties and variability in outcomes can be expressed (Overton et al., 2009; Arthington et al., 2010). A BN is a probabilistic graphical model that is composed of: i) a set of variables represented by nodes in the network; ii) a set of directed links that connect pairs of nodes; and iii) conditional probability tables that quantifies the relationships between the nodes. Compared to many other modelling approaches such as coupled models, system dynamics and agent based models, BNs have the advantages of incorporating different sources of knowledge, quantifying uncertainties, and promoting social learning (Uusitalo, 2007). However, defining a sound model structure in BNs is complicated when our knowledge on the causal links between the interacting variables is limited – a typical situation in wetland ecology. Structure learning algorithms can help explore alternate structures of the network but they generally rely on large datasets (Korb and Nicholson, 2010). Another critical component of BN development is the definition of the states of variables. Many environmental data are continuous variables and these need to be discretised in BNs. Finding a discretisation can be critical for the modelling outcomes; but there is no satisfactory automatic methods developed for BNs (Uusitalo, 2007).

Recursive binary-partitioning is a commonly used statistical method for nonparametric regression and classification that has been widely applied for ecological data (De'Ath and Fabricius, 2000). It generates a decision tree that classifies a response variable based on predictors; in doing so thresholds can be identified to best separate the response variable. Decision trees can identify critical predictors for classification but quite often different trees can describe the same data, especially when many predictors and continuous variables are involved. This can be overcome by using the random forests algorithm. This algorithm was developed by Breiman (2001) as an ensemble classifier. It generates many regression trees and aggregates the results. It is believed to provide a more accurate prediction than a single tree, and can provide more robust prediction compared with many other methods for problems with a large number of variables, nonlinearity and complex interactions (e.g. Knudby et al., 2010; Pino-Mejías et al., 2010). It can provide variable-importance measures by modelling the increase in prediction errors.

In this paper, we develop a coupled BNs and recursive partitioning (BN-RP) method for ecological modelling. Variable-importance measures from random forests and the thresholds identified from regression tree classification are used to assist the structuring of the BNs. In doing so, the structure of the model is statistically more meaningful, whilst the advantages of BNs, such as incorporating existing scientific knowledge and uncertainties, are maintained. A BN model was constructed for the ephemeral vegetation, based on the microcosm experiments conducted as part of the Narran Ecosystem Project (Thoms et al., 2007). The experimental information is used to detect changes in ephemeral vegetation in the Lakes system, using biomass and community types given different inundation regimes.

2. THE NARRAN LAKES

The Narran Lakes Ecosystem (Figure 2) is located within the Condamine Balonne catchment, and is part of the Lower Balonne floodplain region. It is a floodplain wetland complex which consists of four relatively distinct topographical components: Clear Lake, Back Lake and Long Arm in the north, comprising the 'Northern Lakes'; Narran Lake in the south; a large flood-plain area throughout; and a complex network of river channels that dissect the floodplain (Adams and Tyson, 2004). The lakes have a combined surface area of 131.1 km^2, and the channel networks have a combined length of 804.5 km (Thoms et al., 2002). The Narran Lakes Nature Reserve encompasses the area surrounding Back and Clear Lakes, and is listed under the Ramsar Convention. The lakes are important refugia for waterbirds and the success of waterbird breeding in the Narran Lakes region is highly dependent on water inundation of the floodplain and the lakes (Thoms et al., 2007). Any decline in flow volumes is likely to have a negative effect on waterbirds.

The climate of the Narran Ecosystem is semi-arid with an average annual rainfall of 358–425 mm (McGann et al., 2001). The hydrology of the lakes is dominated by flows from the Narran River, local rainfall and evaporation (Thoms et al., 2002). Flooding in the region will generally occur between January and February, and May to June, whilst low flow periods usually occur from September to October (Thoms et al., 2002). Due to recent water resource development, hydrological connections between river channels and associated floodplains have been significantly reduced (Thoms et al., 2002).

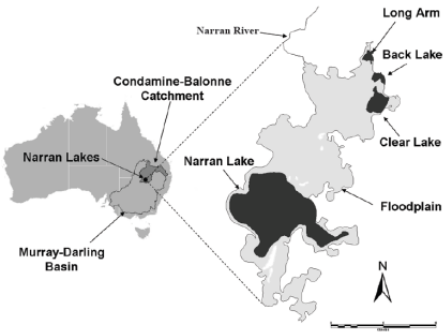

Figure 2. Map of the Narran Lakes Ecosystem, Australia (eWaterCRC, 2007)

3. METHODS

3.1. Data collection

As detailed in Thoms et al. (2008) a mesocosm experiment was conducted during 2005 as part of the Narran Lakes Ecosystem Project to investigate the response of ephemeral plant communities emerging from soil seed banks to a range of flood scenarios. The objectives of this experiment were to determine the role of flood pulse characteristics (i.e. duration, timing, frequency and rate of drawdown) on the productivity and diversity of ephemeral plant communities, as well as the relative influence of long-term flood history in terms of the initial floristic composition of soil seed banks.

Soils were collected from 24 sites across the study area and stratified across the northern and southern regions (12 in each) as well as 4 broad flood frequency classes within each of these regions. Three replicate sites were sampled in each of the 'region – flood frequency' class combination. Soil from each site was divided between 13 pots and each pot then subjected to one of the following annual flood scenarios chosen to represent sensible combinations of varying flood duration, timing, rate of drawdown and frequency (within the constraints posed by limited time and space):

1. 6SF: 6 month summer flood with fast drawdown
2. 6SS: 6 month summer flood with slow drawdown
3. 3SF: 3 month summer flood with fast drawdown
4. 3SS: 3 month winter flood with slow drawdown
5. 6W: 6 month winter flood
6. 3WF: 3 month winter flood with fast drawdown
7. 3WS: 3 month winter flood with slow drawdown
8. 12: 12 month flood
9. 3S3W: 3 month summer flood with fast drawdown and 3 month winter flood with fast drawdown
10. 6 months submerged
11. 3 month summer flood with fast drawdown
12. 3 month summer flood with slow drawdown
13. rainfall only (mimicked based on daily rainfall data from Walgett during 2005).

Pots subjected to treatments 1 to 9 were harvested after 12 months while pots under treatments 10 to 13 were harvested after 6 months. During harvest, all plants were removed from pots, counted, identified and reproductive status was noted. Total biomass, above-ground and below-ground biomass were then measured as dry weights for each species.

3.2. Prior knowledge and conceptual model of hydrological influences on ephemeral herbfields

The conceptual model describes the major hydrological influences on the ephemeral herbfields in the Narran Lakes Ecosystem (Figure 3), where the composition and structure of wetland and floodplain herb communities generally reflect recent flood pulse attributes rather than longer-term flood history. Flood pulse attributes, i.e. timing, depth, duration and rate of drawdown, have a direct influence on processes, particularly the development of aquatic plant communities. Vegetation that responds to drawdown will be indirectly influenced via flood stress and soil moisture levels as drying progresses. Flood timing can directly affect plant community composition in this phase since some species have temperature related germination cues.

Flood history will influence the composition and structure of ephemeral herbfields by shaping the abundance, composition and viability of propagule banks (Capon and Brock, 2006; James et al., 2007). Position in the landscape is included in the conceptual model since survey data suggests that communities in areas adjacent to terrestrial systems will comprise a greater proportion of upland species, e.g. chenopod shrubs. This is likely to be especially important for dry phase vegetation.

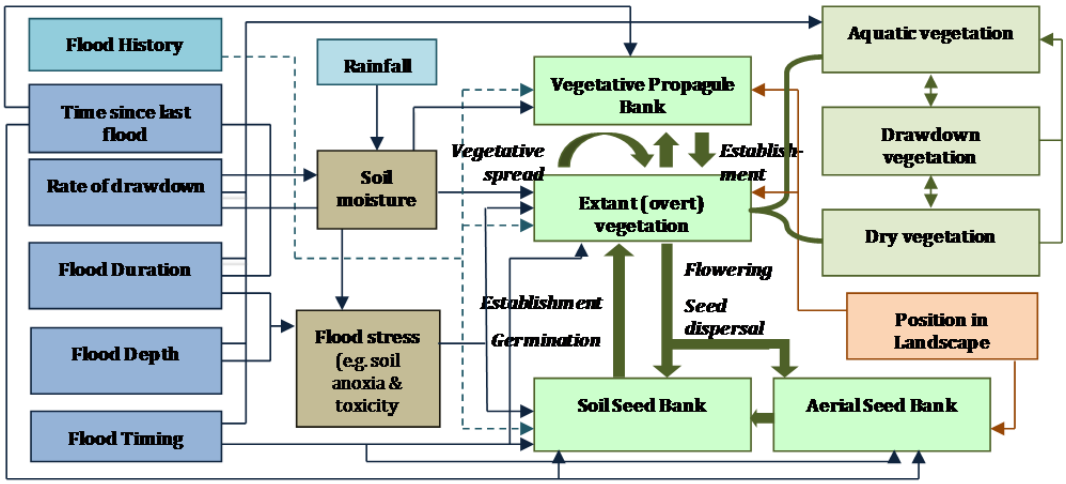

Figure 3. Conceptual model of hydrological influences on ephemeral herbfield communities in the Narran Lakes Ecosystem

3.3. Development of the coupled BN-RP framework

The coupled BN-RP framework (Figure 1) uses recursive partitioning (including regression tree and random forests) to help in defining a BN structure, particularly where there are a large number of interacting and/or continuous variables to consider representing in the network. Variable-importance measures are an output from the random forests analysis, and this assists in identifying the statistically important predictors, whilst the regression trees help quantify thresholds which are statistically meaningful when discretising continuous variables. These types of outcomes from the recursive partitioning analysis are incorporated with existing ecological knowledge to define the BN structure. Conditional probability tables which describe the strength and nature of the relationship between variables are then quantified through BNs data learning algorithm.

Two packages are available in R to perform regression tree analysis: **rpart** (Therneau et al., 2011) and **party** (Hothorn et al., 2011). Using the datasets described in Section 3.1, the response variables considered include total biomass (TBM), percent total biomass below ground (TBM_BG) and above ground (TBM_AG). As a percent of the total biomass, the model also includes annual forbs (AF), annual monocots (AM), perennial forbs (PF), perennial monocots (PM) non-vascular (NV), total forbs (F) and total monocots (M). The predictors are listed in Table 1. These predictors vary in the scales of measurement and numbers of categories. Some predictor variables being considered in the analyses are correlated. To deal with this situation, the *ctree()* and *cforest()* functions in the **party** package can offer statistically unbiased variable selections, representing unbiased tree and variable-importance measures (Hothorn and Zeileis, 2009). Therefore, *cforest()* was used to perform random forests analysis to quantify variable importance. This was followed by using *ctree()* to identify predictor thresholds.

The outcomes of recursive partitioning analysis were used to assist the development of BNs in two ways. Firstly, the important predictors identified through the random forests analysis are those that best explain the response variables statistically. Secondly, the thresholds identified through decision tree analysis provide a guide for discretisation in the BNs. Recursive partitioning can be a useful tool to analyse scientific data, the results of regression tree and random forests analyses (and any results from data mining) must be interpreted in conjunction with scientific knowledge in order to reach a sensible outcome. In light of the objective of the IBIS DSS project, i.e. to assist environmental flow management, the predictors were selected using the following criteria:

- some or all of the predictors are related to flood regime;
- the predictors are consistent with ecological knowledge;
- the predictors have relatively high variable-importance values.

Table 1. Predictor variables used for recursive partitioning analysis.

Variable name	Description	States
Elevation	Elevation	Numeric, ranges from 118.1 to 122.6 m.
Final.Class	Flood frequency: the classification of sites is by region and broad flood frequency (based on the number of years inundated by largest annual event between 1988 and 2004)	Categorical, include NF (north region frequently flooded), NI (north region infrequently flooded), SF (south region frequently flooded) and SI (south region infrequently flooded). Frequently flooded sites were inundated in more than 12 out of 16 years.
Duration	Flood duration during treatment	Numeric, ranges from 3 to 12 months.
Ddown	Water drawdown rate during treatment	Categorical, includes F (fast), S (slow) and N (no drawdown).
Timing	Flood timing during treatment	Categorical, includes S (spring), W (winter), and SW (spring and winter).
Frequency	Flood frequency during treatment	Numeric, includes 1 (once) and 2 (twice).
RI	Annual recurrence interval	Numeric, ranges from 1.259 to 17.
Sand	Percentage of sand	Numeric, ranges from 3.82 to 26.39 %.
Clay	Percentage of clay	Numeric, ranges from 11.54 to 34.32 %.
pH	pH	Numeric, ranges from 7 to 9.5.
Org	Organic matter	Numeric, ranges from 3.11 to 19.61.
Ca	Calcium concentration	Numeric, ranges from 0.54 to 2.3 mg/kg.
K	Potassium concentration	Numeric, ranges from 1 to 1.85 mg/kg.
Mg	Magnesium concentration	Numeric, ranges from 0.64 to 1.14 mg/kg.
Mn	Manganese concentration	Numeric, ranges from 249 to 560 mg/kg.
Na	Sodium concentration	Numeric, ranges from 0.12 to 0.3 mg/kg.
P	Phosphorus concentration	Numeric, ranges from 430 to 850 mg/kg.
N	Nitrogen concentration	Numeric, ranges from 560 to 4640 mg/kg.

The BN software program Netica (www.norsys.com) was used to generate probabilities.

4. RESULTS

The variable-importance measures from random forests analysis are illustrated in a heat map (Figure 4); the brighter the colour, the higher the variable-importance measures. Note that the colour is scaled for the same response variable (i.e. each row). The relative importance of predictor variables varies depending on the response variables. For example, flood timing (Timing) is found to be the most important variable in classifying the proportion of above ground or below ground biomass (%TBM_AG and %TBM_BG). However, for the classification of non-vascular biomass proportion (%TBM_NV), flood duration (Duration) and water drawdown rate (Ddown) become more important. In general, variables that are related to flooding regime such as flood timing, final class (a combination of location and flood frequency), and water drawdown rate were found to have higher variable-importance measures than those related to soil properties such as soil texture and chemistry. An exception is phosphorus concentration which is found to be an important predictor for ephemeral biomass.

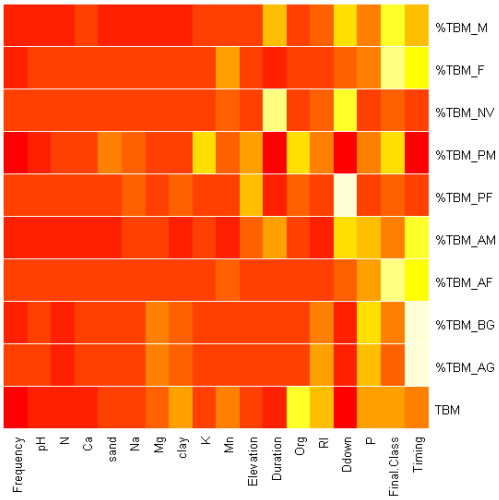

Figure 4. Variable-importance measures for response variables related to ephemeral biomass. (TBM: Total biomass; AG: above-ground; BG: below ground; AF/PF: annual/perennial forbs; AM/PM: annual/perennial monocots; NV: non-vascular).

Although random forests provide variable-importance measures allowing the elemination of insignificant predictors, this analysis does not identify specific thresholds. To do this, we used decision tree analysis with selected variables that has been identifed important in the random forests anlaysis. The regression tree analysis for the proportion of total forbs in total biomass is shown in Figure 5. North region frequently flooded (Final.Class=NF) samples have a higher proportion of total forbs compared to other regions (median total forbs proportion = 13%). In the NF region, samples that were treated with winter flood have much lower total forbs (median total forbs proportion = 11%) than those flooded in spring (Timing=S)

or spring and winter (Timing=SW). A flood duration of 3 months was identified as the critical threshold to separate the proportion of forbs in total biomass: those flooded less than 3 months (Duration ≤ 3) have a median of 92% total forbs while those flooded more than 3 months (Duration > 3) have a median of 56%. The distributions of the response variables in each class (i.e. the box plots in Figure 5) provide valuable information on how to best discretise response variables to maximise model sensitivity. For the example provided in Figure 5 reasonable thresholds to discretise the proportion of total forbs are: below 20%, 20-50%, 50-90% and above 90%.

The BN model for ephemeral vegetation (Figure 6) was constructed with consideration of the purpose of the model (to assist environmental flow management), existing scientific knowledge on how ephemeral vegetation should response to environmental variables, the statistically important variables, and the thresholds which could best separate the response variables. The input nodes of the BN model were selected, and the states of the nodes were discretised with the consideration from the outcomes of recursive partitioning to maximise model sensitivity. This model is implemented the IBIS DSS which links hydraulic behaviour in the Narran Lakes to ecological responses. The IBIS DSS has been designed to explore the likely outcomes of climate and water planning scenarios on ecological characteristics of the Narran Lakes and thus support the OEH plan and manage flows at wetland and valley scales over short term (annual) to long term (decadal) planning scales. Although this model was structured to use the best of the existing data, it is still limited by the uncertainty associated with data itself, which is derived from laboratory experiments for samples from a limited number of sites in a highly variable ecosystem. Validation using field data must be undertaken to test the model before it can be used to assist environmental flow management.

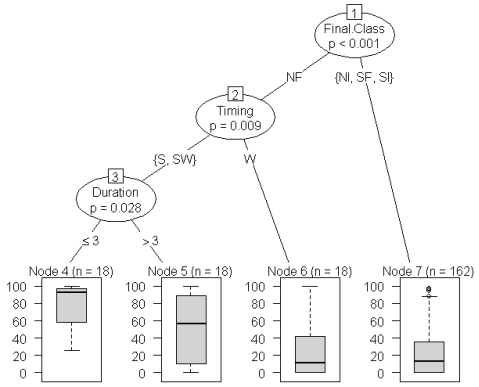

Figure 5. Regression tree for the proportion of forbs in total biomass.

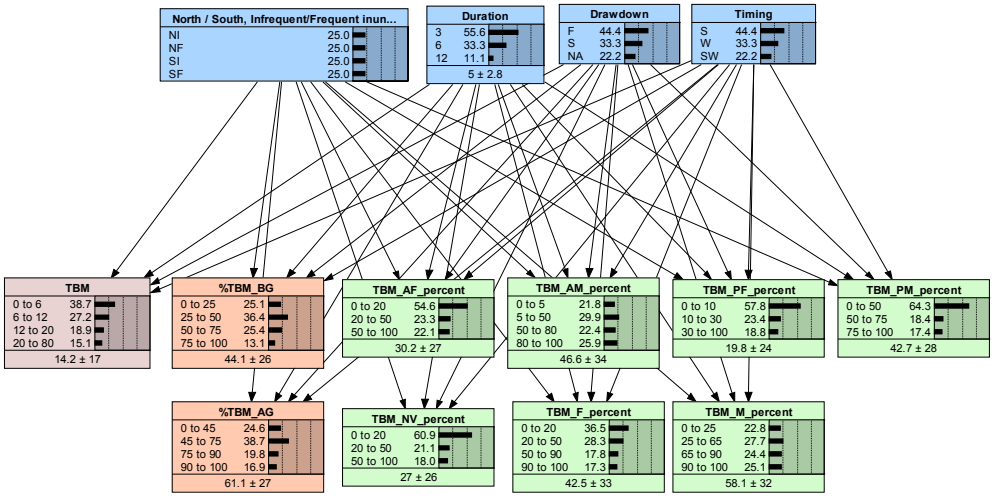

Figure 6. Bayesian network for inundation and vegetation (ephemeral / herbs) variables

5. CONCLUSION

The coupled Bayesian Networks and recursive partitioning method presented here provides a framework to help better structure BNs through the integration of prior knowledge and classification exercise. The model represents an appropriate selection of variables and states discretisation whilst maintaining the strengths of BNs, namely: explicit representation of uncertainty, the capacity to use a broad range of data to populate the network, and representation of ecological processes at a level appropriate to the amount of available data. This method is especially helpful in reducing the complexity of the BNs by eliminating non-critical variables,

and in increasing sensitivity of the model. The model represents a synergy of learning from the recursive partitioning analysis and prior knowledge on ecological response to flooding.

ACKNOWLEDGMENTS

This project was funded through the Rivers Environmental Restoration Program (RERP) which is supported by the NSW Government and the Australian Government's Water for the Future - Water Smart Australia Program.

REFERENCES

Adams, J. and Tyson, D. (2004), Local Knowledge of the Narran Lakes: Oral History as a Line of Evidence in Ecological Understanding, In *Fourth Annual Australian Stream Management Conference,* Launceston.

Arthington, A.H., Naiman, R.J., McClain, M.E. and Nilsson, C. (2010), Preserving the biodiversity and ecological services of rivers: new challenges and research opportunities *Freshwater Biology,* 55, 1-16.

Breiman, L. (2001), Random Forests, *Machine Learning,* 45, 5-32.

Capon, S.J. and Brock, M.A. (2006), Flooding, soil seed bank dynamics and vegetation resilience of a hydrologically variable desert floodplain, *Freshwater Biology,* 51, 206-223.

De'Ath, G. and Fabricius, K.E. (2000), Classification and regression trees: a powerful yet simple technique for ecological data analysis, *Ecology,* 81, 3178-3192.

eWaterCRC (2007), Narran Maps, URL:http:\\www.canberra.edu.au\\centres\\narran\\docs\\resources\\maps\\maps of Narran.pdf 23rd March, 2010, University of Canberra.

Hothorn, S.T. and Zeileis, A. (2009), Party on! A new conditional variable-importance measure for random forests available in the party package, *The R Journal,* 1, 14-17.

Hothorn, T., Hornik, K., Strobl, C. and Zeileis, A. (2011), *party: A Laboratory for Recursive Partytioning. R package version 0.9,* [Online]: http://cran.r-project.org/package=party.

James, C., Capon, S., White, M., Rayburg, S. and Thoms, M. (2007), Spatial variability of the soil seed bank in a heterogeneous ephemeral wetland system in semi-arid Australia, *Plant Ecology,* 190, 205-217.

Knudby, A., Brenning, A. and LeDrew, E. (2010), New approaches to modelling fish-habitat relationships, *Ecological Modelling,* 221, 503-511.

Korb, K.B. and Nicholson, A.E. (2010), *Bayesian Artificial Intelligence, 2nd Edition,* Chapman & Hall/CRC.

McGann, T.D., Kingswood, R. and Bell, D. (2001), Vegetation of Narran Lake Nature Reserve, North Western Plains, New South Wales, *Cunninghamia,* 7, 43-64.

Merritt, W.S., Pollino, C., S., P. and Jakeman, A.J. (2009), Integrating hydrology and ecology models into flexible and adaptive decision support tools: the IBIS DSS, In: *Proceedings of the International Congress on Modelling and Simulation MODSIM 2009, Cairns, July 2009*, Modelling and Simulation Society of Australia and New Zealand.

Overton, I.C., Colloff, M., Doody, T.M., Henderson, B. and Cuddy, S.M. (2009), *Ecological Outcomes of Flow Regimes in the Murray-Darling Basin,* CSIRO for the NWC, Australia.

Pino-Mejías, R., Cubiles-de-la-Vega, M.D., Anaya-Romero, M., Pascual-Acosta, A., Jordán-López, A. and Bellinfante-Crocci, N. (2010), Predicting the potential habitat of oaks with data mining models and the R system, *Environmental Modelling & Software,* 25, 826-836.

Therneau, T.M., Atkinson, B. and R port by Brian Ripley (2011), *rpart: Recursive Partitioning. R package version 3.1-50. ,* [Online]: http://cran.r-project.org/package=rpart.

Thoms, M., Capon, S., James, C., Padgham, M. and Rayburg, S. (2007), *The Response of a terminal wetland system to variable wetting and drying,* Final report to the Murray-Darling Basin Commission, Narran Ecosystem Project, Murray-Darling Basin Commission, Canberra.

Thoms, M., Capon, S., James, C., Padgham, M. and Rayburg, S. (2008), *The Narran Ecosystem Project: the response of a terminal wetland system to variable wetting and drying. Final report to the Murray-Darling Basin Commission,* Murray-Darling Basin Commission, Canberra.

Thoms, M., Quinn, G., Butcher, R., Phillips, B., Wilson, G., Brock, M. and Gawne, B. (2002), *Scoping study for the Narran Lakes and lower Balonne floodpain management study,* CRCFE technical report 3/2002, CRC for Freshwater Ecology, Canberra.

Uusitalo, L. (2007), Advantages and challenges of Bayesian networks in environmental modelling, *Ecological Modelling,* 203, 312-318.