Jul 1st, 12:00 AM

Overcoming data constraints to create meaningful ecological models

Rebecca E. Lester

Peter G. Fairweather

Overcoming data constraints to create meaningful ecological models

Rebecca E Lester & **Peter G Fairweather**
School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide SA 5001,
Australia (rebecca.lester@flinders.edu.au & peter.fairweather@flinders.edu.au)

Abstract: Many techniques used to model ecosystems cannot be meaningfully applied to large-scale ecological problems due to data constraints. Disparate collection methods, data types and incomplete data sets, or limited theoretical understanding mean that a wide range of modelling techniques used to model physical processes or for problems specific to species or populations cannot be used at an ecosystem scale. In developing an ecological response model for the Coorong, a South Australian hypersaline estuary, we combined several flexible modelling approaches in a statistical framework to develop an approach we call 'ecosystem states'. This model uses simulated hydrodynamic conditions as input to predict one of a suite of states per space and time, allowing prediction of likely ecological conditions under a variety of scenarios. Each ecosystem state has defined sets of biota and physico-chemical parameters. The existing model is limited in that its predictions have yet to be tested and, as yet, no spatial or temporal connectivity has been incorporated into simulated time series of ecosystem states. This approach can be used in a wide range of ecosystems, where enough data are available to model ecosystem states. We are in the process of applying the technique to a nearby lake system. This has been more difficult than for the Coorong as there is little overlap in the spatial and temporal coverage of biological data sets for that region. The approach is robust to low-quality biological data and missing environmental data, so should suit situations where community or management monitoring programs have occurred through time.

Keywords: data limitations, ecological response models, statistical modelling

1. INTRODUCTION

Ecological modelling often attempts to simplify ecological systems for the purposes of better understanding the interactions between components, or to assist in the better management of particular ecosystems [Otto & Day 2007]. In aquatic ecosystems, common management-related uses for ecological modelling include determining the impact of altered hydrology (including extractions and dams) or environmental flows, or assessing habitat suitability for particular species. Another increasingly-common use for ecological modelling of aquatic ecosystems is to determine the likely impact of climate change and discern the most appropriate mitigation or adaptation actions.

The methods used for ecological modelling differ, in many cases from those used for other types of modelling in aquatic systems. Unlike hydrodynamic or biogeochemical modelling, ecological modelling is often severely limited by the theoretical understanding of linkages between the components of the ecosystem. Thus, fully deterministic models, which are commonly used in other disciplines, where mathematical relationships describe interactions between ecosystem components [Otto & Day 2007], are often impractical [Wikle 2003]. Some solutions to these challenges in the past have been to develop simplified models of just one aspect of interest or to look to non-deterministic solutions including statistical modelling approaches (e.g. Wikle [2003]) and approaches based on expert opinion or the available literature.

The available data also often constrains the choice of modelling technique for many ecosystem-level tasks. As is the often the case, the amount of data available for ecological

modelling is usually limited. The cost of ecological monitoring and other data collection mean that this is particularly the case for biological components of an ecosystem, especially where remotely-sensed or logged measurements are not possible to collect or are limited to a few variables, e.g. seal tagging for seawater conductivity, temperature and depth data. Differences in methodologies for monitoring between taxa mean that data are collected using disparate methods that may not be directly comparable (e.g. abundances are often reported for birds as counts, while fisheries data are often expressed as the biomass of catch per unit effort), and available data can be a mixture of quantitative, semi-quantitative and qualitative variables. Added to this, there is often limited replication in space and/or time, resulting in a 'patchwork' of available data, with little overlap across studies or taxa with a mixture of spatial and temporal scales of variation at play. Furthermore, the available variables will often not cover all of the parameters of interest. Thus, many modelling approaches that work well in other contexts are simply not applicable due to differences between their data requirements and the available data for ecological systems. In the past, attempts to circumvent these limitations or simply ignore them have resulted in models that have uncertain predictive capabilities and/or have limited transferability to other situations.

The aim of this paper is to illustrate how some of these modelling techniques can be applied to ecological data in a meaningful and applied manner, despite the many data limitations that are common with ecological data sets. We aim to highlight common limitations associated with ecological data sets and some of the methods that we have found effective in dealing with these data that may not be commonly applied in other disciplines. However, there are almost certainly cases in all disciplines where the constraints we describe are relevant, and the methods that are outlined here may be of use. Therefore, we describe the sequence of our modelling technique and then the construction of an ecological response model for a case study where the task was to develop a single ecosystem-level model that could be used to evaluate competing management actions in the light of various future climate-change scenarios. We then outline the advantages and limitations of our application of these modelling techniques, including identifying the types of data sets that are needed to apply this methodology and potential future enhancements of the work done to date.

2. METHODS

2.1 Study Region

The Coorong is the estuary for the largest river system in Australia; the Murray-Darling Basin (see Lester & Fairweather [2009a] for a map). It is a lagoonal complex that stretches approximately 110 km along the South Australian coast and is divided from two large freshwater lakes immediately upstream, and the river itself, by a series of artificial barrages than control freshwater flow into the estuary. The system is an inverse estuary, with a single connection to the ocean at the north-western end of the system, the same end as where the majority of freshwater flows enter the system. Thus, a natural gradient, from estuarine conditions and ecosystems in the north-west to being hypersaline in the south-east exists, although limited inflows of irrigation drainage also occur to the south-east. The region is renowned for its ecological diversity (e.g. through its listing under the Ramsar Convention for Wetlands of International Importance in 1985; Brookes et al. [2009]) but also for its recent ecological decline as a result of ongoing drought and unsustainable upstream diversions of water [Brookes et al. 2009].

The Coorong is a complex system, with estuarine, marine and hypersaline components. Some aspects of the ecosystem are well-understood and well-described; surface hydrology, bird populations, fisheries biology, for example [Webster 2007, Brookes et al. 2009]; but others are less well-understood (e.g. links with groundwater, biogeochemistry, pelagic invertebrates). Thus, a deterministic model would need to be limited to those well-studied components or another approach is necessary. In this instance, we chose to use a statistical modelling approach in order to use as much of the available data as possible.

Data sets for the Coorong were provided by a range of institutions and individuals and their provenance is given in Lester & Fairweather [2009a, b]. While the temporal and spatial coverage of these data sets varied significantly, there was sufficient overlap across the data

sets to enable the ecological response model to be constructed at an annual time step between 1999 and 2007 (which was the most recent data at the time of modelling) and at a quarterly time step between 2005 and 2007.

2.2 Development of the Coorong model

The overall level of complexity and our incomplete understanding of the system, and incomplete data sets meant that a deterministic model for the region was simply not feasible. Instead, statistical modelling techniques were used to develop an ecological response model for the region. State-and-transition models and classification and regression trees (CART) were two techniques likely to be of use for this case study [Lester & Fairweather 2008].

The development of the model (including relevant references and the details of the data sets used) has been described elsewhere [Lester & Fairweather 2009a, b, 2010], so it will only be summarised here (Figure 1). The available data were divided into two sets; one containing biological data and the other physico-chemical (or 'environmental') data. Firstly, the biological data was cluster-analysed to identify groups of fish, birds, macroinvertebrates and macrophytes occurring together in space and time. Cluster membership based on biota was then used as the response variable for a CART analysis of the environmental data set. CART analyses sequentially partition a response variable (i.e. cluster membership) in order to maximise the differences between groups by identifying thresholds in splitting variables drawn from a set of associated independent variables (i.e. the environmental data set). The biological distinctness of cases within each terminal node from CART was tested using analysis of similarities (ANOSIM) tests [Clarke & Gorley 2006], which test for significant differences between grouping factors in multivariate data sets, and those terminal nodes that did not support statistically-distinct biota were combined. Cases that had been excluded from the process to date (due to missing biological data excluding them from the cluster analysis step) were assigned to a terminal node based on their environmental data. The biological data of these new cases was then compared to that of the original cases within each terminal node (or 'ecosystem state') to test the predictive capacity of the model.

The quarterly time step was included in the development of the model because the annual model did not capture recent declines in ecological condition. The two models (annual and quarterly) were combined by determining which of the end ecosystem states were statistically distinct (using ANOSIM) in terms of co-occurring biota and then re-running the CART analysis and subsequent steps in the process. Once a sufficiently predictive model had been identified, the physico-chemical and biological characteristics of each ecosystem state were described.

Information regarding the validation and verification of the model is presented in Lester & Fairweather [2009a]. This model has been applied to investigate a range of management and climate-change scenarios for the region. These include different predictions for the severity of climate change and actions such as changes to flows to the region, changes in dredging effort at the Murray Mouth and new interventions such as pumping of hypersaline water from the South Lagoon to the ocean, based on output for the hydrodynamic model of the region [Webster 2007] applied to the same scenarios [Lester & Fairweather 2009a].

2.3 Applying the approach to the Lower Lakes

Following the success of this technique in predicting the ecological response to assist managers evaluate competing interventions and strategies for the Coorong, we attempted to apply the same modelling technique to the Murray Lower Lakes (i.e. the two freshwater lakes immediately upstream of the Coorong; Lake Alexandrina and Lake Albert).

While the Lower Lakes are geographically close to the Coorong, they differ in several ways that made this next application of the modelling approach a challenge. Firstly, they are predominantly freshwater (apart from occasional saline intrusion through the barrages),

rather than grading from estuarine to hypersaline. Secondly, they are not configured as a distinct linear gradient in the same fashion as the Coorong. Thirdly, and possibly most importantly, they are not nearly as well-studied as the Coorong, despite their proximity. The implementation of a variety of monitoring programs undertaken by the South Australian and Australian governments as a result of recent large scale ecological degradation has begun to redress this situation but the pool of available data was smaller and so represented an opportunity to test the limitations of the technique in a data-poor environment.

The full ecosystem states model for the Lakes is currently under construction, but this attempted application has been included here to highlight the areas in which the technique is robust to sparse data versus where it is not. This will help to elucidate the conditions in which this approach is likely to be of use in the future, and where it may fail.

3. RESULTS

3.1 Coorong model

Twelve out of the 17 available data sets for the Coorong were able to be incorporated into the development of the ecosystem response model. While not all data sets covered the entire Coorong, or the whole time period used for model calibration, sufficient replication in space and time existed to be of use. The cluster and ANOSIM analyses were most sensitive to missing data (as the dissimilarity matrices upon which they are based are undefined for missing values). For the most part, the spatial replication of biological data sets was relatively good, allowing most data sets to be included on that criterion. Replication in time was less consistent but the majority of data sets had some level of replication, particularly near the end of the time period (e.g. 2005 to 2007). Thus, those data sets that were excluded from the development of the annual model (covering 1999 to 2007) were included in the development of a quarterly model (2005 to 2007).

The methods used on the environmental data sets (e.g. CART) were more robust to missing values (through the use of a penalty for missing data; see Lester & Fairweather [2009a]). Therefore, while data sets were extremely patchy in space and time, all data sets that included some degree of replication could be included in the development of the model, although most variables out of 230 in total did not feature in the final model as a primary splitting variable.

The final ecosystem states model for the Coorong has also been described elsewhere [Lester & Fairweather 2009a, b, 2010]. The model included eight possible ecosystem states that described distinct assemblages of biota under particular physico-chemical conditions. The ecosystem states were differentiated first into two groups by their relative values for the average daily tidal range (threshold of 0.05 m); and then the maximum number of days without barrage flows (threshold of 339 days); then either the average annual water level (thresholds of 0.37 and -0.09 m above the Australian Height Datum) or the average annual water depth from the previous year (threshold of 1.99 m) and the average annual salinity (threshold of 64.5 g L^{-1}), depending on the preceding splits. The resulting ecosystem states were effectively divided into four that were tidally-influenced (> 0.05 m) versus four that were not (≤ 0.05 m) because of average annual tidal range appearing as the first splitting variable (i.e. into a 'marine' basin versus a 'hypersaline' basin). The presence of the number of days without barrage flows as the next variable (for both halves of the classification tree) then divided these two basins further based loosely on their relative health. Given that 11 months without barrage flows occurs <1% of years in hydrological models with no water extractions from the catchment, we believe that those states occurring when this threshold is exceeded (>339 days) represent degraded ecological communities. This view is supported by the shifts in abundance of taxa and overall loss of biodiversity amongst these states [Lester & Fairweather 2009a]. Each of the final states has been named based on these two splitting variables for ease of interpretation: Estuarine/Marine, Marine, Unhealthy Marine, Degraded Marine; Healthy Hypersaline, Average Hypersaline, Unhealthy Hypersaline and Degraded Hypersaline (see Lester & Fairweather [2009a]).

Because of the drivers identified, the CART model could then be used to predict the mix of ecosystem states likely to be present under a wide variety of management and climate change scenarios (e.g. see Lester et al. [2009]), based on simulated hydrodynamic conditions within the Coorong [Webster 2007]. It has been used to demonstrate that no engineering interventions suggested so far (e.g. dredging of the Murray Mouth or diverting additional flows from irrigation drainage in the south-east of South Australia into the southern Coorong) would produce similar ecosystem states to those predicted for years with average flows from the River Murray (e.g. see Lester et al. [2009]).

3.2 Lakes model: identifying data sets most likely to be useful

The process to develop an ecosystem states model for the Lower Lakes is currently ongoing. The largest impediment to this process has been the lack of overlap in spatial or temporal scales across data sets. Only 13 of the 32 available data sets for the Lakes were able to be included in model development. This is largely due to the lack of temporal replication for many data sets, although spatial replication for many is also limited. Thus, of 18 distinct biological data sets contributed by various sources, only eight were able to be used in model constructions (with fish data being a conglomerate of several surveys using presence/absence only). Of the 14 available physico-chemical data, only five were able to be included in the model development (compared with six out of 11 and six out of six datasets being useful for biotic and physico-chemical data in the Coorong, respectively).

Data sets that were of most use in the construction of this model (and that for the Coorong) were those that had repeated measures over both space and time. Concerted short-term efforts to measure a set of variables over a wide area or the opposite, of long-term repeated surveys at a single location, while having other uses, were not particularly constructive for the development of an ecosystem model that was representative of the Lakes (or Coorong) as a whole over time. For the clustering and ANOSIM analyses, only cases without missing values can be included (in the absence of a detailed understanding of how various options for filling missing values [e.g. mean replacement] affect the ecosystem states produced), effectively eliminating most of the potential data sources available for many sites in many years (almost two-thirds of the available data sources).

Until the model development process is complete, it is not possible to determine whether the resultant model will be of as much use for assessing competing management options, nor whether the resultant ecosystem states will be intuitive and robust to the same degree as those defined for the Coorong. Thus, we have yet to draw conclusions relating to whether the significantly smaller pool of available, useful data for constructing the Lakes model is sufficient (but not ideal) to construct a workable ecological response model.

4. DISCUSSION

The Coorong ecosystem states model performed extremely well overall. A leave-one-out analysis indicated that the misclassification rate successfully predicted in excess of 80% of cases for three of the four analyses conducted. The model had the most difficulty in differentiating between the Estuarine/Marine and Unhealthy Marine states at an annual time step based on the biological data. Given that some evidence exists (anecdotal and empirical; see Lester & Fairweather [2009a]), that the Estuarine/Marine state may, in fact, represent more than one healthy marine-basin state (but insufficient data exist to resolve this into sub-states), the variability inherent in that that state, as defined, may be contributing to this misclassification rate. Notwithstanding, the remaining misclassification rates are extremely good for ecological models and suggest that the model has value in predicting the ecological condition of the Coorong under known (or modelled) hydrodynamic conditions.

The statistical modelling approach that we applied here combined cluster analysis, CART and ANOSIM to produce a state-and-transition model. This approach had the advantage of being data-derived, so was not reliant on expert opinion or literature from other locations. Nonetheless, the identified splitting variables were intuitive to local researchers and managers. The combination of CART and ANOSIM was an important step to prevent over-

fitting in the model (see Lester & Fairweather [2009a] for additional detail). Other statistical methods that could have been applied (e.g. logistic regression, multiple adaptive regression splines [Lester & Fairweather 2008]) tend to have more assumptions about the distribution of the data, equality of variances and often rely on fitting linear models. The methods described here do not rely on any of these assumptions, and are also relatively robust to missing data, providing significantly more flexibility in model development. This flexibility is likely to be of use in other disciplines in cases where the available data are less than complete or comprehensive. The methods used here may be less effective than some others, like Bayesian belief networks, in their ability to describe events outside the scope of the available data.

This modelling approach also has a substantial advantage over many other approaches in that the resulting ecosystem states are simple to explain and intuitive for managers and members of the public. The ecosystem states arising from the Coorong model are broadly consistent with assemblages of biota that have been observed together by researchers, managers and the general public (although several states occurred infrequently in the calibration data set and so are less well-defined in their delineation). Thus, acceptance and ownership of the results of the model have been quickly forthcoming, and we have had many requests from managers to model particular scenarios that represent options for the management of the system, both in the short and the long term (> 400 scenarios have been modelled so far). Managers, therefore, now have a relatively robust, data-derived ecological response model that allows them to objectively assess the relative merits of competing management proposals, including under a variety of climate-change and sealevel-rise scenarios (as long as we can derive what the scenarios mean for the driving parameters as inputs to our model). Combinations of management options are also possible and the results of these have been particularly informative. The simple nature of the model means that long-term changes can be assessed in a timely manner (e.g. model runs of more than 100 years for 12 locations along the Coorong; Lester et al. [2009]).

As for all models, there are a number of limitations with the Coorong model that need to be considered when results are interpreted. As yet, the predictive capacity of the model has not been independently assessed. We attempted to compare the predicted ecosystem states to data collected in 2008 with limited success due to the small number of surveys conducted during 2008 for which data were available at the time (see Lester & Fairweather [2009a]). We hope to repeat this analysis in the future when more data become available. Other limitations include assumptions that the trajectory of recovery will be the reverse of the trajectory of decline, that all transitions between all states are possible (both of which almost certainly will not hold), and that there is a lack of spatial and temporal dependence on the states that are possible at any one location. However, in the absence of additional data (particularly for periods of ecological recovery), it is currently impossible to quantify the relative importance of these limitations. Uncertainty also exists with respect to how the crisp transitions described in the model (e.g. a time since flow threshold of 339 days) relate in practice (e.g. is 339.5 days considered to be over the threshold, or does some 'fuzziness' exist?). We are currently attempting to assess this, along with determining whether the model can be used retrospectively to describe past ecological conditions (e.g. in 1985 when the region was listed on the Ramsar Convention) and how robust the model is at describing conditions that differ from those in the training data set.

Attempting to apply the ecosystem states model to a nearby region, with a very different history of data collection, has highlighted some data-availability constraints associated with this modelling approach. The ecosystem states approach is substantially more limited by missing data in biological data sets than in physico-chemical data sets. Therefore, logged or remote-sensed data sets such as water level, flow, or meteorological data can be incorporated easily despite occasional equipment (or other) malfunctions. However, regions that have been poorly, or only occasionally, comprehensively-surveyed for biota are less likely to be able to be modelled appropriately using the ecosystem states approach. As we are still in the process of developing the model for the Lakes, it is currently unclear whether the available data are sufficient or whether they fall below the minimum data requirements. Describing this minimum is an area of ongoing research.

The ecosystem states methodology has been shown to be robust to random errors in original data sets (up to approximately 30% error rates; Lester & Fairweather [2009a]), so constraints relating to the quality of the biological data are lower than may be the case for other modelling approaches, provided a range of data exists for multiple time points in multiple locations. Therefore, wetlands that have a history of citizen or management-oriented monitoring (e.g. Frogwatch; www.frogs.org.au) may be able to be modelled even though identification of individual biota to species level may not have occurred and error levels may be relatively high. Appropriate transformations (e.g. use of presence/absence data), as well as an understanding of the error rates likely to be present (through quality assessment of the data collection) are recommended but the data are still likely to be of use if they have been collected in a consistent manner through space and time.

Thus, it is highly likely that our ecosystem states approach will be able to be used in regions where other modelling techniques may not be appropriate. In regions where high-quality biological and physico-chemical data are not available (but data of reasonable quality do exist), this technique may be appropriate where others are not. In particular, it will be useful in a range of circumstances for which deterministic methods are not appropriate due to a lack of understanding relating to linkages amongst ecosystem components or no data on parameters of interest. The approach could also be of use in other disciplines where data are patchy or sparse and mechanisms are poorly understood. It is likely to be most relevant where there is a clear distinction between dependent and independent variables (i.e. here environmental data usually drives biotic response) so that the data can be appropriately divided. Parts of the technique (e.g. the prevention of over-fitting through the use of a combination of methods) may be even more broadly applicable elsewhere.

The technique is unlikely to be of much use, however, in areas where there is a poor or only recent history of replicated dependent data (here, biological monitoring data). The degree to which the modelling approach can be used in situations where very little replicated dependent data exists has yet to be determined but, clearly, the more complete the original data sets are (e.g. of good quality, covering a wide variety of taxonomic groups, replicated in both space and time), the more confident the modeller can be in the generality of the resultant model, and thus the more confident the manager can be in the predictions derived from that model.

5. CONCLUSIONS AND RECOMMENDATIONS

Taking an ecosystem states approach to ecological response modelling addresses many of the common limitations associated with available biological (and physico-chemical) data sets. It allows models to be built in the absence of complete understanding of linkages within the ecosystem and where available data are a mixture of quantitative, semi-quantitative and qualitative, collected using disparate collection techniques and where they are patchy in space and time. The approach is also able to deal with a large number of possible predictive variables (~ 230 used for the Coorong model, reduced to six key influential drivers of transitions between states) and so eliminates the need to select somehow a subset of potential predictive variables that are most likely to be driving the system. Such an approach is likely to have broad applications both within ecological modelling and in other disciplines where similar data constraints are found. However, the technique cannot overcome severe limitations in the replication of biological data in space and time, and is thus not likely to be appropriate where there is not a history of consistent ecological monitoring through time. We therefore recommend that the ecosystem states approach be considered as one of a suite of modelling approaches that be assessed for use in a given application based on the available data and the objectives of the modelling exercise.

ACKNOWLEDGEMENTS

REFERENCES

Brookes JD, Lamontagne S, and 32 others, An Ecosystem Assessment Framework to Guide Management of the Coorong: Final report of the CLLAMMecology Research Cluster. CSIRO: Water for a Healthy Country National Research Flagship, ISBN: 978 0 643 09781 0, 47pp. URL: www.clw.csiro.au/publications/waterforahealthycountry/cllamm/CLLAMM-Final-Report-Ecosystem-Assessment.pdf, 2009.

Clarke KR and Gorley RN, PRIMER v6: User manual/tutorial, PRIMER-E Ltd, Plymouth

Lester RE and Fairweather PG, Review of modelling alternatives for CLLAMM Futures, CSIRO: Water for a Healthy Country National Research Flagship, Adelaide. URL: www.clw.csiro.au/publications/waterforahealthycountry/cllamm/cllamm-modellingalternativescllamm-futures.pdf, 2008.

Lester RE and Fairweather PG, Ecosystem States of the Coorong: An Ecosystem Response Model. Method development and sensitivity analyses, CSIRO: Water for a Healthy Country National Research Flagship, ISBN 978 0 643 09779 7, 126pp. URL: www.clw.csiro.au/publications/waterforahealthycountry/cllamm/CLLAMM-ecosystem-states-of-the-coorong.pdf, 2009a.

Lester RE and Fairweather PG, Modelling future conditions in the degraded semi-arid estuary of Australia's largest river using ecosystem states, *Estuarine, Coastal & Shelf Science*, 85, 1-11, 2009b.

Lester RE and Fairweather PG, Ecosystem response modelling of the Coorong, In: Saintilan N and Overton I (Eds.), *Ecosystem Response Modelling in the Murray Darling Basin*. CSIRO Publishing, pp.343-359, 2010.

Lester RE, Webster IT, Fairweather PG and Langley RA, Predicting the Future Ecological Condition of the Coorong. Effects of management and climate change scenarios. CSIRO Water for a Healthy Country National Research Flagship, Canberra, 104pp. URL: www.clw.csiro.au/publications/waterforahealthycountry/cllamm/ CLLAMM-predicting-Coorong-future-ecological-condition.pdf, 2009.

Otto ST and Day T, *A Biologist's Guide to Mathematical Modeling in Ecology & Evolution*. Princeton University Press, Princeton, 2007.

Webster IT, Hydrodynamic modelling of the Coorong, Water for a Healthy Country National Research Flagship, CSIRO, Canberra. URL: www.csiro.au/files/files/pd9z.pdf, 2007.

Wikle CK, Hierarchical Bayesian models for predicting the spread of ecological processes, *Ecology,* 84, 1382-1394, 2003.
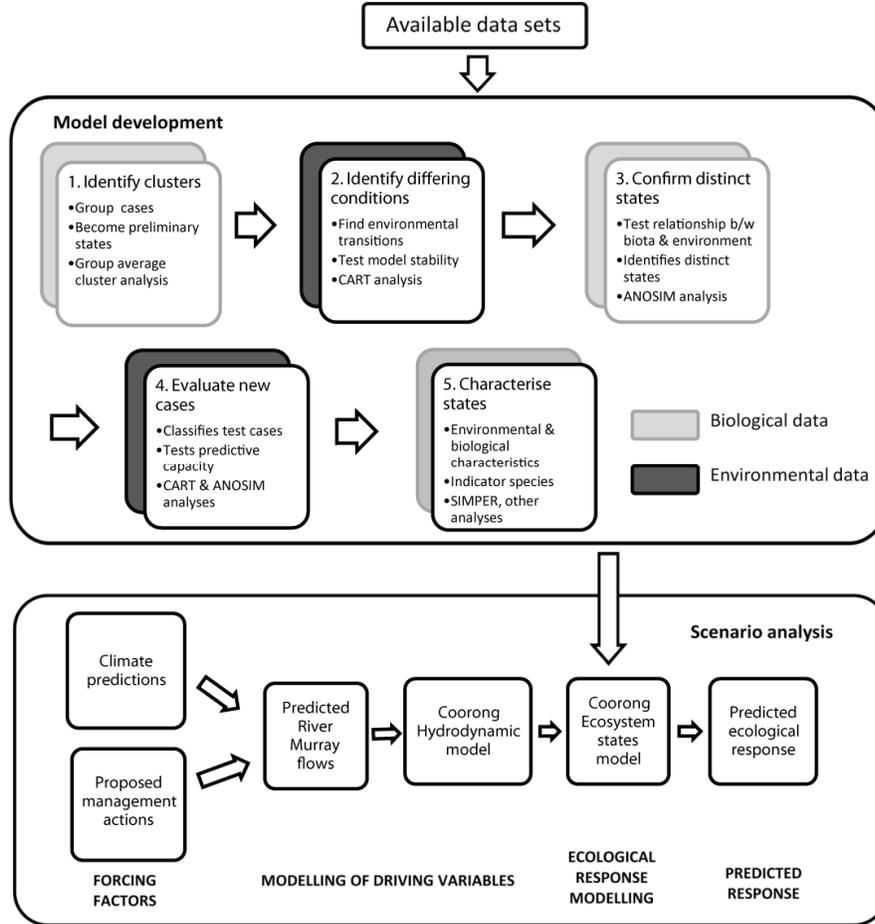
FIGURES

Figure 1. Flow chart describing the model development process and how the ecosystem states model can be used in combination with other existing models for the region to run various scenario analyses, including the generic steps (shown in upper case) that could be followed in applying this method to other case studies or disciplines.