Integrative environmental prediction using Bayesian networks: A synthesis of models describing estuarine eutrophication

Mark E. Borsuk, Craig A. Stow, and Kenneth H. Reckhow
Nicholas School of the Environment and Earth Sciences
Duke University, Durham, North Carolina USA 27708-0328
(mark.borsuk@eawag.ch)

Abstract: The symptoms of coastal and estuarine eutrophication are the result of a number of interacting processes operating at multiple spatial and temporal scales. Thus, models developed to appropriately represent each of these processes are not easily combined into a single predictive model. We suggest that Bayesian networks provide a possible solution to this problem. The graphical structure explicitly represents cause-and-effect assumptions between system variables that may be obscured under other approaches. These assumptions allow the complex causal chain linking management actions to ecological consequences to be factored into an articulated sequence of conditional relationships. Each of these relationships can then be quantified independently using an approach suitable for the type and scale of information available. Probabilistic functions describing the relationships allow key known or expected mechanisms to be represented without the full complexity, or information needs, of highly reductionist models. To demonstrate the application of the approach, we develop a Bayesian network representing eutrophication in the Neuse River estuary, North Carolina from a collection of previously published analyses. Relationships among variables were quantified using a variety of methods, including: process-based models statistically fit to long-term monitoring data, Bayesian hierarchical modeling of cross-system data, multivariate regression modeling of mesocosm experiments, and probability judgments elicited from scientific experts. We use the fully quantified model to generate predictions of ecosystem response to alternative nutrient management strategies.

Keywords: cross-scale modeling, quantitative synthesis, meta-model, risk assessment

1. INTRODUCTION

Eutrophication, in the form of excessive algal growth stimulated by anthropogenic nutrient inputs, is a serious problem in many estuaries and coastal zones [Pelley 1998]. In addition to aesthetic concerns, the enhanced production of algal biomass can have severe ecosystem impacts, including the promotion of bottom-water hypoxia [Cloern 2001]. Documented impacts of hypoxia on aquatic organisms include reduced habitat availability, increased susceptibility to disease and predation, and direct kill events [Dauer et al. 1992]. Thus, to control eutrophication and its consequences, many coastal states are considering watershed management actions intended to reduce riverine nutrient loading [Pelley 1998].

For guidance in this process, decision-makers often ask scientists to provide a predictive link between proposed management actions and ecosystem response [NRC 2001]. However, this link represents a complex causal chain, the entirety of which rarely falls under the domain of a single, coordinated research project. This makes prediction of ecological effects difficult.

Environmental models represent attempts to combine the scientific understanding gained from multiple projects into a single predictive framework. Most models do this by endeavoring to simulate all of the physical, chemical, and biological processes occurring in the system at some pre-determined model scale. However, there is no single scale at which all ecosystem processes operate nor all relevant knowledge applies. Thus, the key to successful prediction lies in choosing scales at which predictable patterns emerge. These patterns may be the collective result of finer scale units, or the consequence of larger scale constraints [Levin 1992], and are likely to differ depending on the specific variables or processes being studied. This complexity creates a serious modeling challenge, and methodologies are required that can link process representations developed at multiple scales and in a variety of forms [Pace 2001]. There is also a need to assess how

uncertainties in each component translate to uncertainty in the final predictions [Reckhow 1994]. Finally, such integrative models must be able to be easily updated to reflect evolving scientific knowledge and policy needs [Walters 1986].

We have found causal Bayesian networks [Pearl 1988] to be one of the most promising methods for such predictive environmental synthesis. The use of a graphical depiction to identify conditional independencies among important system variables facilitates the coherent integration of disparate submodels. The basics of Bayesian networks and their application to environmental prediction are well described in the literature [Reckhow 1999, Borsuk et al. 2002a]. We will not provide extensive background here. Rather, our purpose is to demonstrate the use of this method for integrating lower-level models, developed at multiple scales, into a continuous causal chain. We do this by synthesizing a Bayesian network representing eutrophication in the Neuse River estuary, North Carolina from a collection of previously published analyses. The full model is then used to generate predictions of ecosystem response to alternative nutrient management strategies.

2. BAYESIAN NETWORKS

Briefly, a Bayesian network is a graphical depiction of the relationship among the most important variables in the system of interest. In this depiction, the variables are represented by round nodes, and dependencies between one variable and another are represented by an arrow. The conditional independence, implied by the *absence* of any connecting arrows, greatly simplifies the modeling process by allowing separate submodels to be developed for each conditional relationship indicated by the *presence* of an arrow. These submodels may be derived from any combination of process knowledge, statistical correlations, or expert judgment, depending on the extent of information available about that particular relationship.

In a Bayesian network, each dependency indicated by an arrow is characterized by a conditional probability distribution that describes the relative likelihood of each value of the down-arrow node, conditional on every possible combination of values of its parents. A node that has no incoming arrows is said to have no parents, and such a variable can be described probabilistically by a marginal (or unconditional) probability distribution.

A useful practice for developing the conditional distributions in the network is to view them as functional relationships among variables [Pearl 1999]. In this reading, the links represent autonomous physical mechanisms, as described by mathematical functions, and probabilities are introduced through the assumption that certain variables, or other causes of random disturbances, are unobserved. In other words, each set of arrows pointing to a given node, X_i, represents a function of the form:

$$X_i = f_i(\mathbf{pa}_i, \varepsilon_i) \qquad i = 1,\ldots,n \qquad (1)$$

where $\mathbf{pa}_i$ are the parents of the node X_i, and the ε_i are independent random disturbances with arbitrary distributions. The disturbance terms represent exogenous factors that, for reasons of either choice or ignorance, have not been included in the analysis. Nodes without parents are simply functions of a disturbance term.

This type of functional characterization of the causal relationships in a Bayesian network leads to the same advantages of recursive decomposition as the strictly distributional forms. However, specifying functional equations among variables, rather than conditional distributions, is a task more naturally consistent with both the theory and routine practice of process-oriented environmental science.

3 NEUSE RIVER NETWORK

3.1 Graphical Causal Model

The most publicly visible and ecologically harmful impacts of eutrophication in the Neuse estuary are the result of a number of interacting processes operating at multiple spatial and temporal scales [Cloern 2001]. These include (in approximate order of causal dependence):

1. Abundant algal growth stimulated by warm temperatures and excessive watershed nutrient (specifically, nitrogen) inputs;
2. Increased presence of toxic microorganisms that use algae as a food source;
3. Excessive accumulation of organic carbon produced by algae;
4. Benthic bacterial consumption of accumulated organic matter and associated respiration of dissolved oxygen;
5. Summer salinity stratification leading to reduced vertical mixing and depletion of bottom water dissolved oxygen;
6. Low oxygen-induced shellfish mortality;
7. Reduced fish health and growth rate resulting from the effect of low oxygen on susceptibility to disease, abundance of prey

resources, and the availability of suitable habitat;

8. Occurrence of massive fish kills when cross-channel winds trap fish in low oxygen zones without an escape route, especially when fish health is already compromised.

A Bayesian network can be used to explicitly represent the variables and causal relationships involved in these processes (Figure 1). Each has been quantitatively characterized in previous studies in a manner consistent with available data and scientific knowledge. These lower-level sub-models are described in detail in published papers and will be outlined only briefly below.

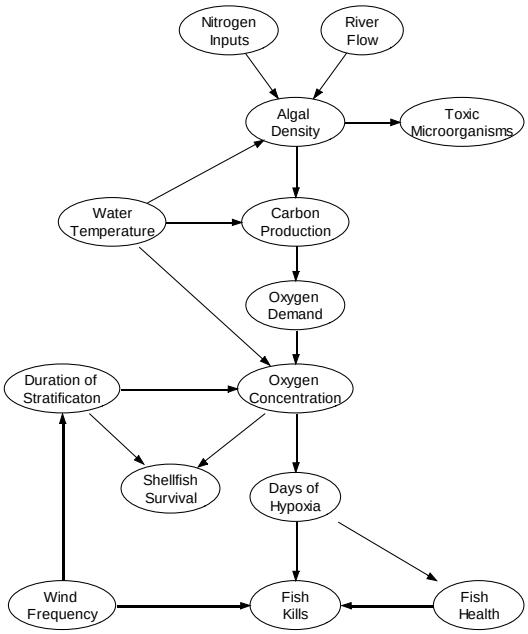

Figure 1. Causal network of the important factors involved in eutrophication in the Neuse estuary.

3.2 Conditional Relationships

1. Nitrogen Input / Algal Density Relationship

A relationship between algal density, as measured by chlorophyll a concentration, estuarine location, water temperature, and incoming Neuse River flow and total nitrogen concentration was developed using a regression model fit to approximately five years of biweekly monitoring data [Borsuk et al. 2002d]. Model results indicated a positive relationship between chlorophyll and nitrogen input concentration for all locations in the estuary. River flow was found to display a piecewise linear, ^-shaped relationship at mid- and lower estuary locations. A positive relationship between chlorophyll concentration and water temperature was found for all sections.

The regression setting used to develop this model provides a natural consistency with the form of the probabilistic functional equation given in eq. 1. Algal density is the response variable, X, in this equation, and water temperature, river flow, and nitrogen concentration are the parents, $\boldsymbol{pa}$. The distributional error term, ε, corresponds to the distribution of regression residuals, assumed to be Normal with a standard deviation estimated by the root mean squared error (RMSE) of the regression. Model parameters can also be respresented as parents in the Bayesian network, with a marginal distribution described by the mean vector and covariance matrix estimated by the regression procedure. In this case, since the model is linear and Normal, the parameter distribution derived in this way is equivalent to the posterior parameter distribution that would result from a Bayesian analysis with non-informative priors [Lee 1997].

2. Algal Density / Toxic Microorganism Relationship

The presence of the toxic dinoflagellate, *Pfiesteria piscicida*, is of particular concern to the public because of the large amount of media attention it has received in recent years. [Burkholder and Glasgow 2001]. When in its non-toxic zoospore stage, *Pfiesteria*'s primary food source is algae [Burkholder et al. 1995], suggesting a linkage between the zoospore and its prey in natural settings. This relationship was recently quantified [Borsuk et al. 2002f] using data obtained from a set of mesocosm experiments by Pinckney et al. [2000]. The relationship between algal density and *Pfiesteria* cell counts was found to be approximately linear after a log-transformation of both variables during the summer season. The model can be interpreted probabilistically in a manner similar to the algal density model above.

3. Algal Density / Carbon Production Relationship

To predict primary productivity from algal density, Borsuk et al. [2002d] used a generalized version of the model originally proposed by Cole and Cloern [1987]. The model, which expresses daily algal carbon productivity as a function of biomass, photic depth, surface irradiance, and water temperature was fit to approximately five years of biweekly monitoring data at 11 mid-channel sampling locations. Photic depth and surface irradiance were found to not be significant terms in the model, perhaps because of the variable irradiance method employed in the determination of productivity.

4. Algal Carbon Production / Sediment Oxygen Demand Relationship

The historical values of algal carbon production do not span the range that may be expected under a significant change in nutrient inputs. Therefore, Borsuk et al. [2001a] relied on cross-system data from 34 estuaries and coastal zones to parameterize a simple, mechanistic model relating carbon production and sediment oxygen demand, including the effects of water column decay and sediment burial. To do this, a hierarchical approach was employed. Both global and system-specific parameters were estimated using Bayes Theorem.

5. Sediment Oxygen Demand / Bottom Water Oxygen Concentration Relationship

A process-based model of oxygen depletion was specified that is consistent with established theory yet is simple enough to be empirically parameterized from available monitoring data [Borsuk et al. 2001b]. The model represents the processes of microbial oxygen consumption and physical reoxygenation, including the effects of temperature and vertical stratification. Nonlinear regression allowed for the direct estimation of rate constants from field data. The resulting model can be used to probabilistically predict the frequency distribution of bottom water oxygen concentrations, conditional on the rate of sediment oxygen demand.

6. Bottom Water Oxygen Concentration / Shellfish Survival Relationship

To relate oxygen status to shellfish abundance in the Neuse River estuary, Borsuk et al. [2002c] developed a survival model for the clam *Macoma balthica*, a critical species in the Neuse ecosystem. Experimental studies have not yet been performed to directly address the sensitivity of this species to low oxygen conditions. Therefore, this sub-model relied upon the expert judgment of two marine biologists to provide the data used in model building. Well-developed methods exist for eliciting expert judgments [Morgan and Henrion 1990], and the marine science literature and the experts' own experience form a solid foundation for accurate assessment. Model parameters were then estimated from the assessed data using Bayes' Theorem. The resulting model probabilistically relates survival of *M. balthica* to time of exposure (duration of stratification) and dissolved oxygen concentration.

7. Dissolved Oxygen Concentration / Fish Health Relationship

Borsuk et al. [2002b] relied upon the elicited judgment of two experienced estuarine fisheries researchers to characterize the relationship between fish population health and the annual extent of bottom water hypoxia. According to the scientists' assessments, the health of the fish population declines nonlinearly with increasing temporal extent of hypoxia. However, a number of factors in addition to oxygen were also believed to affect fish health. Because these other factors were not explicitly included in the analysis, they are manifest as disturbance terms, resulting in the attribution of some likelihood to more than one fish health category a given number of days of hypoxia.

8. Fish Health / Dissolved Oxygen Concentration / Fish Kill Relationship

The probabilities of fish kills of varying magnitudes, conditioned on a given state of fish population health, the occurrence of a strong cross channel wind, and varying bottom water oxygen concentrations, were elicited from the same estuarine fisheries scientists that were questioned for the fish health model [Borsuk et al. 2002b]. Assessment results showed that fish kills are expected to be relatively rare, even with all the conditions being right, with probabilities exceeding 50% only for relatively small kills and a population in poor health. The assessed conditional probabilities of fish kills of varying magnitudes can be used directly in the network model.

3.3 Marginal Probabilities

In addition to the conditional probabilities characterized by the functions described above, full specification of the Bayesian network requires marginal distributions for the outermost variables: water temperature, nitrogen inputs, river flow, and duration of stratification. Because nitrogen reductions in the Neuse are to be expressed in terms of a percent reduction relative to a 1991-95 baseline [NCDWQ 2001], daily data collected during those years at the most downstream river monitoring station served as the basis for the marginal variables. These were represented in the network as a multivariate empirical distribution in order to maintain any underlying dependencies. To predict the effect of a substantial reduction in nitrogen inputs to the Neuse estuary, the marginal distribution for nitrogen inputs was replaced by one in which all values were multiplied by one half. All other functions and marginal nodes in the model were left unchanged, and new distributions were computed for the ecological variables of interest.

3.4. Integrated network

Figure 2 summarizes the full set of marginal and conditional probabilities described in the previous subsections.

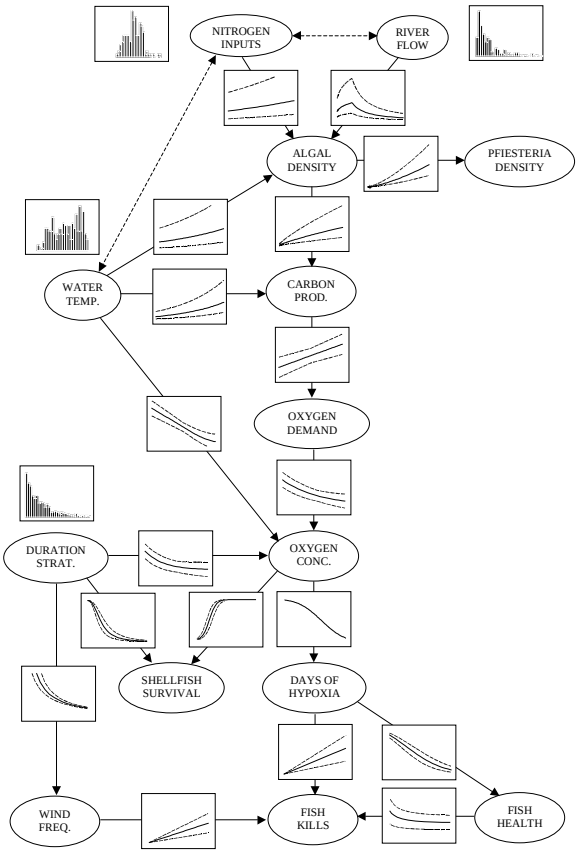

Figure 2. Fully characterized Bayesian network for the Neuse estuary. Relationships between any two variables are shown as bivariate functions representing the partial effect of the parent node. Conditional probabilities are summarized by 90% predictive intervals, as determined by the sub-models

4. RESULTS

The marginal distributions of model endpoints (Figure 3, solid curves) show the relative likelihood of alternative values under the baseline scenario (no nitrogen reduction). These results are in good agreement with historical observation [Borsuk et al. 2002e]. Under the proposed 50% reduction in nitrogen inputs (Figure 3, dashed curves), the values of all ecological endpoints are predicted to improve somewhat, but not necessarily proportional to the nitrogen reductions. For all variables, predictive uncertainty arising from natural variation and knowledge uncertainty is substantial. The magnitude of the combined sources of uncertainty depends on the nature of the variable being predicted. In general, the less observable, less frequent, and further down the causal chain a variable is, the greater the predictive uncertainty. However, this type of variable is precisely the one of most interest to stakeholders (e.g. fish kills, shellfish survival). This observation suggests that a compromise is necessary between achieving policy relevance and predictive precision. Selecting the appropriate degree of compromise and the appropriate target values for the selected variables are tasks that can best be performed by public decision-makers.

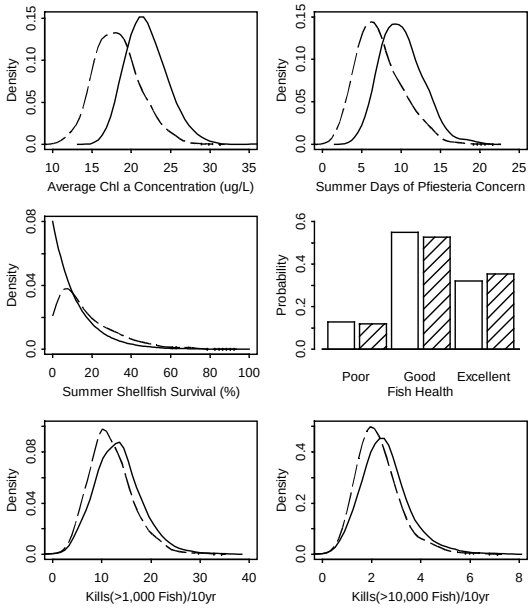

Figure 3. Marginal probability distributions of model endpoints. The baseline scenario is shown as a solid curve, and reduction scenario is shown as a dashed curve or diagonally-striped bar.

5. CONCLUSIONS

Because there is no single scale at which scientists have studied the Neuse Estuarine system, there is no single scale at which all system processes can be represented. Therefore, a characteristic of the Bayesian network that we exploited is its ability to integrate sub-models of disparate scales. Choosing the various scales of representation in a Bayesian network should be a dynamic and iterative process. This is because while the intent is to choose scales that will represent key features of the natural system, it is more often the case that scales are imposed by observational capabilities or technological or organizational constraints [Levin 1992] which may evolve over time. Further, the scale of prediction should correspond to the needs of decision-makers, which may also change with time as they gain understanding of the problem. Such updating of the model is facilitated by the

conditional independencies identified in the causal network representation. These independencies, implied by the lack of a connecting arrow between two nodes, allow for the modularization of the full model into independent causal structures. When the nature of one of these sub-structures is revised, because of either a change in knowledge or a change in environmental conditions, the other structures remain unaltered.

6. REFERENCES

Borsuk, M. E., P. Burkhardt-Holm, and P. Reichert. 2002a. A Bayesian network for investigating the decline in fish catch in Switzerland. IEMSS, Integrated Assessment and Decision Support, Lugano, Switzerland.

Borsuk, M. E., L. A. Eby, and L. B. Crowder. 2002b. Probabilistic prediction of fish health and fish kills in the Neuse River estuary using the elicited judgment of scientific experts. In review.

Borsuk, M. E., D. Higdon, C. A. Stow, and K. H. Reckhow. 2001a. A Bayesian hierarchical model to predict benthic oxygen demand from organic matter loading in estuaries and coastal zones. Ecological Modelling **143**:165-181.

Borsuk, M. E., S. P. Powers, and C. H. Peterson. 2002c. A survival-based model of the effects of bottom-water hypoxia on the density of an estuarine clam population. In review.

Borsuk, M. E., C. A. Stow, R. A. Luettich, H. W. Paerl, and J. L. Pinckney. 2001b. Modelling oxygen dynamics in an intermittently stratified estuary: Estimation of process rates using field data. Estuarine Coastal and Shelf Science **52**:33-49.

Borsuk, M. E., C. A. Stow, and K. H. Reckhow. 2002d. The confounding effects of nitrogen load on eutrophication of the Neuse River estuary, North Carolina. In review.

Borsuk, M. E., C. A. Stow, and K. H. Reckhow. 2002e. Ecological prediction using causal Bayesian networks: A case study of eutrophication management in the Neuse River estuary. In preparation.

Borsuk, M. E., C. A. Stow, and K. H. Reckhow. 2002f. Integrated TMDL development using Bayesian networks. In review.

Burkholder, J. M., and H. B. Glasgow. 2001. History of toxic Pfiesteria in North Carolina estuaries from 1991 to the present. Bioscience **51**:827-841.

Burkholder, J. M., H. B. Glasgow, and C. W. Hobbs. 1995. Fish kills linked to a toxic ambush-predator dinoflagellate: distribution and environmental conditions. Marine Ecology-Progress Series **124**:43-61.

Cloern, J. 2001. Our evolving conceptual model of the coastal eutrophication problem. Marine Ecology-Progress Series **210**:223-253.

Cole, B. E., and J. E. Cloern. 1987. An empirical model for estimating phytoplankton productivity in estuaries. Marine Ecology-Progress Series **36**:299-305.

Dauer, D. M., A. J. Rodi, and J. A. Ranasinghe. 1992. Effects of low dissolved-oxygen events on the macrobenthos of the lower Chesapeake Bay. Estuaries **15**:384-391.

Lee, P. M. 1997. Bayesian Statistics: An Introduction. Wiley and Sons, New York.

Levin, S. A. 1992. The problem of pattern and scale in ecology. Ecology **73**:1943-1967.

Morgan, M. G., and M. Henrion. 1990. Uncertainty. Cambridge University Press, Cambridge.

NRC. 2001. Assessing the TMDL Approach to Water Quality Management. National Academy Press, Washington, D.C.

NCDWQ. 2001. Phase II of the Total Maximum Daily Load for Total Nitrogen to the Neuse River Estuary, North Carolina. North Carolina Department of Environment and Natural Resources, Raleigh, NC.

Pace, M. L. 2001. Prediction and the aquatic sciences. Canadian Journal of Fisheries and Aquatic Sciences **58**:1-10.

Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA.

Pearl, J. 1999. Graphs, Structural Models, and Causality. Pages 95-138 *in* C. Glymour and G. F. Cooper, editors. Computation, Causation & Discovery. AAAI Press, Menlo Park, CA.

Pearl, J. 2000. Causality. Cambridge University Press, Cambridge, UK.

Pelley, J. 1998. Is coastal eutrophication out of control? Environmental Science & Technology **32**:462A.

Pinckney, J. L., H. W. Paerl, E. Haugen, and P. A. Tester. 2000. Responses of phytoplankton and Pfiesteria-like dinoflagellate zoospores to nutrient enrichment in the Neuse River Estuary, North Carolina, USA. Marine Ecology-Progress Series **192**:65-78.

Reckhow, K. H. 1994. Importance of Scientific Uncertainty in Decision-Making. Environmental Management **18**:161-166.

Reckhow, K. H. 1999. Water quality prediction and probability network models. Canadian Journal of Fisheries and Aquatic Sciences **56**:1150-1158.

Walters, C. J. 1986. Adaptive Management of Renewable Resources. Macmillan, New York.