DATA ANALYSIS AND MODEL DEVELOPMENT FOR ECOSYSTEM MANAGEMENT AT LAKE GEORGE, N.Y.

By

Glenn Roberts and Frank DiCesare

(Reprinted from: 1977 Proceedings of the International Conference on Cybernetics and Society, I.E.E.E., NY, NY)

FWI Report 77-11

DATA ANALYSIS AND MODEL DEVELOPMENT FOR ECOSYSTEM MANAGEMENT AT LAKE GEORGE, N. Y.

Glenn Roberts
Electrical & Systems Engineering Department
Rensselaer Polytechnic Institute
Troy, New York 12181

Frank DiCesare
Electrical & Systems Engineering Department
Rensselaer Polytechnic Institute
Troy, New York 12181

Introduction

Lake George is a large oligotrophic fresh-water natural resource located in the Adirondack Mountains of New York State (see Fig. 1). Present activities within the Lake George watershed have the potential to significantly alter the quality of the lake, and its value as a resource. Anticipated problems are large-scale residential and/or commercial (tourist) land development with its associated waste products, including sewage, solid waste, siltation and thermal pollution. This paper describes a methodology for developing a lake ecosystem management model which takes advantage of the large amounts of research data previously collected at Lake George. This methodology will be exemplified in terms of modeling research presently under way.

Background

Because Lake George is just beginning to be seriously impacted by man's activities and without proper environmental management could be modified adversely, it was chosen as a research site in the National Science Foundation's International Biological Program (IBP).

An overriding goal of the previous research at the Lake George site has been the development of a more thorough understanding of the ecosystem as a whole, and of processes within the system. The basis of the program has been the identification and measurement of the significant state variables and transfer rates between compartments of the ecosystem. These data have been used to develop detailed process oriented models which have a high degree of biological realism and reflect the state-of-the-art in process understanding. This process modeling research, which seeks the fundamental biological laws governing each component or sub-component of the ecosystem, has helped provide the foundation for ecosystem modeling research, which seeks to identify the mathematical relationships governing the dynamics of individual components as they relate to other components of the ecosystem.

Because of this close association between process models and system models, it was only natural that the first modeling efforts at Lake George consisted of coupling the complex process models to form a whole ecosystem model (CLEAN).[10] This approach to modeling has been used in past efforts at Lake Texoma[11], Lake Wingra[7], and the San Joaquin River[3] to name a few. An alternative approach is to simulate the important inter-compartmenal mass flows within the ecosystem, without attempting to describe the complex actions in each compartment. Efforts at Lake George have shown that a simple 6^{th} order set of differential equations can reproduce the data very well[2], and Williams' model of Cedar Bog Lake has met with similar success.[13]

There are advantages and disadvantages to each of these two approaches. The first approach, that of coupling complex process models, usually produces a model having a very large number of parameters and

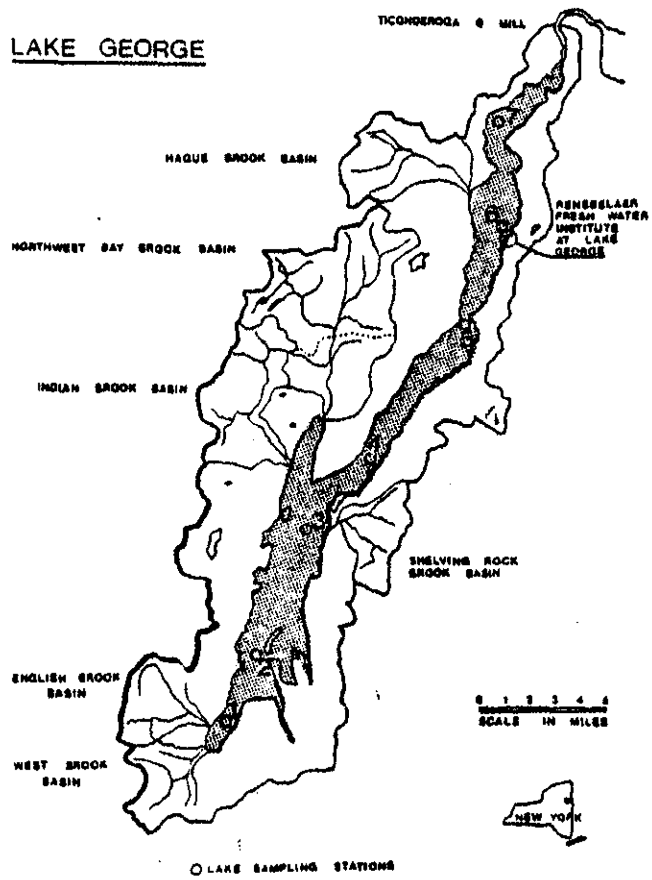


Figure 1

requiring a considerable amount of data for accurate calibration. If these calibration data are available, the model may be used to study certain internal dynamics of the ecosystem, which are difficult or impossible to monitor. Although results obtained in this manner often cannot be validated, they can provide important insight into the complex mechanics of the system, and can raise questions which stimulate further research. The second approach, that of simulating only mass flows among compartments, produces a less complicated model having few parameters and requiring relatively little data for calibration. Because of their inherent simplicity, these models are more easily understood and applied.

In the past, both of these types of models have been developed with the primary goals being the stimulation of thought processes among ecologists and the explanation of observed phenomena. Such models may be thought of as research oriented models. The temptation exists to use these models for predictive purposes; however, if one's goal is environmental management, both the methodology for modeling and the resultant models will be different.

This work supported by the National Science Foundation under Grant DEB-76-12294 and administered by the Rensselaer Fresh Water Institute at Lake George.

In this paper, we will discuss our methodology for construction of data-based policy analysis models for lake ecosystems using as a case study the modeling research being conducted at Lake George.

The Modeling Process

Consider the modeling process as depicted below

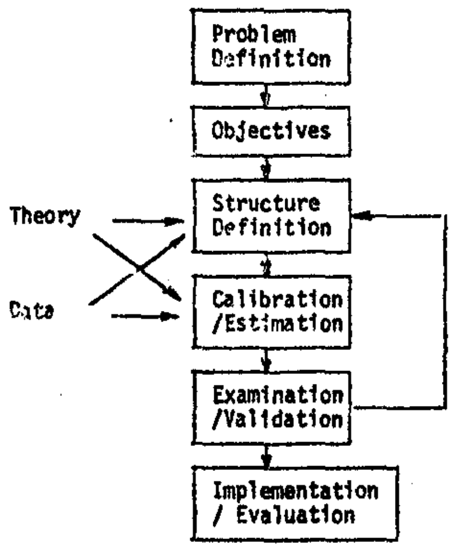


Figure 2

We envision a six-step process defined as follows:

1) Problem Definition - The examination of the system to determine what aspects may be problematic. The examination of models of system behavior to determine areas which may deserve further attention.

2) Specification of Objectives - The specification in light of the above problems, of the particular goals of the research, and the desired capabilities and restrictions of the model.

3) Specification of Model Structure - Specifications are based on inputs from objectives, understanding of theories, and analysis of the data. Since model structure must always be biased by our own goals and theories we should note these assumptions and biases carefully.

4) Calibration/Parameter Estimation - The determination, given the model structure and possibly some knowledge of the physical significance of the model parameters, of the parameter values which cause the model to "best" reproduce observed system behavior.

5) Examination/Validation - The determination of whether or not the particular model structure being considered has led to a model which meets our goals. If not, we revise the model structure.

6) Implementation - The use and evaluation of the model as a tool to address the problems stated in step number 1.

The topics we have touched upon thus far have dealt directly only with model development, that is structure and calibration. We have, however, touched indirectly upon the preliminary steps, problem definition, and statement of objectives. It is here that the modeler must decide exactly what he wants his model to be able to do. The goals he sets will influence all future decisions in the modeling process. In the authors' opinion, often little or no thought is given to these areas in hopes that the final model will be very general and useful for many things.

Problem Statement

This research is simultaneously addressing problems at two levels. On the policy level, there exists no easily used model to test land-use and water pollution control strategies. At the modeling level there exists no unified methodology for taking advantage of the existing data and our understanding of the ecosystem to develop predictive tools.

Regarding the policy level, the present Lake George water is of drinking quality. However, in recent years man's impact on the lake has become more apparent as indicated by occasional algae blooms at the more highly developed south end of the lake. At this level, the problem is to develop a model which can estimate the impact of further development and of water pollution control strategies. Attempts to use existing research models have met with only limited success.

At the modeling level, the problem is to develop a methodology for policy analysis model development.

Objectives

Based on these problems we have produced a set of goals and objectives for our work.

Our principal objective is to have a model which can be used to assess man's impact on the lake ecosystem. In particular, development may cause serious changes in the inputs of nutrients (e.g. phosphorus and nitrogen) and organic matter and may raise the temperature of the water. Increased human activity would directly lead to more human related wastes and by-products and any change in the lake basin such as cutting the forest or paving streets would change the amount and character of the precipitation and stream inputs. Man can, of course, limit these inputs by land use control and by waste water treatment.

The impact on the lake is perceived by changes in the levels of biotic components of the ecosystem. (One of the first signs of further eutrophication is higher level of algae).

This assessment model should have the following properties: it should take full advantage of the dat available as reflected by its structure and calibration; it should reflect, insofar as possible, the macro-ecological structure of the system; it should b easy for the user to understand in terms of mathemati cal construct and behavior; and it should be calibrated in a manner which provides us with some quanti tative measures of performance, in order that we may be more objective when comparing various model forms.

Concomitant objectives of this research are the examination and comparison of various methods of para meter estimation and of several different structural forms and simply to develop a better understanding of the system and its dynamics from a policy viewpoint.

A Review of Model Development

Let us consider the two-part process which we refer to in Figure 2 as "model development". The process consists of defining the structure and calibration. Each of these procedures can be thought of as being somewhere between highly inductive and highly deductive as shown in the figure below.

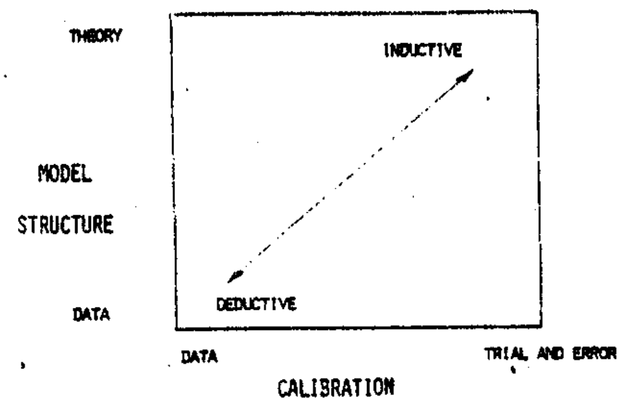


Figure 3

The problem is really one of storing in the model information about the real system. This information is stored in the structure and in the parameters. It is obtained from both the data and the theory and is incorporated through the structuring and calibration processes. Our goal is to maximize the amount of information contained in the model subject to our own constraints of time and complexity. The model structure may range from being determined entirely by the theory to being based merely on observed patterns in the data. The calibration procedure may be an iterative method wherein selected parameters are adjusted until the model produces a "good fit" or it may be a more systematic method with all parameters being identified simultaneously such that certain error quantities are minimized, and statistical measures of performance are obtained.

Most ecological modeling projects to date have relied heavily on inductive type reasoning and trial and error calibration which places them somewhere in the upper right corner of Fig. 3. In this sense these models are unbalanced and should incorporate more of the information contained in the data, hence we would like to be nearer the upper left corner. Extensive efforts have been devoted toward the development of more quantitative calibration techniques for ecosystem models, but these efforts have often met with only limited success since the highly theoretical models also tend to be highly complex.

Model Structure

As we stated earlier, a great deal of effort in the past has been centered around coupled process models[11,10,3] which tend to be tightly coupled, high order sets of non-linear differential equations involving perhaps thousands of parameters. One justification for this type of work has been a striving for biological reality on a micro level, but a second justification has often been a lack of data. When some data are lacking, we lose information on the system and are forced to devise theories to make up for this loss, thus the representation becomes more theory-based (as opposed to data-based).

Theory-based models can tend to be unbalanced in terms of accuracies of the process models. Those lake processes which are well studied and understood are represented by tested process equations, but other processes which are not as well studied are modeled based on a series of assumptions and guesses, the justification being that this is the best that can be done. The resulting model should be considered to be only as reliable as its least reliable sub-model much as a chain is only as strong as its weakest link or a series of numerical calculations is only as accurate as the least accurate numerical entry. This problem may, however, be somewhat alleviated by the feedback coupling in the relationships.

These considerations have led us to believe less complex structures such as inter-compartmental mass-flow and species population models to be the best types of structures given that we are interested in lake management. We are still striving for biological reality, for without it we are getting the right answers for the wrong reasons, however, this reality is on a macro-ecological level, that is it pertains only to those compartmental actions and interactions necessary for realistic model outputs for the desired range of inputs.

Calibration

Referring again to Fig. 3, many modelers find that their work lies somewhere in the upper right corner of the picture. Due to a lack of data and/or a high degree of model complexity they are forced to calibrate by eye, using their judgement to know which parameters to adjust and by how much to adjust them. There may in fact be an infinite number of parameter values for which the model will "fit" the data, this due to the fact that there are actually more parameters than data points. From a statistical point of view the problem is one of insufficient degrees of freedom in the data to accurately estimate the model parameters.

We at Lake George have found ourselves in a position of having an extensive data base and desiring a model of simple structure, low complexity, and a few parameters. Working from such a position, we are able to experiment with various parameter estimation methods designed to give us the best choice of parameters without any guess-work.

Lake George Model Development

As mentioned earlier, data have been collected at Lake George under various research projects over the past nine years. These projects have been sponsored by several agencies in addition to the National Science Foundation and have been continuously coordinated by the R.P.I. Freshwater Institute. This program of research has resulted in the accumulation of an extensive set of data. Realizing that these data possess a wealth of information, one of the first priorities of our current research project was to organize and verify the data such that they could be used in both hypothesis testing and model building.

Based on an examination of these data and knowledge of current and forecasted land-use patterns, the decision was made to concentrate our efforts on the study of the south basin of Lake George. This is represented by station 1 (see Figure 1) of six stations where lake data have been collected. Table 1 gives a list of these data.

The data sets were carefully reviewed in preparation for structuring our model. The following are some of the observations noted, along with comments on how they may affect the structure. The reader is cautioned that these comments do not represent tested hypotheses, but merely the thoughts and ideas which help to guide us in the modeling process.

Measured Variable	Princip. Investig.	No. of Data Points	Time Span	Comments
Phytoplankton(biomass)	Howard	7	9-13-72 to 8-29-73	Includes species breakdown.
-nytoplankton(numbers)	Williams	30	11-28-67 to 7-22-70	Only total Plankton diatoms.
$_{max}$(max. phyto.prod.)	Stross	29	7-10-69 to 11-7-72	Values can be used to calculate mean daily production.
Zooplankton(numbers)	McNaught	40	6-19-70 to 10-11-73	Includes species breakdown.
Decomposer Bacteria(numbers)	Clesceri, L.	10	6-21-72 to 11-7-72	Data generally include nitrate, ammonia, filtrable reactive phosphorous, and dissolved & particulate organic carbon.
Lake Chemistry	FWI	30	1973, 1975, & 1976	
Precipitation Loadings	FWI	28	1972, 1973	Includes total nitrogen & phosphorous
Stream Loadings	Fuhs	25	70.210 - 71.194	Includes dissolved inorganic nitrogen,orthophosphate, and dissolved & particulate organic carbon.
Incident Solar Radiation	Colon	66	68.336 - 71.154	Reported as weekly averages.
Water Temperature	FWI	46	1972, 1973, 1976	

Table 1

1) There appear to be predictable annual fluctuations in key variables (see Fig. 4). Because of this,

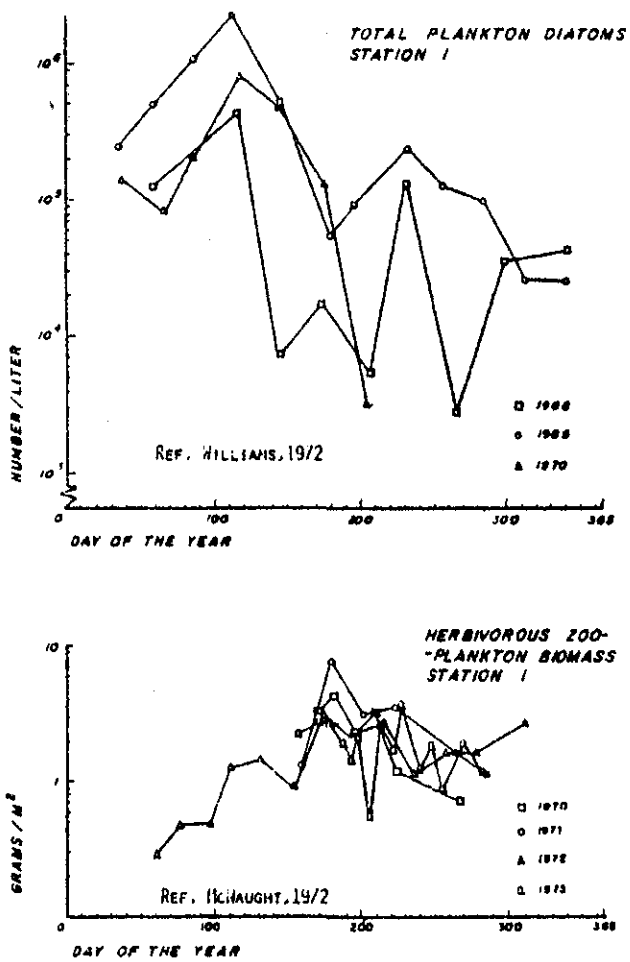


Figure 4a

we feel that we can explain most of this variation through recurring phenomena such as the dynamics of solar radiation, water temperature, stream loadings,etc. and through feedback within the system. These observations and assumptions have been made by modelers in the past.

To relate this observation to the structure, we have been considering using some form of difference equation:

$$\underline{x}(k+1) = \underline{f}[\underline{x}(k),\underline{u}(k)] \quad (1)$$
$$\underline{y}(k) = \underline{g}[\underline{x}(k)]$$

where the vector $\underline{x}$ represents the vector of state variables, $\underline{u}$ represents the driving variables and $\underline{y}$ represents the output variables (usually the species' biomass).

We are aware of both the pros and cons of using linear models [1, 11] and we feel that we must examine this family of models first and then, if we find it necessary, go on to more complex formulations. From the theory we know that the primary productivity rate (in this case the rate of growth of phytoplankton) is a function of the level of solar radiation and the biomass of the phytoplankton. Also generally accepted in theory is the fact that the growth rates of all biotic components are temperature dependent. Since temperature and light exhibit annual fluctuations, a form where temperature and light form time varying coefficients of a set of difference equations would be implied.

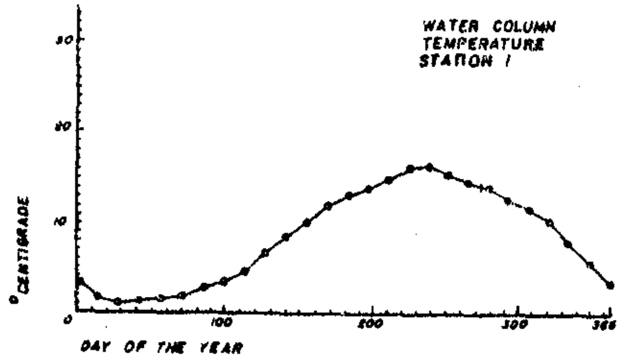
$$\underline{x}(k+1) = A(k)\underline{x}(k) + B(k)\underline{u}(k) \quad (2)$$

On the other hand, in observing the data as shown in figure 4 one can hypothesize that the lake responds basically as an underdamped constant coefficient system to a spring pulse in nutrients (e.g. phosphorus) and light. This could lead to a linear time invariant donor controlled ecosystem model [9].

$$\underline{x}(k+1) = A\,\underline{x}(k) + B\,\underline{u}(k) \quad (3)$$

2) There exists a large variance in individual species responses and a lower variance when these species are aggregated; thus, as expected, the assemblage of species exhibits a greater predictability in behavior than

any single member species.



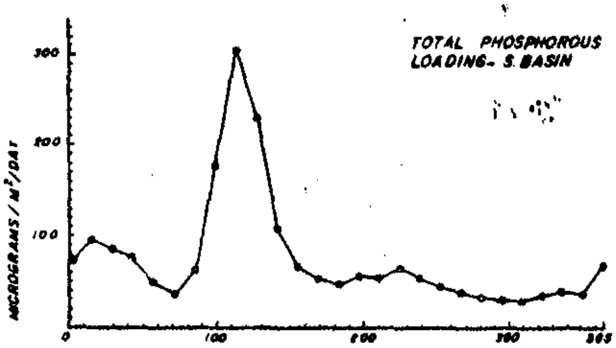


Figure 4b

The implication for modeling is that if appropriate assemblages are chosen as variables, then a model of relatively low order will result in explanation of the aggregate variations. Of course the main concern must be that this macro-ecological representation is adequate for the policy analysis functions of the model. If this is the case, then there is no reason to work at any greater level of detail.

3) As can be seen in Figure 4 there appears to be a consistent lag between the fluctuations of phytoplankton and zooplankton. This lag has been attributed to the predation of phytoplankton by zooplankton. However, the lag observed in the data is longer than represented by previous models of Lake George and raises questions as to the cause effect relationships previously assumed.

The data show the phytoplankton peaking in late April and the zooplankton reaching a maximum in late June (just as the phytoplankton reach a minimum) some 60 days later. We shall have to investigate, therefore, whether this consistently observed zooplankton peak is indicative of the prey-predator relationship between phytoplankton and zooplankton or is related more to other recurring lake phenomena such as variations in temperature and organic matter.

4) As a result of regression analysis on some of the data, we have noticed a linear correlation between phosphorus loadings (external inputs) and the biomass of phytoplankton. While previous studies have shown that phosphorus can be an important control variable[5] we know that these observed variations are at least partially explained by other factors such as light, temperature, and zooplankton predation. The question therfore is: how much of the phytoplankton dynamics is explained by these observed variations in phosphorous loadings, and how much by other factors? By using statistical calibration techniques we hope to be able to analyze the significance of the many such sources of variation.

5) As we expected, the solar radiation data exhibit large day-to-day variations due to cloud cover. These fluctuations are at a higher frequency than the sampling rate of the biotic data. The assumption which we will make is that the lake system acts as a low-pass filter which averages these effects especially in terms of the output of interest. The morphometry (depth, volume, etc.) of Lake George suggests that this assumption is reasonable, i.e., the lake is deep and has a large volume of water. Most observers note that it does not respond instantaneously to single events, as opposed to more shallow, eutrophic lakes which can exhibit strong sensitivity to rapid changes in outside factors such as light and external loadings.

6) Figure 4 shows that some of the data sets vary over several orders of magnitude annually. This poses particular difficulties in attempting to use simplified forms. These authors have obtained some promising results by using an exponential transformation in representing the output from the state variables, that is,

$$y_i(k) = \exp\,[x_i(k)] \qquad (4)$$

Thus the state variables of the equations are actually the logarithms of the biomass levels we wish to observe (for some reason past modelers have often assumed that the state-variables must be the biomass levels themselves).

Calibrations

The calibration process should do more than just assign values to the model parameters, it should supply us with numerical measures of various aspects of model performance. We need these performance measures in order to compare different models on a quantitative scale and to place confidence limits on our parameters and our predictions. Trial-and-error based calibration methods do not supply these important statistics and do not necessarily even provide the best estimates of parameters, hence we have begun to look toward highly data based calibration techniques.

We at Lake George are now at the stage in our modeling efforts where we are beginning to review existing calibration methods and assess their value for various situations. The current emphasis in calibration centers around various autoregressive techniques[12] which have several desirable qualities. These methods have been refined by researchers in the field of econometrics where simple regression techniques often fail due to high autocorrelation of the residual error terms. Ordinary least squares regression methods assume that the pattern of the residuals is random. Mathematically this amounts to the well known assumption

$$E(\epsilon_t, \epsilon_{t-s}) = 0 \quad \forall\, s \neq 0 \qquad (3)$$

Since we are dealing with time series of data, it is highly unlikely that this assumption will hold. The solution to this problem is to postulate an autoregressive scheme relating the ϵ_t's to themselves.

One such scheme is the simple first order equation proposed by Durbin[4]

$$\epsilon_t = \rho\epsilon_{t-1} + u_t \qquad (4)$$

where the u_t's now obey (3). Other schemes can involve higher order lags in order to follow oscillations in the residuals[6].

Since these auto-regressive techniques seem to show much promise, we plan to rely on them, at least in the early stages of our work. If we find that these methods do not live up to our expectations, there are other methods we may try such as Generalized least squares, Kalman filtering, Delta approximation, etc.

Examination

Refer again to Figure 2. We have discussed the first four steps of our modeling process, which brings us to model examination. We have shown that our procedure involves looking at many different formulations and also that the calibration process yields measures of performance. In the examination process we will compare these measures, which might include confidence intervals, F values, and goodness-of-fit measures. We will also examine such things as coefficient signs and magnitudes to determine whether or not they are reasonable from an ecosystem viewpoint and whether they will cause the model to be stable. Based on this examination some formulations may be rejected or needed structural changes may become obvious, in either case, we will return to the model structuring stage.

Evaluation

From the above examination procedure we will obtain a model or a set of models which are reasonable formulations and meet the objectives we stated at the outset. The sixth and final step in our procedure is to apply and evaluate these models as tools for policy impact analysis.

As part of this research project, studies are being made at Lake George to quantitatively relate level of development to changes in controlling factors of the lake ecosystem, especially nutrient and organic loadings and water temperature. By coupling this information with existing predictions of land-use development, we can note how these system inputs will be perturbed for various development scenarios. By using our models, we can analyze how these changes to lake inputs can be expected to modify lake characteristics, most notably the overall water quality in terms of algae and pollutants.

Summary Comments

In this paper we have:

1) Argued the need for a methodology for developing data-based ecosystem management models. Other lake ecosystem models have either highly complex coupled process submodels or system models embodying particular system interaction formalisms. Since they are in general not validated and rely heavily on theory we question their use for management purposes.

2) We have proposed a separate methodology which can take advantage of information contained in previously unavailable data sets. This methodology is not unlike other systems modeling methodologies. In particular we emphasize the need to develop the model with specific policy issues in mind.

We stress the use of data in both developing the model structure as well as in parameter identification.

3) We have partially demonstrated the use of this methodology considering the data set from Lake George. We have related our observations in the data to model structures which conform to our objectives.

It is too early in this research to offer conclusions either about the methodology or the Lake George ecosystem management model. Further work will hopefully yield an example of this methodology and a useful policy analysis tool for Lake George.

Present efforts are concentrated in the area of calibration of various model structures and evaluation of them quantitatively through performance measures and qualitatively with regard to macro-ecological realism and their ease of understanding and use.

References

1. Bledsoe, L. J., "Linear and Nonlinear Approaches for Ecosystem Dynamic Modeling", Systems Analysis and Simulation in Ecology, Vol. IV, B.C.Patten (Ed.), Academic Press, N. Y., 1976.

2. Bloomfield, J. and DiCesare, F., "The SIMPLE Model" unpublished manuscript, R.P.I., Troy, N.Y., 1975.

3. DiToro, D. M., O'Connor, D. J. and Thomann, R.V., "A Dynamic Model of the Phytoplankton Population in the Sacramento-San Joaquin Delta", in Advances in Chemistry Series #106, American Chemical Society, Washington,D.C. 1971.

4. Durbin, J., "Estimation of Parameters in Time - series Regression Models", J. Royal Statist. Soc., Series B, Vol. 22, pp 139-153, 1960.

5. Gray, A. C., Jr., "Nutrient Effects on Phytoplanktonic Primary Production in the Lake George Ecosystem", Doctoral Thesis, R.P.I., Troy, N.Y., 1976.

6. Kashyap, R.L.and Rao, A.R., "Dynamic Stochastic Models from Empirical Data", Academic Press, N.Y.,1976.

7. Loucks, O.L., Prentki, R. T., Watson, V.J., Reynolds, B.J., Weiler, P.R., Bartell, S.M. and D'Alessio, A.B., "Studies of the Lake Wingra Watershed: An Interim Report", Report IES-78 Institute for Environmental Studies, Univ. of Wisconsin at Madison.

8. McNaught, D.C., "Determination of Secondary Production by Herbivores in Lake George", IBP Memo Report 72-68, Internation Biological Program, 1972.

9. Mulholland, R.J., and Keener, M. S., "Analysis of Linear Compartment Models for Ecosystems", Journal of Theoretical Biology, Vol. 44, pp 105-116, 1974.

10. Park, R. A., Scavia, D. and Clesceri, N.L., "CLEANER, The Lake George Model", Report #FWI 74-15, The Fresh-water Institute, R.P.I., Troy, N.Y.

11. Patten, B. C., Egloff, D. A. and Richardson, T.H., "Total Ecosystem Model for a Cove in Lake Texoma", in Systems Analysis and Simulation in Ecology, Vol. III, B.C. Patten, (Ed.), Academic Press, N.Y., 1975.

12. Poole, R. W., "Stochastic Difference Equation Predictors of Population Fluctuations", Theoretical Population Biology, Vol. 9, pp 25-45, 1976.

13. Williams, R. B., "Computer Simulation of Energy Flow in Cedar Bog Lake, Minnesota Based on the Classical Studies of Lindeman", in Systems Analysis and Simulation in Ecology, Vol. I, B.C.Patten, (Ed.), Academic Press, N.Y. 1971.

14. Williams, S. L., and Clesceri, Nicholas, L., "Diatom Populations Changes in Lake George, N.Y.", reports FWI 72-1 through FWI 72-8, Fresh-water Institute, R.P.I. Troy, N.Y., 1972.