Ecological Modelling, 20 (1983) 113–133
Elsevier Science Publishers B.V., Amsterdam – Printed in The Netherlands

CONFIRMATION OF WATER QUALITY MODELS

KENNETH H. RECKHOW

School of Forestry and Environmental Studies, Duke University, Durham, North Carolina 27706 (U.S.A.)

STEVEN C. CHAPRA

Environmental Engineering Division, Civil Engineering Department, Texas A&M University, College Station, Texas 77843 (U.S.A.)

(Accepted for publication 28 February 1983)

ABSTRACT

Reckhow, K.H. and Chapra, S.C., 1983. Confirmation of water quality models. *Ecol. Modelling*, 20: 113–133.

Water quality simulation models, whether descriptive or predictive, must undergo confirmatory analyses if inferences drawn from the models are to be meaningful. Current practices in the confirmation of simulation models are examined and criticized from this perspective. In particular, labeling this process "verification" or "validation" (truth) probably contributes to the often inadequate efforts, since these states are unattainable. The evaluation of scientific hypotheses, or water quality simulation models, may proceed according to inductive logic, the hypothetico-deductive approach, or perhaps according to a falsification criterion. The result of successful testing is at best confirmation or corroboration, which is not truth but rather measured consistency with empirical evidence. On this basis a number of statistical tests are suggested for model confirmation. The major difficulty to overcome, before confirmation becomes meaningful, is the generally inadequate data for establishing rigorous statistical tests.

INTRODUCTION

Mathematical model development begins with the conceptualization of the functions and relationships of the characteristics of the issue or system under study and proceeds with specification of the mathematical relationships, estimation of the model parameters, and validation of the model as a reliable representation. The process may be iterative. All steps are important; the validation step, however, may be the most important because it provides confirmation (or lack thereof) that the conduct of the other steps resulted in a reliable model. Ironically, validation is also the step that is most

often inadequately conducted in water resource model development.

In fact, validation, or the ascertainment of truth, is inconsistent with the logic of scientific research. As Anscombe (1967) notes, "The word valid would be better dropped from the statistical vocabulary. The only real validation of a statistical analysis, or of any scientific enquiry, is confirmation by independent observations." The testing of scientific models may be considered an inductive process, which means that even with true premises, we can at best assign high probability to the correctness of the model. Philosophers of science have long debated the appropriate criteria for the effectiveness of arguments of this nature, considering characteristics such as the severity of tests and the goodness-of-fit.

How can this be translated into statistical terms for practical applications? Generalization of verifying criteria is possible only to a limited extent; beyond that, issue-specific criteria must be determined. Still, guidelines may be proposed for the composition of tests that are rigorous and for the selection of goodness-of-fit tests and acceptance levels. Both model developers and model users should benefit from careful consideration and application of criteria for model confirmation.

These issues are not only of academic interest. In the past 15 years, a number of water quality simulation models have been developed and then promoted as predictive methods to aid in the management of environmental quality. In most cases, however, these models have not been subjected to a rigorous validation or confirmation. Therefore, the model user often has no assurance that the model will yield reliable and informative predictions. This has potentially serious consequences, since the inadequately confirmed planning model may lead to the implementation of economically inefficient or socially unacceptable water quality management plans. It is the purpose of this paper to outline some philosophical and statistical issues relevant to the problem of confirmation, and then to recommend appropriate applications of statistical confirmatory criteria.

PHILOSOPHY AND SCIENTIFIC CONFIRMATION

Until the 1950s, virtually all scientists and philosophers of science viewed the advancement of science and the scientific method as endeavors dominated by empiricism and logic. The empiricism of Hume and the deductive logic of Russell and Whitehead formed the basis for the approaches of the logical positivist and, later, the logical empiricist. In particular, the logical empiricists have enjoyed widespread support during the twentieth century.

Logical empiricism (see Hempel, 1965) is based on the presupposition that observations and logic advance scientific knowledge. For example, when a scientific hypothesis (model) is proposed under logical empiricism, observa-

tions of relevant phenomena are acquired, and inductive or deductive logic is used to determine the degree of confirmational support. Inductive arguments, we may recall, cannot strictly be proven as true, but can at best be assigned a high likelihood of being correct. In statistical inference, inductive arguments are often associated with reasoning from the specific to the general. Deductive logic (reasoning from the general to the specific), on the other hand, must yield true conclusions if the premises are true and the arguments are valid.

Logical empiricists are divided on the importance and appropriate applications of deduction and induction in science. For example, under the hypothetico-deductive approach (Kyburg, 1970), a scientific hypothesis is proposed and criteria that can be tested are deduced logically. The scientist then must be concerned with constructing rigorous tests which, depending on rigor and the results of testing, confer a degree of confirmation upon the hypothesis. When competing hypotheses are offered, philosophers have recommended acceptance of the simplest one that is consistent with the empirical evidence, possibly because it is most probable (Kyburg, 1970).

Inductive logic, on the other hand, is important in a class of problems concerned with statistical explanation (Salmon, 1971). Scientists and philosophers who subscribe to this approach argue that there are many scientific analyses in which the information content of the conclusion exceeds that of the premises. In those circumstances, inductive logic is appropriate, and we, at best, can assign high probability to the conclusion based on the premises. Alternatively, using reasoning similar to Bayes' Theorem we may state that the degree of confirmation, or the probability of a hypothesis [$P(\mathrm{H})$], is conditional on the available empirical evidence (E):

$$P(\mathrm{H}|\mathrm{E}) = \frac{P(\mathrm{H}, \mathrm{E})}{P(\mathrm{E})} \tag{1}$$

In contrast, Edwards (1972) advocates a likelihood interpretation (without Bayes' Theorem) for the support provided to a hypothesis from a set of data; Rosenkrantz (1977), however, takes a strictly Bayesian view and conditionalizes this likelihood with an often information-less prior.

Popper (1968) has proposed a variation on the hypothetico-deductive approach that has undergo a variety of interpretations since its introduction (Brown, 1977). Popper rejects the notion that induction should be called logic, since the nature of induction is to support a conclusion that contains more information than the premises. Therefore, if we accept the logical empiricist view that science is based on logic, then deductive logic is necessary. Consistent with the hypothetico-deductivists, Popper requires the deduction of observational consequences of a scientific hypothesis; but in a break from previous thought, he bases scientific knowledge on a criterion of

falsification rather than confirmation. This means, according to Popper, that scientific statements are distinguished, not by the fact that they can be confirmed by observation, but rather by the fact that they can be falsified by observation. Popper believes that candidate hypotheses should be subjected to severe tests, and from among the successful hypotheses, the one that is deemed most falsifiable is the one that should tentatively be accepted. Although this may at first seem counter-intuitive, it is reasonable since, following the application of severe tests, the hypothesis that was most likely to be falsified yet survived is the hypothesis receiving the greatest empirical support. Popper would then say that this highly falsifiable hypothesis had been corroborated through the application of rigorous tests. Like confirmation, corroboration has a vague quantitative meaning, associated with the severity of the applied tests and the degree of test success.

Finally, to complete this discussion of the philosophy and methods of science, we must consider the thinking of some philosophers during the past 30 years (see Kuhn, 1970; or Brown, 1977). Specifically, it has been suggested that the logical empiricist notion of observation and logic, as being fundamental to scientific research, biases our view of science. When these presuppositions are eliminated, other criteria may become important, such as the consensus of opinion of the scientific community (Brown, 1977). Under one view of this new philosophy, most scientific research is "normal science" (Kuhn, 1970), in which the existing theoretical framework determines the research and the nature of the scientific inferences drawn from the observations. In contrast to normal science, scientific "revolutions" change the basic theoretical framework. These ideas represent an important new philosophical view of the conduct of science. Without rejecting this view, however, we may justifiably consider logical empiricism as the dominant theoretical framework at present. This means that methods proposed for the testing of mathematical models must draw their support from the philosophy and methods of the logical empiricists.

Independent of the preference we may have for a particular logical empiricist approach for evaluating scientific hypotheses, there are consistencies among the approaches that the scientist should note well. Without doubt, tests must be rigorous. This means that the hypothesis should be subjected to conditions that are most likely to identify its weaknesses or falsity. Mathematical simulation models must be tested with data that reflect conditions that are noticeably different from the calibration conditions; without this, there is no assurance that the model is anything more than a descriptor of a unique set of conditions (i.e., those representing the calibration state). To assess the degree of confirmation or corroboration that a hypothesis or model should enjoy, a statistical goodness-of-fit criterion is necessary. Finally, the modeler should prepare a set of candidate model

formulations and then base the model choice, in part, on the relative performance of the models on statistical tests and on consistency with theoretical system behavior. Comparison of rival models/hypotheses is an important step in the testing of scientific hypotheses.

The severity of the tests employed is often dependent on the intended use of the hypothesis or model. For example, the user normally faces a risk associated with the application of an incorrect model and a cost associated with testing candidate models. A cost/risk trade-off determines the appropriate level of test severity. Likewise, the statistical criterion is use-dependent. Specifically, it is noted below that the needs of a particular application generally determine the best criteria for assessing statistical goodness-of-fit.

SOME PRACTICAL ISSUES

The selection of a statistical test for the confirmation of a mathematical model may be facilitated through consideration of the following issues:

1. What characteristics of the prediction are of interest to the modeler? The answer may be one or more of: mean values, variability, extreme values, all predicted values, and so forth. If one of the limited, specific responses is given, then the test statistical criterion should focus on that specific feature.

2. Is it intended that the model be primarily descriptive (identifying hypothesized cause–effect relationships) or primarily predictive? Different statistical tests are appropriate in each case.

3. What is the criterion for successful confirmation? In statistical inference, mean square error is often adopted, although many statistical tests (e.g. nonparametric methods) do employ other error criteria. In some situations, a decision theoretic approach such as regret minimization is warranted (see Chernoff and Moses, 1959).

4. Are there any peculiar features to the model application of concern? This is a "catch-all" question intended to alert the model user to the fact that each application is unique, and therefore the confirmation process must be designed on that basis. For example:

   a. When prediction and observation uncertainty are considered, are all error terms quantified? It should be noted that model specification error is rarely estimated for water quality simulation models. This means that the corresponding prediction error is underestimated. Omission or mis-specification of any prediction/observation error terms will influence model confirmation.

   b. Are the assumptions behind any of the statistical tests violated? In particular, since time series or spatial series of data are often examined in model confirmation, autocorrelation may be a problem. When the validity of the statistical procedures is sensitive to a violated assumption (as is

generally true for the assumption of independence), then some modifications or alternatives must be considered.

CONFIRMATION OF SIMULATION MODELS: LITERATURE REVIEW

Although there have been few, if any, rigorous attempts at confirmation of a water quality simulation model, this is not because of a complete lack of attention to this issue in the recent literature. General discussions on the importance of model confirmation, or on confirmation as a step in simulation model development, are noteworthy in this regard (Van Horn, 1969; Naylor and Finger, 1971; Mihram, 1973; Davis et al., 1976; Caswell, 1977). In addition, the Environmental Protection Agency recently sponsored a workshop on this topic (Hydroscience, 1980). Unfortunately, the discussion

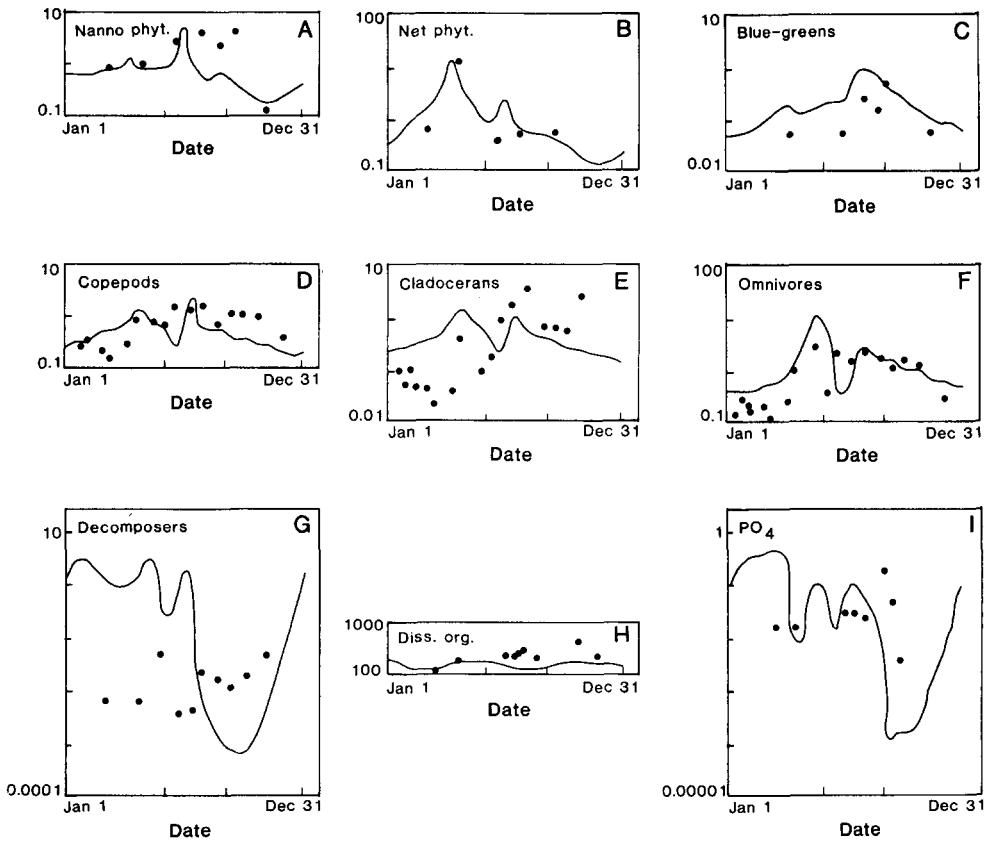

Fig. 1. An example of a comparison of observations with predictions for a water quality simulation model.

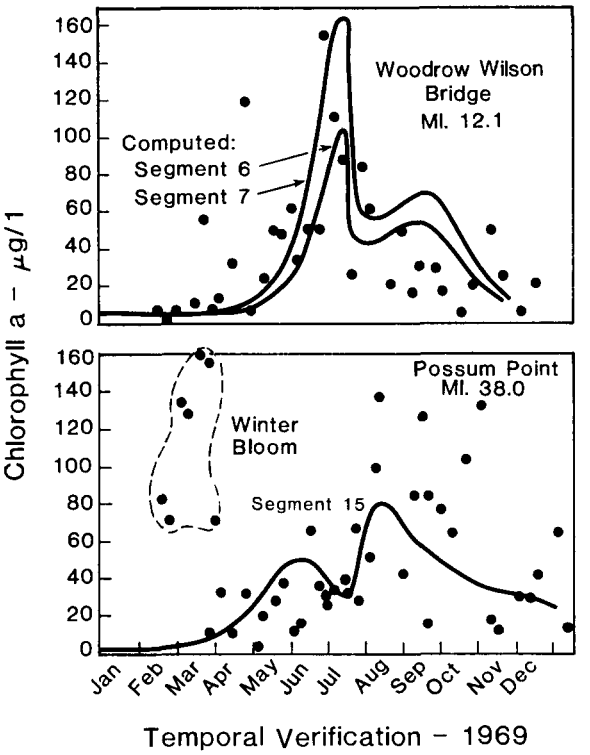

Fig. 2. An example of a model verification plot.

groups convened as part of this workshop generally offered a mixed or mild endorsement of rigorous statistical confirmatory criteria. While some of the papers presented at this workshop contained strongly-worded statements on the importance of confirmation (e.g. Velz) or presented statistical criteria for confirmation (e.g. Thomann, or Chi and Thomas), the workshop missed an opportunity to produce and to promulgate a set of confirmatory guidelines necessary for proper model development.

Despite claims by some that statistical confirmatory criteria should generally not be recommended because of the unique demands of each model application, there are in fact statistical tests that may be adapted to virtually any situation. A number of researchers (Mihram, 1973; Shaeffer, 1980; Thomann, 1980, Thomann and Segna, 1980; Thomann and Winfield, 1976) have suggested various statistics that are useful for model confirmation. Thomann's work, in particular, stands out as one of the few statistical statements on confirmation that are specific to water quality simulation models. In the field of simulation modeling in hydrology, James and Burges (1981) have prepared a useful practical guide to model selection and calibra-

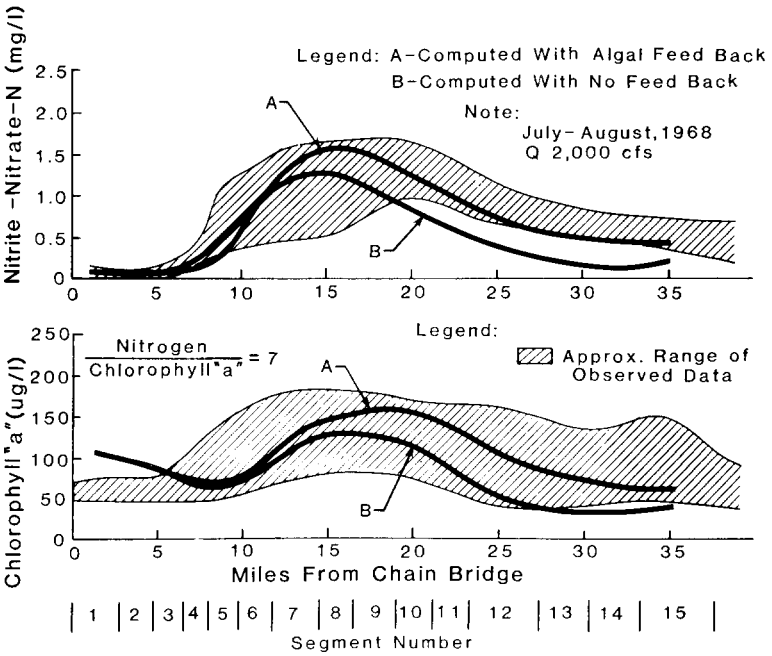

Fig. 3. An example of a model verification plot with observation error shown.

tion. Many of the statistical tests that they propose and apply to calibration are equally applicable for confirmation. In further support of the argument against those reluctant to adopt statistical confirmatory criteria, consider the alternative. Figures 1–3 were selected from water quality simulation literature as examples of current practice in prediction/observation comparison or model "validation." Commenting on Fig. 1, the authors described the prediction/observation match as "good." Ignoring this debatable judgment on the fit, we may still question any conclusion on model fit in the absence of statistical tests (which are clearly possible with these quantitative measures). Figures 2 and 3 are specifically labelled as "verification" plots, yet there is again no statistical goodness-of-fit criterion. In Fig. 3, the authors take the commendable step of displaying a shaded region for the observations. Regrettably, this region is so large that it may detract from meaningful confirmation by preventing model discrimination. A shaded observation region in conjunction with a time series for central tendency would be a more useful representation of the observations.

Several investigators (Meyer, 1971; Miller, 1974; Miller et al., 1976; Hornberger and Spear, 1980; 1981; Spear and Hornberger, 1980; Majkowski et al., 1981) have advocated the use of sensitivity analysis for the confirma-

tion and evaluation of simulation models. It is clear that a measure of the importance and error levels of model terms can be helpful during model development and can also provide some insight into the confirmation process. Gardner et al. (1981) caution that the standard form of sensitivity analysis (partial derivative with respect to the parameter of concern) may result in a misleading approximation of effect when nonlinearity and parameter covariance are large. In contrast, Hornberger and Spear (1980, 1981) avoid this problem by using Monte Carlo simulation in their sensitivity analyses. Factorial methods from experimental design can also be useful in simulation model studies of single variable and interaction effects. Sensitivity analysis, then, does not yield a measure of model confirmation, but it can provide information that is extremely useful in model testing and development.

Several studies (Beck, 1980; Fedra, 1980; 1981; Fedra et al., 1980) at the International Institute for Applied Systems Analysis have examined issues in the development, calibration, confirmation, and prediction of water quality simulation models. Sensitivity analysis and Monte Carlo simulation have been used to examine the impact of error terms on the prediction error. Among the strong conclusions apparent from this work (and from examination of Fig. 1) is that data are often inadequate for effective calibration and confirmation of mathematic models. This situation clearly limits the degree to which we may apply statistical confirmatory criteria and ultimately affects the reliability of planning models and methods.

STATISTICAL METHODS FOR CONFIRMATION

Several statistical methods may be found useful for assessing the degree of confirmation of a mathematical model. Some of the more common tech-

TABLE I

Statistical methods for confirmation

1. Deterministic Modeling	2. Stochastic Modeling
(a) measures of error	(a) deterministic method for particular percentile
(b) t-test	(b) probability density function "slices"
(c) Mann–Whitney–Wilcoxon test	(i) chi square test
(d) regression	(ii) Kolmogorov Smirnov test
(e) cross-correlation	(iii) comparison of moments
(f) graphical comparisons—box plots	(iv) graphical comparisons—box plots

niques are listed in Table I. Before selecting a technique (or, for that matter, before acquiring data for model development), the investigator should consider the issues presented in previous sections of this chapter. For example, it is likely that model applications are primarily concerned with only certain features of the model. It is appropriate, then, for confirmation to focus on those features of concern. In addition, statistical assumptions must be considered. Common assumptions include normality, homogeneity of variance, and independence. Many procedures are robust to mild violations of the first two assumptions, but not to lack of independence. Transformations may often be applied to achieve approximate normality or to stabilize variance, while some robust and nonparametric procedures mentioned below may be useful under non-normality.

Violation of the independence assumption poses more difficult problems. Predictions and observations in water quality simulation are often time series, and autocorrelation may be present in one or both of these series. In a dependent (autocorrelated) series, the information content is less than in an equivalent-length, independent series, because each data point, to some degree, is redundant with respect to the preceding point. This means that confidence intervals and significance tests that falsely assume independence will be overly optimistic, i.e. the intervals will be too small.

For a single data series x_i, the autocorrelation coefficient for lag k is defined as:

$$r_k = \frac{\text{covar}[x_i, x_{i+k}]}{\text{s}(x_i)\text{s}(x_{i+k})} \tag{2}$$

where covar is a covariance and s is a standard deviation. When predictions and observations are compared, the Durbin–Watson test may be used to examine the residuals for autocorrelation (see Wonnacott and Wonnacott, 1981). However, as Wonnacott and Wonnacott note (page 232), the estimate of autocorrelation from residuals tends to be low. In fact, Lenton et al. (1973) observe that the small sample sizes usually found in water resources can cause large estimation errors for the autocorrelation coefficient.

Fortunately, when autocorrelation is found and quantified, there are some steps that can then be taken to permit application of many of the standard statistical tests. Yevjevich (1972) presents several relationships for calculating the effective sample size, which is the size of an independent series that contains the same amount of information as contained in the autocorrelated series. For example, in the common situation when the lag one autocorrelation coefficient is positive and the first-order Markov model is appropriate, the effective sample size $N(e)$ is:

$$N(e) = N\left(\frac{1 - r_1}{1 + r_1}\right) \tag{3}$$

where N is the actual sample size.

A less efficient but computationally easy alternative to $N(e)$ is to use or to aggregate data covering intervals greater than the period of autocorrelation influence. For example, weekly data may exhibit autocorrelation, but monthly data may not; therefore, confine the analysis to monthly data.

A particularly useful method for time series confirmation is cross-correlation of the prediction and observation series. Here, too, autocorrelation is important, but methods have been developed to address the problem. Yevjevich (1972) presents an equation from Bartlett (1935) for the effective sample size in two autocorrelated series:

$$N(e) = \frac{N}{1 - 2r_i(x)r_i(y) + \ldots + 2r_k(x)r_k(y)} \tag{4}$$

where $r_i(x)$ is the lag i autocorrelation for series x, and $r_i(y)$ is the lag i autocorrelation for series y. $N(e)$ then may be used in the significance test for the cross-correlation coefficient. Note that $N(e)$ equals N if one of the two series contains no autocorrelation.

An alternative solution to autocorrelation in cross-correlation analysis is prewhitening (Box and Jenkins, 1976; McCleary and Hay, 1980). Under this procedure, the Box–Jenkins methods are used to transform each series into a white noise process. Cross-correlation analysis is then performed on the two white noise series.

To summarize this discussion on autocorrelation, considerable attention has been devoted to this topic because of the impact of lack of independence on statistical tests of significance. Further, even those water quality simulation model studies that have employed statistical confirmatory criteria—for example, Thomann and Segna, 1980—have neglected autocorrelation in situations where it is probably present. Since "data generated by dynamic simulation (models) are usually highly correlated" (Naylor, 1971) and since time series of natural phenomena may also be correlated, the attention paid to autocorrelation seems appropriate.

Following consideration of these application-specific and statistical issues that help to determine the model terms and statistical methods to be involved in confirmation, the investigator is likely to employ one or more of the techniques listed in Table I. Graphical examination of data sets or series is usually a necessary part of any statistical analysis, and it certainly can be helpful in model confirmation. However, it is not listed in Table I because of concern that confirmation will begin and end with a graphical study and thus not advance beyond Fig. 1–3 mentioned above.

One view of model output and, hence, confirmation approaches, leads to the separate groupings of deterministic and stochastic modeling. Most water quality simulation models are deterministic, and this limits the set of

available statistical methods. The trend toward error analysis in modeling is important for model confirmation, although difficulties in the estimation of model error may restrict the confirmation study. Some of the methods listed in Table I under deterministic modeling use aggregated data (prediction and observation samples), and some of the methods are appropriate for data series. Under "measures of error," we may include various weighting functions for the difference between the predictions and the observations. For example, the relative error is:

$$\text{relative error} = \frac{|x_{\text{obs}} - x_{\text{pred}}|}{x_{\text{obs}}} \tag{5}$$

Another alternative, the squared error, is:

$$\text{squared error} = (x_{\text{obs}} - x_{\text{pred}})^2 \tag{6}$$

In each case, the average value (e.g., mean square error) is the appropriate form for expressing these error terms.

However, to assess the degree of confirmation (beyond a relative comparison of models), we need to use a test of statistical significance such as the t–test (parametric) or the Mann–Whitney–Wilcoxon test (nonparametric). Both tests require assumptions of independent identically-distributed (i.i.d.) observations, but the t–test adds a normality assumption. While the t-test is fairly robust to violations of the normality assumption (Box et al., 1978), neither test is robust to violation of independence.

The t–test is conducted from Student's t distribution, which, for a null hypothesis of no difference between the mean of the observations ($\bar{x}_{\text{obs}}$) and the mean of the predictions ($\bar{x}_{\text{pred}}$), is expressed as:

$$t = \frac{\bar{x}_{\text{obs}} - \bar{x}_{\text{pred}}}{s_{\text{dif}}/\sqrt{n}} \tag{7}$$

where s_{dif} is the standard deviation of the paired differences ($x_{\text{obs}} - x_{\text{pred}}$), and n is the number of ($x_{\text{obs}} - x_{\text{pred}}$) pairs.

The nonparametric alternative to the t–test is the Mann–Whitney or Wilcoxon test. This procedure, which may be preferred under certain conditions of nonnormality, although perhaps not strongly so (see Box et al., 1978), is based on the relative ranks achieved when the data are ordered. Both the t–test and the Mann–Whitney test are clearly described in Snedecor and Cochran (1967).

A second set of related statistical methods useful for model confirmation are regression and correlation. Here, the method chosen would be used to relate one data series to another, and thus autocorrelation is again a problem. Following analysis and adjustment (if necessary) for autocorrela-

tion, the investigator may regress the predictions on the observations. The fit may be assessed through the standard error statistic, or perhaps using the reliability index proposed by Leggett and Williams (1981). This index reflects on a plot of predictions vs. observations, the angle between the 1 : 1 line (line of best fit), and a line through each data point from the origin. Cross-correlation (see Davis, 1973) is calculated in a manner similar to that for the Pearson product moment correlation. Davis (1973) provides a test statistic for assessing the significance of the cross-correlation coefficient between two series. Remember that it is important to adhere to the assumptions behind the statistical methods; failure to do this under certain conditions can lead to faulty inferences concerning confirmation.

The final method presented here for deterministic model confirmation (from a by no means exhaustive list of options in Table I) is the box plot (McGill et al., 1978; Reckhow, 1980). The box plot is an extremely informative method for graphing one or more sets of data for the purpose of comparing order statistics. For each data set, the plot displays the median, the relative statistical significance of the median, the interquartile range, and the minimum and maximum points (see Fig. 4). With aggregated (non-series)

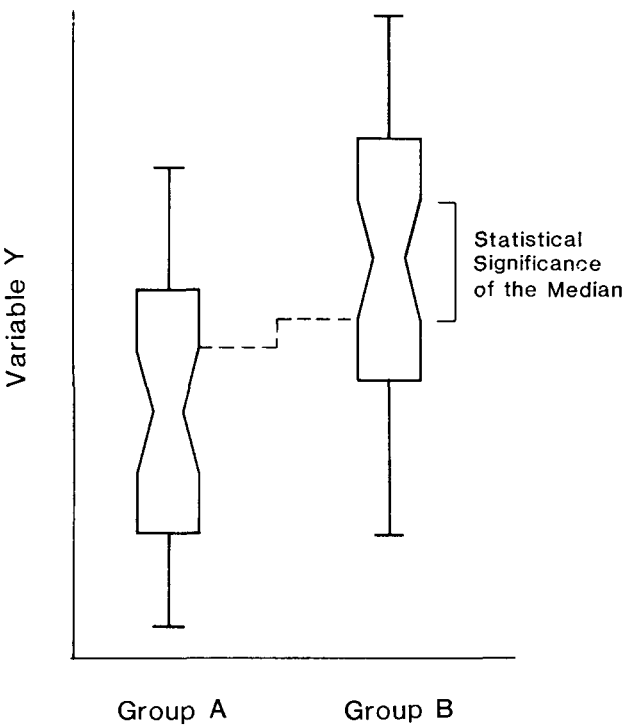

Fig. 4. Box plots possessing significantly different medians.

data, the box plot yields perhaps the best visual comparison of two or more data sets. A detailed example illustrating box plot construction and several applications is presented in Reckhow and Chapra (1983). McGill et al. (1978) and Reckhow (1980) also describe the construction and interpretation of box plots.

Before examining some statistical options for confirming stochastic models, consider one of the important practical issues that the model developer must face with a multivariate deterministic model. Specifically, the model developer may have calculated a confirmation statistic (e.g., a cross-correlation coefficient) for each variable in the model and/or for certain features of the model such as extreme values. How might these confirmatory statistics be aggregated into a single confirmation measure?

First, the modeler must realize that to aggregate statistics and to make the confirmation decision on the basis of a single measure means the loss of potentially valuable model evaluation information. If this loss is acceptable, then the modeler must decide on an aggregation scheme for the individual confirmation statistics, for example, for the cross-correlation coefficients. This decision should be based on the relative importance of the model characteristics (e.g., model variables) for which confirmation statistics are available. The confirmation statistics then are aggregated using weights reflecting this importance. For example, if chlorophyll and dissolved oxygen are deemed most important in an aquatic ecosystem model, then the con-

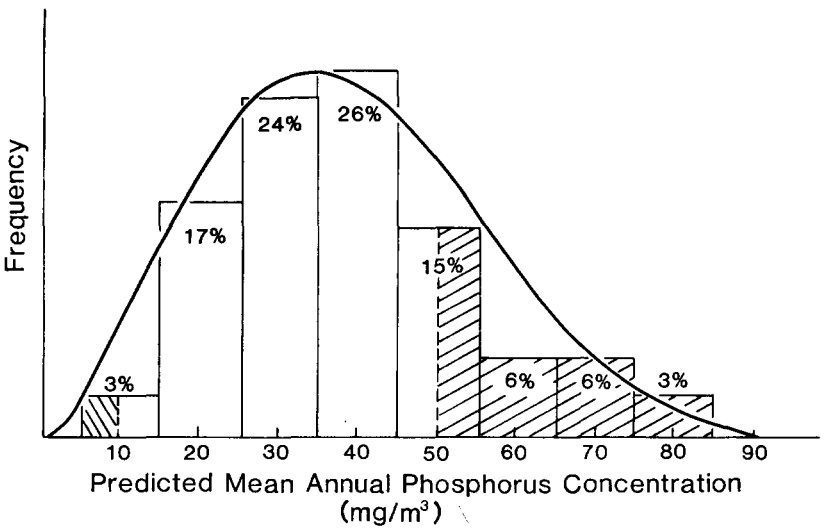

Fig. 5. A comparison of a histogram with a probability density function for the chi-square test.

firmatory statistics for these two variables should receive the highest weights. The final confirmation measure is a weighted sum of, for example, cross-correlation coefficients:

$$\text{Confirmation Measure} = \Sigma w(x) r(x) \tag{8}$$

where x is the model characteristic (variable), w is the weight reflecting the importance of x, and r is the cross-correlation confirmation statistic for x. The particular scheme presented above is merely meant to suggest one option for aggregating statistics; others are certainly possible.

Less common, at present, than the deterministic simulation model, the stochastic model is nonetheless important, and is quite amenable to a number of statistical confirmatory approaches. In fact, all the methods discussed above for the deterministic model are appropriate for a number of

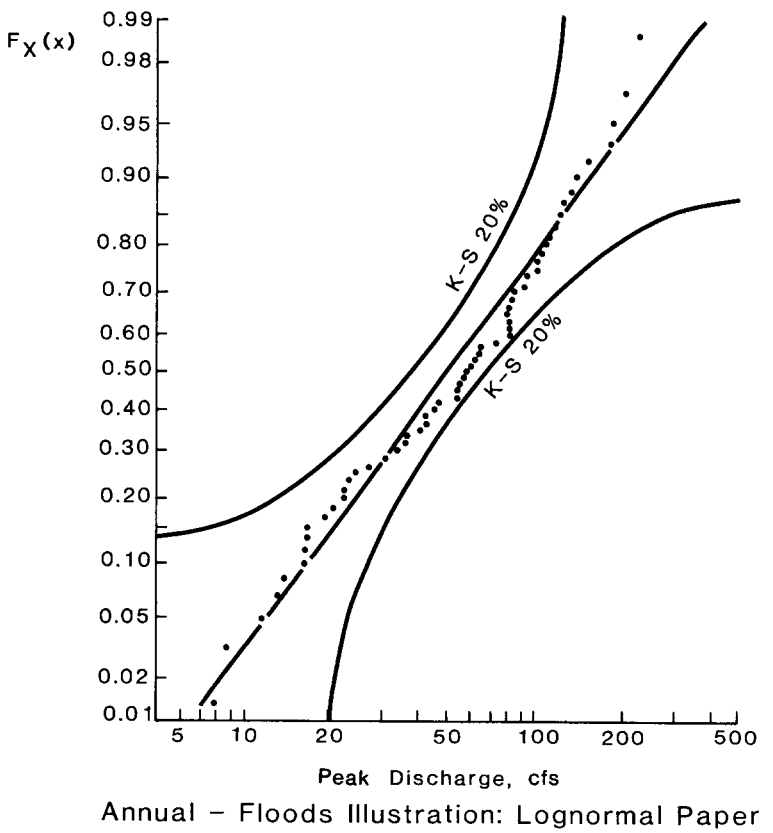

Fig. 6. A lognormal probability plot illustrating the Kolmogorov–Smirnov test. (Benjamin and Cornell, 1970)

features (e.g. the time stream of mean values) of the three-dimensional prediction and observation surfaces. In addition, we can take a two-dimensional slice of these three-dimensional distributions. Several statistical goodness-of-fit tests listed in Table I can then be employed.

The chi-square test (Fig. 5) and the Kolmogorov–Smirnov test (Fig. 6) yield test statistics based on the comparison of two distributions. The chi-square statistic, χ^2, is calculated as the sum:

$$\chi^2 = \sum_i \frac{(n_{i,\text{obs}} - n_{i,\text{pred}})^2}{n_{i,\text{obs}}} \tag{9}$$

where $n_{i,\text{obs}}$ is the number of observations in cell i, and $n_{i,\text{pred}}$ is the number of predictions in cell i. The chi-square test is conducted from the probability density function (pdf); each pdf is divided into a number of cells from which the chi-square counts are made. The somewhat arbitrary nature of the pdf division can unfortunately affect the results of the test. Benjamin and Cornell (1970) provide an excellent discussion of the merits of the chi-square test, as well as a table of chi-square statistics for the significance of the test.

The Kolmogorov–Smirnov test is perhaps preferable to the chi-square test because it is based on the cumulative distribution function (cdf). This removes the arbitrariness and investigator influence, because cells are not required; rather, the data are ordered and the deviations of the order statistics are examined. In Fig. 6, the fit of a model is examined and the

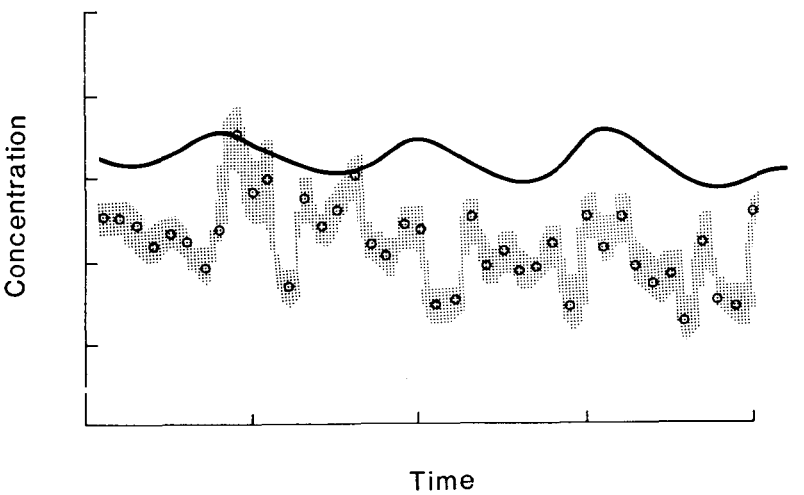

Fig. 7. Prediction–Observation comparison for model confirmation: continuous point observations.

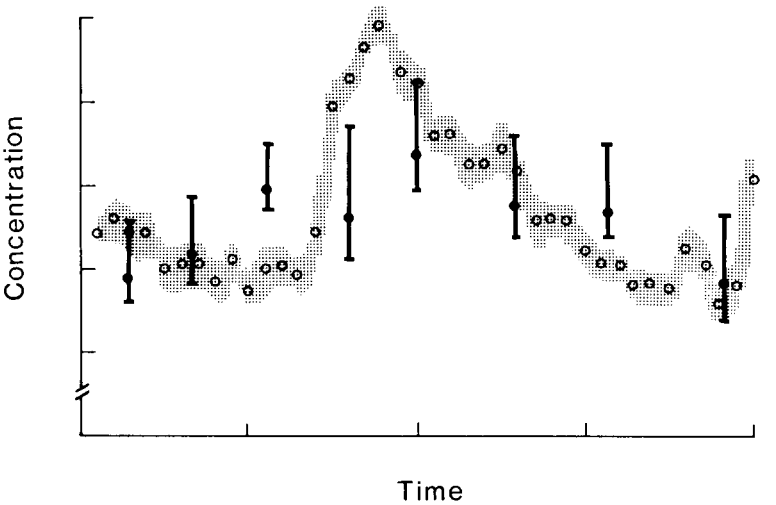

Fig. 8. Prediction–Observation comparison for model confirmation: error bars on observations.

goodness-of-fit is portrayed graphically with the Kolmogorov–Smirnov statistic.

Either of these methods may be used when the data are arranged in histogram or cumulative distribution form. Test statistics are available to assess the degree of confirmation. Benjamin and Cornell (1970) present a superb discussion on these goodness-of-fit methods for probability models.

Other statistics or procedures may certainly be considered to support model confirmation. The comparison of moments, particularly higher moments, can be useful in some situations (see Benjamin and Cornell, 1970). In addition, the box plot is quite effective for examining and displaying differences between distributions.

From a practical standpoint, the stochastic water quality simulation modeler probably does not have two three-dimensional distribution representing predictions and observations, respectively. Rather, he may have a continuous prediction "cloud" representing one standard error around the central points, and either continuous point observations (Fig. 7), or discrete observations with error bars (Fig. 8). If the prediction error cloud does not include all error (is model error included?), then the error region is too small and "rejection" is more likely. Otherwise, the modeler for Fig. 7 could cross-correlate the two central time series or test the overlap of the error region with the observation line. For Fig. 8, the modeler could compare the points and regions statistically during the discrete observation times. Remember that the preferred approach involves comparison of a number of candidate models.

In closing this discussion of statistical methods for confirmation, two additional model types deserve mention: cross-sectional regression models and descriptive (causal) models. Confirmation of cross-sectional regression models does not often pose the problem experienced with simulation models. This is particularly true if the cases studied are truly representative (e.g. a random sample) of the population of concern. Nonetheless, "shrinkage of the coefficient of multiple correlation" (Stone, 1974) between model development and application data sets is to be expected. The methods of cross-validation (viz. calibrate on half of the data and confirm on the other half; if no significant difference occurs, recalibrate using all the data) and the jackknife are useful for both regression model confirmation and estimation of a non-shrinking standard error or correlation. Stone (1974) and Mosteller and Tukey (1977) discuss these methods in detail.

To this point, the proposed statistical confirmatory methods are largely intended for predictive applications of models. Another important use of mathematical models is as descriptors of hypothesized causal relations. The methods presented above may also be found useful for confirming descriptive (causal) models. In addition, though, a causal confirmation process may be proposed.

The causal mode confirmation procedure is based on a comparison of synthetic data generated from the model with actual observations. The statistical methods employed are path analysis (Kenny, 1979) and the confirmatory approaches of Joreskog for linear structural equations (Jöreskog and Sörbom, 1978; Kenny, 1979). The following procedure is recommended:

1. Experimental design and/or Monto Carlo simulation are used to generate "data" from each model. It is important that a number of plausible candidate models be evaluated.

2. The statistical tests presented in Table I may be used to compare real and simulated output data.

3. Using path analysis and LISREL (Jöreskog and Sörbom, 1978), construct linear structural equation models for both the real and synthetic data. These models are intended to represent causal behaviour, from the point of view of the real data and from the point of view of each candidate simulation model.

4. Use the statistical tests presented in Table I to compare the real and synthetic structural equation models. The degree of confirmation of the causal model is then a measure of confirmation of the descriptive nature of the original simulation model. Several evaluative criteria may be posed:

(a) What "parts" of the model are most consistent (inconsistent) with observation? To what extent?

(b) What changes might be appropriate? This suggests an iterative approach

alternating model development with causal confirmation analysis.
(c) How do the models (and model subroutines) compare in performance?
It appears likely that there are many applications of path analysis and confirmatory methods for structural equations in the areas of simulation model development and testing.

CONCLUDING COMMENTS

To end this discussion of the philosophy and statistics of simulation model confirmation, a few points deserve restating.

1. Inadequate model confirmation increases the risks associated with the application of the model. Admittedly, there is a data cost and an analysis cost associated with model confirmation. This cost is to be compared with the risk resulting from the use of an unconfirmed model.

2. If confirmation is to be meaningful:
(a) rigorous statistical tests must be applied; and
(b) calibration-independent data are needed.

3. A number of plausible candidate models (or model sub-routines) should undergo confirmation. Comparison of the performance of the candidates aids the modeler in the determination of the degree of confirmation.

Finally, it must be recognized that the proposed confirmation criteria can rarely be applied in practice to the extent outlined in this paper. This realization does not reduce the importance of these criteria. Rather, a confirmation goal has been proposed, and the modeler may assess the extent of achievement of this goal. This degree of confirmation, estimated in terms of test rigor, test success, and data set independence, represents a measure of confidence to be assigned to the model as a predictive tool.

REFERENCES

Anscombe, P.J., 1967. Topics in the investigation of linear relations fitted by the method of least squares, J.R. Stat. Soc., B, 29: 1–52.

Bartlett, M.S., 1935. Some aspects of the time correlation problem in regard to tests of significance, J.R. Stat. Soc., 98: 536–543.

Beck, M.B., 1980. Hard or Soft Environment Systems? International Institute for Applied Systems Analysis Working Paper 80–25.

Benjamin, J.R. and Cornell, C.A., 1970. Probability, Statistics and Decision for Civil Engineers. McGraw Hill, New York, 684 pp.

Box, G.E.P., Hunter, W.G. and Hunter, J.S., 1978. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley, New York, 653 pp.

Box, G.E.P. and Jenkins, G.M., 1976. Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco, CA, 553 pp.

Brown, H.I., 1977. Perception, Theory, and Commitment. University of Chicago Press, Chicago, IL, 202 pp.

Caswell, H., 1977. The Validation Problem. In: B.C. Patten (Editor), Systems Analysis and Simulation in Ecology, Vol. IV. Academic Press, New York, pp. 313–325.

Chernoff, H. and Moses, L.E., 1959. Elementary Decision Theory. John Wiley and Sons, New York, 364 pp.

Davis, D.R., Duckstein, L. and Kisiel, C.C., 1976. Model Choice and Evaluation from the Decision Viewpoint. In: Theories of Decision in Practice. Crane, Russak and Co., New York, pp. 341–351.

Davis, J.C., 1973. Statistics and Data Analysis in Geology. John Wiley and Sons, New York, 550 pp.

Edwards, A.W.F., 1972. Likelihood. Cambridge University Press, Cambridge, UK, 235 pp.

Fedra, K., 1980. Estimating Model Prediction Accuracy: A Stochastic Approach to Ecosystem Modeling. International Institute for Applied System Analysis Working Paper 80–168.

Fedra, K., 1981. Hypothesis Testing by Simulation: An Environmental Example. International Institute for Applied System Analysis Working Paper 81–74.

Fedra, K., Van Straten, G. and Beck, M.B., 1980. Uncertainty and Arbitrariness in Ecosystem Modeling: A Lake Modeling Example. International Institute for Applied Systems Analysis Working Paper 80–87.

Gardner, R.H., O'Neill, R.V., Mankin, J.B. and Carney, J.H., 1981. A comparison of sensitivity analysis based on a stream ecosystem model. Ecol. Modelling, 12: 173–190.

Hempel, C.G., 1965. Aspects of Scientific Explanation. The Free Press, New York, 504 pp.

Hornberger, G.M. and Spear, R.C., 1980. Eutrophication in Peel Inlet I. The problem-defining behavior and mathematical model for the phosphorus scenario. Water Res., 14: 29–42.

Hornberger, G.M. and Spear, R.C., 1981. An approach to the preliminary analysis of environment systems. J. Environ. Manage., 12: 7–18.

Hydroscience, 1980. Workshop on Verification of Water Quality Models. U.S. Environmental Protection Agency, EPA–600/9–80–016.

James, L.D. and Burges, S.J., 1981. Selection, calibration, and testing of hydrologic models. Am. Soc. Agric. Eng. Monogr., 68 pp.

Jöreskog, K.G. and Sörbom, D., 1978. LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood. User's Guide, Version IV. National Educational Resources, Chicago, IL, 165 pp.

Kenny, D.A., 1979. Correlation and Causality. John Wiley and Sons, New York, 277 pp.

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, 2nd edn. University of Chicago Press, Chicago, IL, 210 pp.

Kyburg, Jr., H.E., 1970. Probability and Inductive Logic. Macmillian, London, 272 pp.

Leggett, R.W. and Williams, L.R., 1981. A reliability index for models. Ecol. Modelling, 13: 303–312.

Lenton, R.L., Rodriguez-Iturbe, I. and Schaake, Jr., J.C., 1973. A Bayesian Approach to Autocorrelation Estimation in Hydrologic Autoregressive Models. MIT Water Resources Rpt., p. 163.

Majkowski, J., Ridgeway, J.M. and Miller, D.R., 1981. Multiplicative sensitivity analysis and its role in development of simulation models. Ecol. Modelling, 12: 191–208.

McCleary, R. and Hay, Jr., R.A., 1980. Applied Time Series Analysis for the Social Sciences. Sage Publications, Beverly Hills, CA, 331 pp.

McGill, R., Tukey, J.W. and Larsen, W.A., 1978. Variations of box plots. Am. Stat., 32: 12–16.

Meyer, C.F., 1971. Using experimental models to guide data gathering. ASCE Journal, Hydro. Div., HYIO: 1681–1697.

Mihram, G.A., 1973. Some practical aspects of the verification and validation of simulation models. Oper. Res. Q., 23: 17–29.

Miller, D.R., 1974. Sensitivity analysis and validation of simulation models. J. Theor. Biol., 48: 345–360.

Miller, D.R., Butler, G. and Bramall, L., 1976. Validation of ecological system models, J. Environ. Manage., 4: 383–401.

Mosteller, F. and Tukey, J.W., 1977. Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading, MA, 588 pp.

Naylor, T.H., 1971. Computer Simulation Experiments with Models of Economic Systems. John Wiley and Sons, New York, 502 pp.

Naylor, T.H. and Finger, J.M., 1971. Validation. In: T.H. Naylor (Editor), Computer Simulation Experiments with Models of Economic Systems. John Wiley and Sons, New York, pp. 153–164.

Popper, K.R., 1968. The Logic of Scientific Discovery. Harper and Row, New York, 480 pp.

Reckhow, K.H., 1980. Techniques for exploring and presenting data applied to lake phosphorus concentration. Can. J. Fis. Sci., 37: 290–294.

Reckhow, K.H. and Chapra, S.C., 1983. Engineering approaches for Lake Management, Vol. 1. Data Analysis and Empirical Modeling. Ann Arbor Science Publishers, Ann Arbor, MI, 340 pp.

Rosenkrantz, R.D., 1977. Inference, Method and Decision: Toward a Bayesian Philosophy of Science. D. Reidel Pub., Dordrecht, Holland, 262 pp.

Salmon, W.C., 1971. Statistical Explanation and Statistical Relevance. University of Pittsburgh Press, Pittsburgh, PA, 117 pp.

Shaeffer, D.L., 1980. A model evaluation methodology applicable to environmental assessment models, Ecol. Modelling, 8: 275–295.

Snedecor, G.W. and Cochran, W.G., 1967. Statistical Methods. The Iowa State University Press, Ames, IA, 593 pp.

Spear, R.C. and Hornberger, G.M., 1980. Eutrophication in Peel Inlet II. Identification of critical uncertainties via generalized sensitivity analysis. Water Res., 14: 43–49.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J.R. Stat. Soc., B, 36: 111–147.

Thomann, R.V., 1980. Measures of Verification. In: Workshop on Verification of Water Quality Models. Hydroscience, Inc., U.S. Environmental Protection Agency, EPA-600/9-80-016.

Thomann, R.V. and Segna, J.S., 1980. Dynamic Phytoplankton-Phosphorus Model of Lake Ontario: Ten Year Verification and Simulations. In: R.C. Loehr, C.S. Martin and W. Rast (Editors), Phosphorous Management Strategies for Lakes. Ann Arbor Science, Ann Arbor, MI, pp. 153–190.

Thomann, R.V. and Winfield, R.P., 1976. On the Verification of a 3-Dimensional Phytoplankton Model of Lake Ontario. In: W.R. Ott (Editor), Environmental Modelling and Simulation. U.S. E.P.A., 600/9-76-016, Washington, DC, pp. 568–572.

Van Horn, R., 1969. Validation. In: T. H. Naylor (Editor), The Design of Computer Simulation Experiments. Duke University Press, Durham, NC, pp. 232–251.

Wonnacott, T.H. and Wonnacott, R.J., 1981. Regression: A Second Course in Statistics. John Wiley and Sons, New York, 556 pp.

Yevjevich, V., 1972. Probability and Statistics in Hydrology. Water Resources Publications, Fort Collins, CO, 302 pp.