Preprint of: J.K. Vanclay, Skovsgaard, J.P. and O. García, 1996. Evaluating forest growth models. *In:* M. Köhl and G.Z. Gertner (eds) Caring for the Forest: Research in a Changing World: Statistics, Mathematics and Computers. Proceedings of the Meeting of IUFRO S4.11-00 held at IUFRO XX World Congress, 6-12 August 1995, Tampere, Finland. WSL, Birmensdorf, p. 11-22.

Evaluating Forest Growth Models

J.K. Vanclay, J.P. Skovsgaard and O. Garcia
Center for International Forestry Research, Bogor, Indonesia,
Danish Forest and Landscape Research Institute, Hørsholm, Denmark, and
Instituto Forestal, Santiago, Chile.

Abstract

Effective model evaluation is not a single, simple procedure, but comprises several interrelated steps that cannot be separated from each other or from the purpose and process of model construction. We draw attention to several statistical and graphical procedures that may be used both with data used for model calibration and with data used in the evaluation of the model. We emphasize that the validity of conclusions depends on the validity of assumptions. These principles should be kept in mind throughout model construction and evaluation.

1. Introduction

Model evaluation is an important part of model building, and some examination of the model should be made at every stage of model design, fitting and implementation. It should not merely be an afterthought or an acceptance trial. A thorough evaluation of a model involves several steps, including two which are often called *verification* and *validation*. In forest growth modelling, these usually denote *qualitative* and *quantitative* tests of the model, but there are some objections to these terms:

1. They are value-loaded, and it is preferable to use neutral language to assess model performance (see e.g. Oreskes *et al*. 1994).
2. The same terms are used in other branches of mathematics and logic to denote other meanings: a model is valid if the logic is correct, and verified if it is "true".
3. Verity implies truth, but it is impossible to prove a model "true" (except in the special case of a closed system). The only truth that can be established in a growth model is (e.g. in the context of Goulding 1979), that the model is a faithful representation of what the modeller intended. Similarly, the only sense of validity that can be demonstrated for a model is the "reasonableness" of the statistical assumptions.

Thus it is appropriate to avoid these terms, and to use alternatives such as *criticism* and *benchmarking* respectively. The former should involve examination of the structure and properties of a model, with or without supplementary data, to confirm that it has no internal inconsistencies and is biologically realistic. Benchmarking requires comparisons with data to quantify the performance of the model. Thorough benchmark tests should involve data which are in some sense unlike the data used to fit the model, but useful insights can also be obtained with the calibration data. These tests cannot prove a model to be "correct", but may

be used in attempts to *falsify* inferences made from the model. The quality of a model can only be evaluated in relative terms, and its predictive ability always remains open to question. However, the failure of several attempts to falsify a model should increase its credibility and build user confidence. This is the role of model evaluation. Thus model evaluation ia an on-going procedure which continues as long as the model remains in use.

Soares *et al*. (1995) and Vanclay (1994) recently reviewed some ways to evaluate forest growth models. Here, we give a brief overview of the framework they suggest, and offer some new insights into the role of benchmarking. We stress that model evaluation should not be a mere mechanical procedure to examine a model's technical credentials, but should also involve philosophical considerations by modellers and model users.

2. Procedures for Evalating Models

Model evaluation should reveal any errors and deficiencies in the model, in part, by establishing (Vanclay 1994):

1. whether the equations used adequately represent the ecophysiological processes involved,
2. if the equations have been combined correctly in the model,
3. whether the numerical constants obtained in fitting the model are the best estimates,
4. whether the model provides realistic predictions throughout the likely range of application,
5. if the model satisfies specified accuracy requirements, and
6. how sensitive model predictions are to errors in estimated coefficients and input variables.

An evaluation requires more than a decision regarding the acceptability of a model for a defined use. It should provide as much information as possible about the model's behavior and predictive ability, to allow users to decide if it is adequate for their intended uses. It should also reveal where future data collection and model revision efforts may be most useful.

Evaluation should not be a mere afterthought to model construction, but should be considered at every stage of model design and construction, when component functions are formulated and fitted to data, and when these components are assembled to provide the completed model. Here we deal primarily with regression techniques, but recognize that other approaches may also be used in modelling. Model evaluation includes both theoretical and empirical issues, and is dealt with in standard texts on applied regression analysis (e.g. Gilchrist, 1984; Ratkowsky, 1990). Key aspects may be grouped under five interrelated headings (with some selected examples):

1) Examine the model and its components in terms of logic structure and from theoretical and biological views (e.g. Hamilton, 1990; Oderwald and Hans, 1993) to see if they are:

- parsimonious
- biologically realistic
- consistent with existing theories of forest growth, and
- predict sensible responses to management actions.

2) Ascertain the statistical properties of the model in relation to data (e.g. Bates and Watts, 1988; Ratkowsky, 1983; Seber and Wild, 1989), including:
   - nature of the error term (i.e. additive or multiplicative, independence, etc.),
   - estimation properties of parameters in model functions.

3) Characterize errors (e.g. Gertner, 1987; Mowrer, 1991; Power, 1993; Reynolds, 1984; Reynolds and Chung, 1986) in terms of
   - accuracy,
   - nature of residuals (distribution, dependencies on initial stand conditions and length of projection),
   - confidence intervals and critical errors,
   - contributions by each model component to total error, and
   - propagation of errors through the model.

4) Test, using statistical approaches (e.g. D'Agostino and Stephens, 1986; Gregoire and Reynolds, 1988; Mayer and Butler, 1993; Power, 1993; Reynolds *et al*., 1988) for:
   - bias and precision of the model and its components,
   - goodness-of-fit of predicted size distributions,
   - patterns in, and distribution of residuals, and
   - correlations over time and between components.

5) Conduct sensitivity analyses to determine those model components which have the greatest influence on predictions (e.g. Botkin, 1993; Jørgensen, 1986; Van Henten and Van Straten 1994).

   These analyses need not be sequential, but all relevant aspects should be examined in each model component and in the assembled model. Each of these steps could involve both graphical analyses as well as statistical indices.

2.1 Logical and Biological Consistency

Each model component and the model as a whole should be logically consistent and biologically realistic. Many model properties can be examined for consistency, e.g. (after Oderwald and Hans 1993):

1. Do variables included in, and omitted from the model agree with expectations?
2. Do the sign and magnitude of coefficients should agree with expectations?
3. Are extrapolations outside the range of the development data reasonable?
4. Are transformations of model predictions reasonable (e.g. Do model forecasts of future diameters also provide reasonable estimates of diameter increments, future volumes, mean increment curves, etc.)?
5. Are any contradictions present within the model?
6. Do derivatives, limits, maxima, minima, inflections, etc. agree with expectations?

Care is required in resolving an apparent discrepancy between model predictions and expectation: it may be the expectations, and not that the model, that is wrong!

Matrix plots of simulated stand development trajectories showing a range of property-time and property-property relationships (Leary 1988) may offer useful insights into model behaviour, and may provide an efficient way to reveal discrepancies in model predictions.

Parameter estimates and model forecasts should agree with both empirical data and current understanding of growth processes. Experienced foresters and other experts may indicate areas where model predictions are deficient. Several researchers have advocated formalizing this procedure as a Turing test in which experts are asked to discriminate between simulated and real world data, but this does not provide a good basis for comparison. If the real and simulated data are sufficiently alike to offer a realistic test, they should be amenable to statistical testing which avoids potential difficulties with personal bias. Conversely, if the data are unsuited to statistical testing, it is likely that they will contain certain identifiable features which may make the distinction easy.

Simulations at extremes of stand condition may be particularly revealing. Optimization studies may provide a discriminating test of a model, as Monserud (1989) reported that his optimizer was remarkably efficient at exploiting seemingly minor quirks in the Prognosis model (Stage 1973, Wykoff et al 1982, Wykoff 1986) to arrive at unrealistic solutions. Thus optimization studies coupled with expert insights may provide a good basis for model criticism. However, a model should not be rejected simply because it behaves in a counter-intuitive fashion; it may be our preconceptions that are wrong. Thus discrepancies should cause a critical reappraisal of the model, the data, and of preconceptions.

2.2 Statistical properties

With linear regression models, $\boldsymbol{Y = Xb + e}$, it is usually assumed that the random errors e are additive, independent and identically normally distributed with zero mean and constant, but unknown variance ($e_i \sim N(0,s^2)$). Departures from these assumptions may result in parameter estimates with undesirable statistical properties. Several transformations and weighting techniques may be used where data do not satisfy these assumptions, but some problems may remain (e.g. multiplicative errors in models with additive terms that preclude logarithmic transformations).

In forestry applications, several measurements are often taken from each sampling unit (e.g. measurements on a single tree, trees on a plot, or remeasures of a plot). These repeated measurements are not statistically independent, and ordinary least squares techniques may underestimate the variance of parameters, leading to the acceptance of more complex models than would otherwise be indicated. West (1995) gave an overview of some ways to deal with this problem.

Parameter estimates of non-linear growth models may not possess the same desirable statistical properties as their linear counterparts (i.e. unbiased, normally-distributed, minimum variance estimators). However, non-linear models which are "close-to-linear" approach these properties asymptotically, and many models may be reparameterized so that they behave in a close-to-linear fashion (Ratkowsky, 1983, 1990).

In the standard regression model, the explanatory variables are assumed to be free of error. This assumption is rarely tenable in forest growth models, where there is joint variation in the variates, and this means that derived relationships could be grossly in error. It is possible to correct for this (e.g. Seber and Wild, 1989; Weisberg, 1985), but the procedures may be tedious. Failure to account for the nature of the response variable will lead to inflated estimates of variance, but the effect can be minimized by ensuring a large range of each explanatory variable relative to its error.

Most forest growth models are constructed from several equations independently fitted to data. Simultaneous estimation of all model components minimizes overall model errors and provides a variance-covariance matrix for the model as a whole (e.g. Gallant, 1987; Seber and Wild 1989), but few forest growth models have been constructed in this way (e.g. Furnival and Wilson 1971, García 1984, Leary 1970).

The standard regression assumptions are ideals that real situations (models and data in conjunction) may approach without ever exactly attaining. Fortunately, least-squares techniques tend to be relatively robust in practice (at least for parameter estimation, if not for assessing precision). Irrespective of this, evaluation of a model, before and after fitting to data, should include the appraisal of the statistical properties of the model and the data.

2.3 Characterizing model error

One of the most efficient ways to examine model performance is to plot residuals or standardized residuals for all possible combinations of tree and stand variables to detect possible autocorrelation and other dependencies. Such plots may be interpreted visually, but formal tests are also available (e.g. Draper and Smith, 1981; Weisberg, 1985).

Two simple criteria, in conjunction, provide a summary of the overall model performance: average model bias ($\Sigma(\bar{y}_i - y_i)/N$) and mean absolute difference ($\Sigma|\bar{y}_i - y_i|/N$). Average model bias measures the expected error when several observations are to be combined by totalling or averaging, and mean absolute difference measures the average error associated with a single prediction. Error dependencies on projection length or initial forest condition can be shown graphically. Regression analysis and principal component analysis may help to detect possible dependencies. These techniques apply equally when checking the model against data used for model calibration, and when testing the model with independent data.

The error structure and the contribution of each model component to total error may be more revealing than a mere evaluation of total model performance. Thus a map of variance components of the model may help to identify weaknesses and define priorities for future research.

2.4 Statistical tests

Many statistical tests of model performance have been suggested, but no single criterion can incorporate all aspects of model evaluation, and it is desirable to use several simple tests to examine different facets of model behavior.

One simple but efficient technique is based on linear regression of observed versus predicted data. Some useful insights into the quality of predictions may be given by R^2 and the slope and intercept of the fitted line, and a good test for bias is the simultaneous F-test for slope=1 and intercept=0 (e.g. Dent and Blackie, 1979; Mayer and Butler, 1993, Mayer *et al*. 1994).

Another useful technique is to compare predictions directly with observed data using a statistic analogous to R^2, and sometimes called modelling efficiency:

$$EF = 1 - \frac{\Sigma\left(y_i - \hat{y}_i\right)^2}{\Sigma\left(y_i - \bar{y}\right)^2}$$

This statistic provides a simple index of performance on a relative scale, where 1 indicates a "perfect" fit, 0 reveals that the model is no better than a simple average, and negative values indicate a poor model indeed.

In addition to overall appraisals, it is desirable to partition data (e.g., by age, site index or stand density), and examine model performance in each of several strata (e.g. Mayer and Butler, 1993). The most revealing insights may be obtained by devising strata based on a knowledge of the biological system, as well as model and data characteristics. However, the absence of inadequacies in any particular stratification does not imply that weaknesses will not be found in an alternative stratification.

2.5 Sensitivity Analyses

A sensitivity analysis attempts to reveal model inputs, parameters and submodels which, when perturbed, cause the greatest fluctuations in model predictions. These studies may reveal model components with low and high sensitivity, both of which are of interest. Insensitive components may contribute little toward model predictions and could be targets for omission from the model during model revisions. Conversely, it is useful to know about model components with high sensitivity, because these may have the greatest impact on model predictions. All model parameters and inputs should be estimated accurately, but particular care is required with the most sensitive variables.

In theory, the sensitivity of model parameters can be examned analytically (e.g. by taking derivatives), but in practice this may be complicated by the interaction of various model components and feedback loops. Thus sensitivity analyses are often carried out as simulation studies in which the parameters or components are changed to observe corresponding effect on predicted outputs. In practice, meaningful sensitivity studies are difficult, as the estimate of sensitivity depends both on the values of the inputs and the model parameters, so that many simulations may be necessary to complete the picture. This may be a tedious undertaking, especially where there are many parameters. Results of sensitivity tests may reveal parameters critical to model predictions, and parameters which may be redundant. Knowledge of sensitive parameters may guide applications (especially extrapolations) and the planning of model enhancements.

Similarly, it is important that users have a knowledge of the model's sensitivity to inputs. Studies of error propagation (Gertner 1987, Mowrer 1991) may reveal model limitations, and are particularly useful in offering insights into the interaction of errors in the input data and in the simulation. One application of stochastic simulation studies is to investigate the "quality" of predictions. Variance approximation provides an efficient alternative to such studies, and enables the variance of predictions to be estimated deterministically. It also enables the variance of the input data to be incorporated into the analysis. Mowrer and Frayer (1986) and Gertner (1987) used a simple first-order Taylor series to estimate the errors propagated through growth and yield projections.

3. Benchmark Tests

It has become customary in the evaluation of growth models to reserve some data to provide an "independent" benchmark test of the model (e.g. Snee 1977, West 1981, Shifley 1987). This raises some controversial questions about the merits of setting data aside for "independent" tests, about the nature and amount of data used for such comparisons, and about the nature of the population of interest. In effect, benchmarking involves a compromise between the best possible parameter estimates (using all the data for calibration) and the best possible estimates of precision (reserving some data for benchmarking). Two options seem to offer the best of both worlds: 1) to reserve data, benchmark and then to recalibrate using the full data set; and 2) to use re-sampling techniques such as cross-validation.

3.1 Partitioning Data

Benchmarking in its purest form requires independent data against which to test the model. The most convincing test would use data from controlled and replicated trials measured over a long period, but such data are rarely available. Growth modellers have to decide whether it is worthwhile splitting data into two subsets, one for development, and the other for the testing the model. This is not a trivial decision, especially when data are scarce. Setting some data aside may help to test the model, but may result in inferior parameter estimates.

The role of an independent benchmark sample cannot be divorced from the nature of the model. If the model fitting exercise is intended to reveal possible causal parameters (e.g. in medical epidemiology), then the costs of independent benchmarking may be greater than the benefits (Hirsch 1991). Partitioning data to allow benchmarking may help to reduce type I errors (i.e. falsely rejecting the null hypothesis, and thus e.g. incorrectly concluding that a variable contributes little and should be omitted from the model), but fewer data for calibration mean a reduction in the precision of parameter estimates, and an increase in type II errors (i.e. falsely accepting the null hypothesis and thus e.g. including irrelevant variables in the model). However, if empirical data are used to calibrate a model deliberately formulated to represent biological processes, then the goal is a different one: namely to accurately estimate parameters rather than to identify possible explanatory variables. In this latter case, benchmark data may serve a more useful role in illustrating the robustness of the model. Clearly, an assessment of the utility of independent benchmark data cannot be divorced from the purpose of the model.

If a decision is made to partition a data set, the modeller must avoid the temptation to weaken the tests, for example, by reducing the number of data available for benchmarking, despite a desire to find the model acceptable. The outcome of benchmark tests can be influenced by the selection of data: "like" data will provide a more optimistic result than comparisons with "unlike" data from another population. Thus the most convincing demonstration of model quality can be made only if the test data are in some sense unlike the development data. A single sample split into two parts is no substitute for test data from controlled, replicated trials. Vanclay (1994, p.88) discussed the dangers of constructing a growth model from passive monitoring data in which stand density and site productivity were confounded. Splitting such data into calibration and benchmark sets would not reveal the fallacy of a positive correlation between stand density and tree growth; this can only be refuted (empirically) using data from thinning and spacing trials.

Unfortunately, the ideal, a series of properly replicated trials, is rarely available. However, data which are spatially (e.g. different location), temporally (e.g. more recent), or logistically (e.g. collected by a different agency) independent may provide a convincing test if they can be reserved without compromising the range of site and stand conditions represented in the model. Plots established for long periods with regular remeasurement, particularly those remaining undisturbed (i.e. no thinning), may prove useful as a discriminating test. Objective procedures (e.g. Snee 1977) may be used to select these data to minimize the dangers of bias. Following testing, the benchmark and calibration data should be pooled and the model recalibrated to obtain the best parameter estimates.

One possible frustration with benchmarking may arise when the initial calibration of the model seems inadequate in benchmark trials, since there is no way to test if recalibration using the pooled data will result in a significant improvement. The change in parameter estimates may serve as a guide (and may even serve as a good benchmark criterion), but do not reveal if the recalibration is "adequate". However, if the model is the best that can be obtained with existing resources, it must be considered acceptable, even if inadequate in some sense, since there is no alternative other than to invest more resources and wait for new data and techniques. Perhaps the real test of a model is if forest managers have sufficient confidence in it to use it as the basis for management decisions.

3.2 Resampling Procedures

An efficient alternative to independent benchmark data is to mimic these tests with resampling techniques such as cross-validation, boot-strapping and jack-knifing (e.g. Efron and Gong 1983, Weisberg 1985). Cross-validation is the logical generalization of partitioning the data for model calibration and benchmarking. Rather than omitting some data, each datum is deleted in turn and the model is fitted to the remaining *n-1* data. Benchmark tests are averaged from the individual deleted data. If the test statistic is squared error and the model is linear, the cross-validation estimate of true error is *n* times the PRESS statistic computed by many regression packages. The boot-strap and jack-knife are similar, especially as sample size increases, but are computationally more complex.

One shortcoming of any resampling procedure lies in its dependence on the data. The sample should adequately represent the variability and other characteristics of the population of interest, or the resampling procedure will not provide an adequate test of the model. Unfortunately, these are the very circumstances under which the model itself should come under heaviest criticism.

Despite the efficiency of re-sampling procedures, it seems impossible to avoid the use of some benchmark data, since resampling to test a complete model involving many relationships and assumptions seems impractical.

4. Other considerations

A technical appraisal of a model does not constitute a complete evaluation. There are several other important qualitative aspects which should also be considered. Some of these aspects include:

1. Does the model satisfy the needs of clients?
2. Are the underlying concepts sound, and visible to users (in the model or documentation)?
3. Have they been implemented faithfully, or constrained by resources or technology? (e.g. Has the IF ... THEN ... ELSE ... ENDIF structure of the computer language led to the use of on–off behaviour rather than a gradual phasing in and out, even though the latter may be more appropriate?).
4. Is the model parsimonious, satisfying the general principle of science (Ockham's razor) that *entities should not be multiplied beyond necessity*?

Some of these aspects have been explored more thoroughly in the social sciences where it is more difficult to obtain quantitative benchmark data than in the nautral sciences. Thus it is interesting to explore some experiences of that discipline. In a review of several models for social policy analysis, Meadows and Robinson (1985, p.370) observed that "tests tend to be weak, marginal, unsymmetrical and very biased. In part this is due to oversized models whose complete testing would be impossibly expensive and tedious. It is also due to a general lack of imagination, motivation, training, client pressure and agreed-upon methods for testing." Although this criticism was levelled specifically at the social sciences, it also applies to some extent, in forest growth modelling. Meadows and Robinson (1985, p.392-402) collated specific advice to overcome these limitations, including (and followed by our responses):

1. Modellers should think more and wield tools less (Majone 1977) – the "new tool" syndrome is a hazard that is also prevalent in growth modelling;
2. Models should be given to an independent evaluation agency for testing (Quade and Boucher 1968 p.352) – several independent evaluations have been published in refereed journals (e.g. Reynolds 1984, Oderwald and Hans 1993, Soares *et al.* 1995);
3. Modellers should test *each* part of their model, not just the summary output (Biggs and Cawthorns 1962) – this may be tedious and time-consuming, but is important to gain a good insight into the model (e.g. Hann 1980);
4. Modellers should test their results against the real world, rather than against a set of artificial rules or formulas (Brewer 1973) – this seems to be one thing that is done well on the rare occasions that forest growth models are thoroughly benchmarked.
5.

It is disconcerting to reflect how this advice remains as necessary, and as rarely applied today, as it was some 20 years ago when first offered.

Meadows and Robinson (1985, p.407) concluded with a warning that "modelling efforts often succumb to a slow ... drift ... away from what is important to what is ... tractable, away from unconventional viewpoints and toward established wisdom. At each little decision point ... the guiding question should be 'would it help solve the *problem*?'. ...[T]he criterion for decision should always be what will most help real-world decisions, not what the modeller will find easy or fun, or what the client will find ... uncontroversial." A decade later, this warning remains timely and pertinent.

Some readers may find our stance too idealistic, but while we accept that modelling may be constrained by knowledge, data and resources, we echo the sentiments of Ziman (1978): "[one] learns how easy it is to persuade oneself of the validity of a model which later turns out to be false, and comes to realize that even in very strongly mathematical and well-defined scientific issues, it may take a long time, much criticism and the death of many promising conjectures before a reliable theory is [established]".

5. Synthesis of evaluation procedures

These few simple suggestions are not intended as a comprehensive review of model evaluation procedures, but merely highlight some important and sometimes overlooked aspects. We stress that evaluation is not one simple procedure, but consists of a number of interrelated steps that cannot be separated from each other or from model construction. Several statistical tests, as well as graphical procedures, may be useful, both with data used for model calibration and with data used for "independent" evaluation of the model. However, the validity of conclusions depends on the validity of assumptions. These principles should be kept in mind throughout model construction and evaluation.

Acknowledgments

Jerry Leech and Stanley Wood provided helpful suggestions and thought-provoking comments on the draft manuscript.

References

Bates, D.M. and Watts, D.G., 1988: Nonlinear regression analysis and its applications. Wiley, N.Y., xiv+365 p.

Biggs, A.G. and Cawthorns, A.R., 1962, quoted in P.W. House and J. McLeod, Large-Scale Models for Policy Evaluation. Wiley, NY, p.73.

Botkin, D.B., 1993. Forest Dynamics: an ecological model. Oxford Univ. Press, xv+309 p.

Brewer, G.D., 1973. Politicians, Bureaucrats and the Consultant. Basic Books, NY.

D'Agostino, R.B. and Stephens, M.A. (eds), 1986. Goodness-of-fit Techniques. Marcel Dekker, N.Y., xviii+560 p.

Dent, J.B. and Blackie, M.J., 1979. Systems Simulation in Agriculture. Applied Science Publishers, London.

Draper, N.R. and Smith, H., 1981. Applied Regression Analysis. Wiley, N.Y., 709 pp.

Furnival, G.M. and Wilson, R.W., 1971. Systems of equations for predicting forest growth and yield. In G.P. Patil, E.C. Pielou and W.E. Walters (eds) Statistical Ecology. Penn. State Univ. Press. Volume 3, p. 43-57.

Gallant, R.A., 1987. Nonlinear Statistical Models. Wiley, N.Y., xii+610 pp.

García, O., 1984. New class of growth models for even-aged stands: Pinus radiata in Golden Downs Forest. N.Z. J. For. Sci. 14:65-88.

Gertner, G., 1987. Approximating precision in simulation projections: an efficient alternative to Monte Carlo methods. For. Sci. 33:230-239.

Gilchrist, W., 1984. Statistical Modelling. Wiley, Chichester, xv+339 pp.

Goulding, C.J., 1979. Validation of growth models used in forest management. N.Z. J. For. 24:108-124.

Gregoire, T. and Reynolds, M.R., 1988. Accuracy testing and estimation alternatives. For. Sci. 34: 302-320.
Hamilton, D.A., 1990. Extending the range of applicability of an individual tree model. Can. J. For. Res. 20:1212-1218.
Hann, D.W., 1980. Development and evaluation of an even- and uneven-aged ponderosa pine/Arizone fescue stand simulator. USDA For. Serv., Res. Pap. INT-267. 95 p.
Hirsch, R.P., 1991. Validation samples. Biometrics 47:1193-1194.
Jørgensen, S.E., 1986. Fundamentals of Ecological Modelling. Elsevier Amsterdam, 389 pp.
Leary, R.A., 1970. Systems identification principles in studies of forest dynamics. USDA For. Serv., Res. Pap. NC-45, 38 pp.
Leary, R.A., 1988. Some factors that will affect the next generation of forest growth models. In A.R. Ek, S.R. Shifley and T.E. Burk (eds) Forest Growth Modeling and Prediction. Proceedings of IUFRO Conference, 24-28 Aug 1987, Minneapolis, MN. USDA For. Serv., Gen. Tech. Rep. NC-120, pp. 22-32.
Majone, G., 1977. Pitfalls of analysis and analysis of pitfalls. Urban Analysis 4:235.
Mayer, D.G. and Butler, D.G., 1993. Statistical validation. Ecological Modelling 68:21-32.
Mayer, D.G., Stuart, M.A. and Swain, A.J., 1994. Regression of real world data on model output: an appropriate overall test of validity. Agricultural Systems 45:93-104.
Meadows, D.H. and Robinson, J.M., 1985. The Electronic Oracle: computer models and social decisions. Wiley, Chichester, xv+445 pp.
Mowrer, H.T., 1991. Estimating components of propagated variance in growth simulation model projections. Can. J. For. Res. 21:379-386.
Oderwald, R.G. and Hans, R.P., 1993. Corroborating models with model properties. For. Ecol. Manage. 62:271-283.
Power, M., 1993. The predictive validation of ecological and environmental models. Ecological Modelling 68:33-50.
Quade, E.S., and Boucher, W.I. (eds), 1968. Systems Analysis and Policy Planning. Elsevier, NY.
Ratkowsky, D.A., 1983. Nonlinear Regression Modeling. Marcel Dekker, NY, viii+276 pp.
Ratkowsky, D.A., 1990. Handbook of Nonlinear Regression Models. Marcel Dekker, NY, ix+241 pp.
Reynolds, M.R., 1984. Estimating the error in model predictions. For. Sci. 30: 454-469.
Reynolds, M.R. and Chung, J., 1986. Regression methodology for estimating model prediction error. Can. J. For. Res. 16:931-938.
Reynolds, M.R., Burk, T.E. and Huang, W., 1988. Goodness-of-fit tests and model selection procedures for diameter distribution models. For. Sci. 34: 373-399.
Seber, G.A.F. and Wild, C.J., 1989. Nonlinear regression. Wiley, N.Y., xx+768 p.
Shifley, S.R., 1987. A generalized system of models forecasting Central States growth. USDA For. Serv., Res. Pap. NC-279. 10 p.
Snee, R.D., 1977. Validation of regression models: methods and examples. Technometrics 19:415-428.
Stage, A.R., 1973. Prognosis model for stand development. USDA For. Serv., Res. Pap. INT-137. 32 p.
Van Henten, E.J. and Van Straten, G., 1991. Sensitivity analysis of a dynamic growth model of lettuce. J.Agric. Engineering Res. 59:19-31.
Weisberg, S., 1985. Applied Linear Regression, 2nd ed. Wiley, NY, xiv+324 pp.

West, P.W., 1981. Simulation of diameter growth and mortality in regrowth eucalypt forest of southern Tasmania. For. Sci. 27:603-616.
West, P.W., 1995. Application of regression analysis to inventory data with measurements on successive occasions. For. Ecol. Manage. 71:227-234.
Wykoff, W.R., 1986. Supplement to the user's guide for the stand prognosis model - version 5.0. USDA For. Serv., Gen. Tech. Rep. INT-208. 36 p.
Wykoff, W.R., Crookston, N.L. and Stage, A.R., 1982. User's guide to the stand prognosis model. USDA For. Serv., Gen. Tech. Rep. INT-133. 112 p.
Ziman, J., 1978. Reliable Knowledge: an exploration of the grounds for belief in science. Cambridge Univ. Press, 197 pp.