Ecological Modelling, 68 (1993) 21–32
Elsevier Science Publishers B.V., Amsterdam

Statistical validation

D.G. Mayer [a] and D.G. Butler [b]

[a] *Queensland Department of Primary Industries, G.P.O. Box 46, Brisbane, Australia*
[b] *Queensland Department of Primary Industries, P.O. Box 102, Toowoomba, Australia*

(Received 22 April 1992; accepted 25 November 1992)

ABSTRACT

Mayer, D.G. and Butler, D.G., 1993. Statistical validation. *Ecol. Modelling*, 68: 21–32.

Validation is a necessary step for model acceptance. No single combination of validation tests will be applicable across the diverse range of models and their uses. Choice of technique is important, as some contain problems and inconsistencies. Subjective assessment can be useful as a guide. Within visual techniques, observed vs. predicted plots are shown to have superior diagnostic capabilities compared to the more widely-used time-series plots. Mean absolute error is demonstrated as a more robust deviance measure than mean absolute percent error, and within the statistical tests a nominated sub-set of some simpler statistics should reveal most of the required information. The modelling efficiency is proposed as the best overall measure of agreement between observed and simulated values.

INTRODUCTION

Validation has been defined as a comparison of the model's predictions with the real world to determine whether the model is suitable for its intended purpose (McKinion and Baker, 1982). Model validation is a mandatory step in the complex task of simulation. General overviews of this process can be found in most simulation texts (Jørgensen, 1986; Bratley et al., 1987; Ripley, 1987), with more comprehensive views somewhat rarer (Dent and Blackie, 1979). Note that there are no absolute criteria; validation relates to the potential applications and users of the model, not the model itself (McCarl, 1984).

A range of potential errors and problems exist in the overall process of validation, as discussed elsewhere. These include the necessity for on-going

Correspondence to: D.G. Mayer, Queensland Department of Primary Industries, G.P.O. Box 46, Brisbane, Qld. 4001, Australia.

validation as opposed to a one-off exercise (Law and Kelton, 1982); acceptable levels and costs of Type I and Type II statistical errors (Dent and Blackie, 1979); and sources of acceptable or suitable data sets for validation (McCarl, 1984). Also, models can never be proven valid, only invalid (Harrison, 1990). Failure to prove a significant difference between real and model data may only be due to insufficient replication or lack of power of the applied statistical test.

This study assumes that the necessary data have been obtained, and concentrates on techniques and statistics for comparing validation data with model predictions. It is shown how some commonly-used methods are open to misinterpretation and potential abuse, and better alternatives are evaluated. The modelling efficiency, a dimensionless statistic which parallels the co-efficient of determination, is proposed as the best overall measure of model performance.

VALIDATION TECHNIQUES

A wide range of methods has been proposed and used in many different fields of study. In many cases the choice of technique is restricted by the potential uses and testing requirements of the model, the type of data that the model generates, or the availability of real-world data. Validation techniques can be grouped into four main categories, namely subjective assessment, visual techniques, deviance measures and statistical tests.

Subjective assessment

These techniques involve evaluation by a number of experts in the field of interest. They include the Turing-type tests (Law and Kelton, 1982), where the experts are presented with both simulated and real-world data series and asked to distinguish between these. This application in particular is open to misinterpretation. Whilst the model may perform well, there may be certain identifiable features contained in either the simulated or real-world data which make the distinction easy.

Due to their very nature, subjective tests are prone to personal bias. Some members of a panel may pass less-than-acceptable results; equally, others may be over-critical. Successes in subjective assessment would appear to be more a function of panel selection than model performance, and this technique's major advantage would appear to be as a complement to more objective measures.

Visual techniques

Graphical displays of data feature in these methods, typically plots of both simulated data (usually continuous, and represented by a line) and

observed data (usually discrete, and represented by points) against a common independent variable. These are most commonly presented as time series plots, but may also be applied across, for example, depth of water or soil profile, altitude, pathogen or pest numbers, or concentration of a critical nutrient or toxic element.

These plots have been recommended as an informative method of data presentation (Dent and Blackie, 1979), and are widely used in many disciplines. There are, however, problems with these displays, despite their popularity. Take, for example, Fig. 1a, where the data appear to be at least approximately in agreement with the line. This is despite the fact that they are independent — the line was calculated as $y = \sin(x^2) - 0.07x^2 + 0.7x + 0.2$, and the points (equally spaced across the X-axis) were chosen from the uniform distribution $0.5 \leqslant x \leqslant 2.5$, using the first 8 pairs of random numbers from Kendall and Babington-Smith (1954). Figure 1b looks to be a reasonably good fit; the points here were selected for statistical properties that will be illustrated later.

The apparent acceptance of variable-quality data presented in this way is enhanced by the undiscriminating reader, who views the deviance of each point as the distance to the nearest line (in any direction). The true deviance is the perpendicular distance to the line, which is often far larger, especially if the simulated line oscillates across the available space.

A preferable alternative is to plot the observed (y) vs. predicted ($\hat{y}$) data directly, with the line $y = \hat{y}$ marked (to indicate the position of the 'perfect fit'). The fitted linear regression line should not be presented, as this process equates to a re-calibration of the model, and successfully hides any systematic departure (i.e. bias) from the line $y = \hat{y}$. If the data can somehow be stratified (e.g. according to location, soil type, species, age, sex,

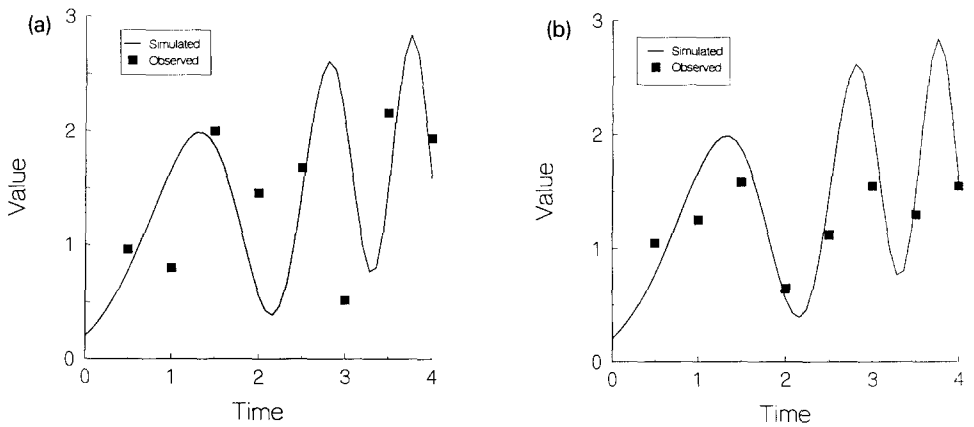

Fig. 1. Hypothetical data (see text) vs. time.

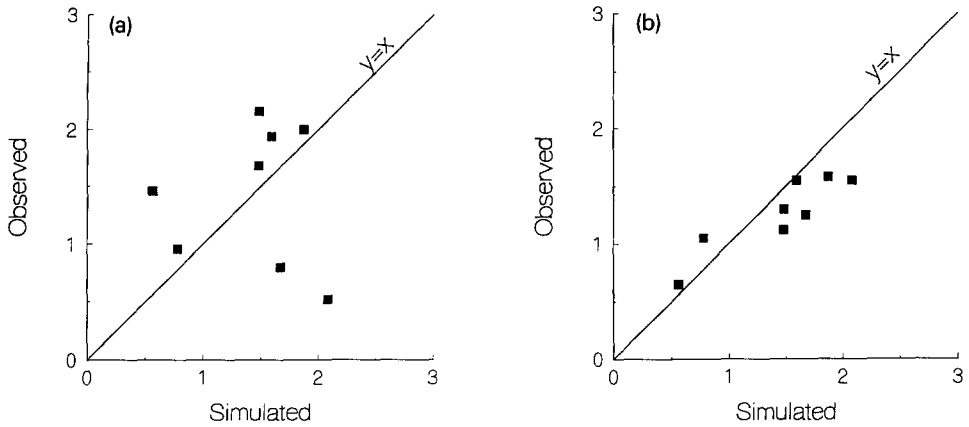

Fig. 2. Observed vs. predicted plots of data from Fig. 1.

etc.), it is advantageous to plot each stratum with a different symbol. This plot directly presents goodness of fit as vertical deviations from the 'perfect' line, and indicates any biases present (either overall, or in certain sections of the data). This type of graph presents very similar information as a residuals $(y - \hat{y})$ vs. predicted $(\hat{y})$ plot, widely used in statistical diagnostics (Draper and Smith, 1966). For validation purposes, y itself is preferable to the residual, as this way the predicted and observed values can be read off directly.

Figure 2 shows the data from Fig. 1 plotted this way. As expected, Fig. 2a shows the 'model' to be worthless, with a random scatter of data. Figure 2b demonstrates a strong positive relationship, with the first indication of two problems — firstly, insufficient data sampling, as there are no data in the higher range of predicted values; and secondly, model bias, with evidence of under-prediction at higher values.

Deviance measures

These are applicable when observed and simulated data can be paired according to time, location, treatment, etc. Deviance measures are based on the differences between the simulated and observed values. A balanced average of all data is normally used, but in some circumstances data weighting may be applied, for example, 'worst case' measures based on the maximum or extreme values (Jørgensen, 1986). Miller (1974) and Miller et al. (1976) propose a method for internal model validation based on an extended sensitivity analysis combining deviations, but this does not appear to have been much used. For binary data, a simple deviance measure is percent of the observational units with a correct value (Brammer, 1989).

For numerical data, two commonly used measures are mean absolute error (MAE) and mean absolute percent error (MA%E) (Schaeffer, 1980), defined as:

$$\text{MAE} = \left(\Sigma \mid y_i - \hat{y}_i \mid\right)/n, \quad \text{and} \tag{1}$$

$$\text{MA\%E} = 100\left[\Sigma\left(\mid y_i - \hat{y}_i \mid / \mid y_i \mid\right)\right]/n, \tag{2}$$

where y_i represent observed values, $\hat{y}_i$ simulated values, and n the number of pairs.

MA%E is widely used, although by different names, in demographic studies (Siegal, 1972; Cohen, 1986; Smith, 1987) and other fields (Hameed, 1974). As MAE is in the same units as the data and MA%E is relative, both can be informative measures. Note that MAE may also be used for mean *algebraic* error (Smith, 1987), in which case the absolute signs would be omitted from Eq. (1). In this case, it is not a true measure of deviance, as positive and negative deviations cancel each other out. Rather, its overall average measures any bias of the model by comparing the distribution of the differences against zero. As bias is better measured by other statistical tests, we do not recommend this usage, and adopt MAE as its original (absolute) definition.

An alternative to using absolute differences is to use second moments (Picard and Cook, 1984), and, using its square root, derive the root mean square error RMSE as

$$\text{RMSE} = \left\{\left[\Sigma(y_i - \hat{y}_i)^2\right]/n\right\}^{0.5}. \tag{3}$$

Algebraically, RMSE > MAE (due to the influence of squaring larger values), with these measures being approximately equal if the absolute differences are of similar magnitude. With squared deviations, RMSE can be useful in deriving statistical properties. As a summary measure of the relative degree of deviations, either MAE or RMSE can be used.

Kleijnen (1987) recommends MA%E, and suggests 10% as an upper limit on acceptability. Given that model validity depends very much on both the type of model and its intended uses (Bratley et al., 1987), it is impractical to set a single absolute limit. In many cases these measures are used merely to compare different models or techniques.

A potential problem exists in Eq. (2) with the division by y_i. Obviously, MA%E is undefined if any observed value (y_i) equals zero, but this restriction does not apply to the simulated values ($\hat{y}_i$). Problems also occur with low values of y_i, as MA%E tends towards infinity as any y_i tends towards zero. Given that MA%E can be heavily influenced by single low values of y_i, it should be viewed with caution unless all y_i are of similar

magnitude. Given this limitation, a better relative deviance measure may be the mean absolute error relative to the observed mean ($= MAE/\bar{y}$), which can also be used if MA%E is undefined. An alternate relative measure is the 'general standard deviation' of Jørgensen et al. (1986), which equates to ($RMSE/\bar{y}$). This measure would provide very similar information to ($MAE/\bar{y}$).

Statistical tests

Applicability of statistical tests depends very much on the types of data available. For example, if sample population distributions are measured and simulated, then overall population tests are applicable, such as the unpaired *t*-test or a non-parametric distribution test (Conover, 1980). If, however, paired samples are available, appropriate methods include the paired *t*-test, regression analysis or the non-parametric sign test. Stochastic models, which generate a distribution for comparison with each observation, also introduce statistical problems (Dent and Blackie, 1979). The usual approach here is to use the mean of the generated distribution, although more complex alternatives are available (Reynolds et al., 1981; Whitmore, 1991).

A wide range of potential statistical tests exists, as outlined by McCarl (1984). Many are specialised applications to particular disciplines and model types. Of the 'general' statistics available, we favour the simpler, more easily understood ones. We also prefer parametric to non-parametric statistics, provided the underlying parametric assumptions hold with either the data or their transformation, because of the former's power in validation applications (Reynolds et al., 1981). Note that it is usual to adopt two-tail significance testing, thus allowing for deviations in either the positive or negative direction.

As a first step, the spread of the distributions can be tested with the *F*-statistic for variance ratios, and the locations of the distributions can be checked by calculating their respective means and the appropriate *t*-test (Kleijnen, 1987). The overall distribution tests (Conover, 1980) simultaneously test both these properties, but we recommend they be detailed individually for clarity. A range of more complex parametric statistics, based on the ones above, is outlined in Reynolds et al. (1981). They demonstrate that these are useful in identifying data problems which would otherwise escape detection. We do not support their view, and will show later that the simpler statistics perform equally well on their example.

One potential problem with these techniques is the non-independence of data. If the observations are taken from different experimental units these statistical tests are appropriate. If, however, the observations are repeated samplings of the same experimental unit (as often occurs in time-series

data with insufficient time-lag), they will be correlated, violating the assumption of independence. In those cases, two possibilities are available. Firstly, Feldman et al. (1984) define a chi-square statistic for testing between samples, allowing for dependence. To apply this statistic, it is necessary to estimate the population variances and covariances, and Feldman et al. (1984) concede that the required data will rarely be available. Instead, they advocate using estimates of these from a stochastic model. Whilst theoretically correct, this technique may be difficult for generalist modellers to implement.

The second available alternative for dependent time-series data relies on a number of time series (for example, different years, locations, varieties, treatments, etc.) being available. Split-plot analysis of variance can be applied (Cole and Grizzle, 1966), with time being the split factor. The main-plot analysis will be valid, but may have too few error degrees of freedom to be sufficiently powerful. Within the split-plot section, the 'time' main effect and 'main plot by time' interaction effects will be approximate only, but will at least indicate which effects are dominant.

Given independent paired data, regression analysis of observed vs. predicted data is also a useful tool for model validation. To conform with statistical assumptions, it is usual to take the observations as the Y-variate (as these data contain natural variability), and model predictions as the non-variable Xs (deterministic models contain no variation, and stochastic models can be re-run many times to minimise variation) (Harrison, 1990). A number of useful statistics are available from regression analysis — R^2 indicates the degree of fit, significance of the quadratic term can be used as a test for curvature, and the fitted constants indicate any observed biases in the model. The simultaneous F-test for slope = 1 and intercept = 0 (Dent and Blackie, 1979) is particularly useful in identifying bias.

As indicated in the Visual Techniques section, the problem with the line of best fit is that it does not relate the observed data to the 'perfect fit' line; rather to a 're-calibration' of the model. A data set could be extremely close to linear, but spatially removed from the $y = \hat{y}$ line. The regression fit would be excellent, although the simultaneous F-statistic for bias would be significant.

A dimensionless statistic which *directly* relates model predictions to observed data is the modelling efficiency, EF (Loague and Green, 1991), defined as:

$$\text{EF} = 1 - (\text{SS about } y = \hat{y})/(\text{Corrected SS of y})$$

$$= 1 - \Sigma(y_i - \hat{y}_i)^2/\Sigma(y_i - \bar{y})^2 \qquad (4)$$

where SS is sum of squares. Along similar lines, Greenwood et al. (1985)

proposed using (SS about $y = \hat{y}$) as a percentage of the corrected SS of y, as a measure of model performance. This is effectively 100(1 − EF). From Eq. (4), it is obvious how EF parallels the widely-used co-efficient of determination,

$$R^2 = 1 - (\text{SS about line of best fit})/(\text{Corrected SS of } y) \quad (5)$$

R^2 is interpreted as the proportion of variation explained by the fitted regression line, and EF is a similar measure against the set line $y = \hat{y}$. For both statistics, a 'perfect fit' results in zero for the sum of squares about the respective lines, giving unity as the upper bound and desired value. The degree of fit declines as these statistics fall away from one. For regression, the line of best fit cannot be worse than $y = \bar{y}$, so R^2 has a lower bound of zero. In Eq. (4) the data are compared with a fixed line, so this restriction is removed. Thus, EF has a (theoretical) lower bound of negative infinity. A value of zero indicates the fit to $y = \hat{y}$ is equal to the fit to $y = \bar{y}$, with values of EF less than zero resulting from a worse fit. Use of this statistic in validation is an extension of the R^2 statistic for the class of non-linear regression models which do not include $y = \bar{y}$ as a possible case. In these situations, negative values are also possible.

The calculated EF is thus an overall indication of goodness of fit. Any model giving a negative value cannot be recommended, with preferable values close to one indicating a 'near-perfect' model.

APPLICATION TO DATA SETS

Taking the data presented in Fig. 1, a number of relevant statistics were calculated, as presented in Table 1. Data set *a* was previously dismissed on visual appraisal, and the statistics confirm this — high deviance measures, no relationship between observed and predicted, and EF value less than zero. On the time-series plot, data set *b* appeared a good fit, and some statistics confirm this — comparatively low deviance measures, a non-sig-

TABLE 1

Statistical measures of validation applied to data from Fig. 1

Data set	Deviance measures		Paired *t*-test	Linear regression				Modelling efficiency
	MAE	MA%E		R^2	Slope	Intercept	Bias [a]	
a	0.61	70.0	0.0 [ns]	0.001 [ns]	−0.04	1.50	1.8 [ns]	−0.78
b	0.27	21.8	1.9 [ns]	0.805 **	0.55	0.46	10.6 *	−0.12

[a] Simultaneous *F*-statistic for slope = 1 and intercept = 0.
[ns] not significant; * $P < 0.05$; ** $P < 0.01$.

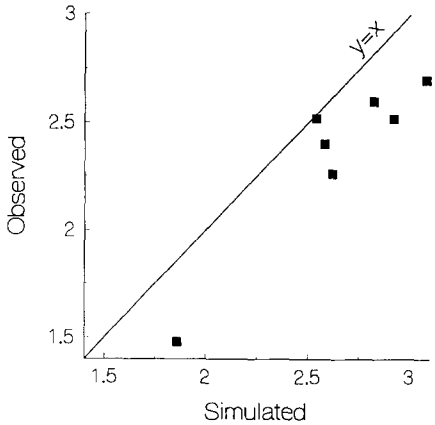

Fig. 3. Nitrate levels (ppm) in the Pigeon River, from Summers et al. (1991).

nificant t-test for differences, and a regression R^2 of 80%. The observed vs. predicted plot, however, indicated potential bias, which was confirmed by the significance ($P < 0.05$) of the simultaneous F-test. However, serious problems are indicated by the negative EF value, showing that $y = \bar{y}$ is a closer fit to these data. For this hypothetical model, we would conclude that whilst a good relationship exists between observed and predicted values, the model displays bias, and re-calibration would be necessary.

A second example demonstrates how both visual presentation and statistical tests can be open to mis-interpretation. Plotted against geographical distance, the NO_3 validation data of Summers et al. (1991), their fig. 10c, looks a reasonably good fit. However, an observed vs. predicted plot as shown in our Fig. 3 (ignoring the initial 'pre-discharge' calibrated point) reveals consistent over-prediction. The statistics applied by Summers et al. (1991), namely individual t-tests of slope against one and intercept against zero, were non-significant [for data in Fig. 3, $b = 0.98 \pm 0.16$ (standard error), $a = -0.22 \pm 0.43$]. However, significant bias is indicated by alternate statistics, namely a simultaneous test of the above conditions ($F_{2,5} = 11.1$; $P = 0.017$), and by a t-test for paired observations ($t_6 = 5.1$; $P = 0.004$). Also, the modelling efficiency is reasonably low at 0.35.

Our third example uses the data from 63 pine plantation plots of Daniels and Burkhard, as listed in Reynolds et al. (1981). These authors used comparatively complex statistical techniques to find previously undetected deficiencies in the model's predictions. For our analysis, we initially view an observed vs. predicted plot, with age classes (as grouped by Reynolds et al., 1981) being plotted as different symbols (Fig. 4). To the discerning eye, it is apparent that the model is over-predicting at younger ages and under-predicting at older ages.

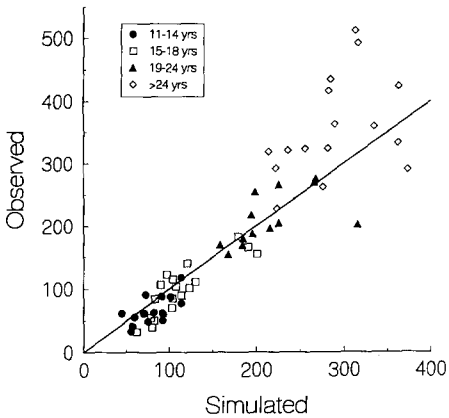

Fig. 4. Pine plantation yields (m^3/ha) from Reynolds et al. (1981), by tree age.

TABLE 2

Validation measures for data of Reynolds et al. (1981), both overall and by age class

Age class (years)	Combined	11–14	15–18	19–24	> 24
Number of observations	63	16	18	13	16
Mean (simulated values)	171	80	114	215	289
Mean (observed values)	180	67	101	211	356
Paired t-test	−1.3 ^{ns}	3.1 **	2.6 *	0.4 ^{ns}	−3.5 **
Linear regression					
R^2	0.83	0.41	0.75	0.32	0.16
Slope	1.24	0.66	0.93	0.53	0.62
Intercept	−31.6	13.4	−5.6	96.0	176.7
Bias [a]	6.3 *	5.6 *	3.1 ^{ns}	1.8 ^{ns}	5.8 *
Modelling efficiency	0.80	−0.12	0.64	0.07	−0.64

[a] Simultaneous F-statistic for slope = 1 and intercept = 0.
^{ns} not significant; * $P < 0.05$; ** $P < 0.01$.

This opinion is confirmed by the statistics presented in Table 2—overall, the t-test is non-significant, and both EF and R^2 are quite good. The F-test for bias is, however, significant ($P < 0.01$), and the slope greater than one. Considering the individual ages, the t-tests show significant differences in three out of the four classes, and EF appreciably better than zero in only one case. Taken overall, these results indicate a reasonable model that still contains unacceptable biases, as was also concluded by Reynolds et al. (1981).

CONCLUSIONS

Due to the complexities of models and data types, there is no set combination of validation techniques which is applicable across all mod-

elling situations. In most cases, a number of validation measures are necessary to appreciate 'the whole picture'. Of the available methods, the simpler ones in many cases are both adequate and preferable.

Whilst time-series plots can be informative, a better diagnostic alternative is to plot the observed vs. predicted data, with the line $y = \hat{y}$ marked. For deviance measures, mean absolute error or root mean square error are recommended as more stable statistics than mean absolute percentage error. The parametric t-test of means, and linear regression analysis of the observed vs. predicted plot (including simultaneous F-test for bias) are the most useful general statistical methods. The modelling efficiency, a statistic based on the co-efficient of determination, directly compares predictions with real-world observations, and is proposed as an important overall measure of fit.

ACKNOWLEDGEMENTS

We are grateful to Don Abel and Mark Silburn for raising the initial query. Thanks also to Tony Swain and Pat Pepper for their statistical advice.

REFERENCES

Brammer, R.F., 1989. Unified image computing based on fractals and chaos model techniques. Opt. Eng., 28: 726–734.

Bratley, P., Fox, B.L. and Schrage, L.E., 1987. A Guide to Simulation. Springer-Verlag, New York, NY.

Cohen, J.E., 1986. Population forecasts and confidence intervals for Sweden: A comparison of model-based and empirical approaches. Demography, 23: 105–126.

Cole, J.W.L. and Grizzle, J.E., 1966. Applications of multivariate analysis of variance to repeated measurements experiments. Biometrics, 22: 810–828.

Conover, W.J., 1980. Practical Nonparametric Statistics. Wiley, New York, NY.

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

Draper, N.R. and Smith, H., 1966. Applied Regression Analysis. Wiley, New York, NY.

Feldman, R.M., Curry, G.L. and Wehrly, T.E., 1984. Statistical procedure for validating a simple population model. Environ. Entomol., 13: 1446–1451.

Greenwood, D.J., Neeteson, J.J. and Draycott, A., 1985. Response of potatoes to N fertilizer: Dynamic model. Plant Soil, 85: 185–203.

Hameed, S., 1974. Modelling urban air pollution. Atmos. Environ., 8: 555–561.

Harrison, S.R., 1990. Regression of a model on real-system output: An invalid test of model validity. Agric. Syst., 34: 183–190.

Jørgensen, S.E., 1986. Fundamentals of Ecological Modelling. Elsevier, Amsterdam.

Jørgensen, S.E., Kamp-Nielsen, L., Christensen, T., Windolf-Nielsen, J. and Westergaard, B., 1986. Validation of a prognosis based upon a eutrophication model. Ecol. Modelling, 32: 165–182.

Kendall, M.G. and Babington-Smith, B., 1954. Tracts for Computers, No. XXIV, Tables of Random Sampling Numbers. Cambridge University Press, Cambridge.

Kleijnen, J.P.C., 1987. Statistical Tools for Simulation Practitioners. Marcel Dekker, New York, NY.

Law, A.M. and Kelton, W.D., 1982. Simulation Modeling and Analysis. McGraw-Hill, New York, NY.

Loague, K. and Green, R.E., 1991. Statistical and graphical methods for evaluating solute transport models: Overview and application. J. Contam. Hydrol., 7: 51–73.

McCarl, B.A., 1984. Model validation: An overview with some emphasis on risk models. Rev. Market. Agric. Econ., 52: 153–173.

McKinion, J.M. and Baker, D.N., 1982. Modeling, experimentation, verification and validation: Closing the feedback loop. Trans. Am. Soc. Agric. Eng., 25: 647–653.

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.

Picard, R.R. and Cook, R.D., 1984. Cross-validation of regression models. J. Am. Stat. Assoc., 79: 575–583.

Reynolds, M.R., Burkhart, H.E. and Daniels, R.F., 1981. Procedures for statistical validation of stochastic simulation models. For. Sci., 27: 349–364.

Ripley, B.D., 1987. Stochastic Simulation. Wiley, New York, NY.

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

Siegal, J.S., 1972. Development and accuracy of projections of population and households in the United States. Demography, 9: 51–68.

Smith, S.K., 1987. Tests of forecast accuracy and bias for county population projections. J. Am. Stat. Assoc., 82: 991–1003.

Summers, J.K., Kazyak, P.F. and Weisberg, S.B., 1991. A water quality model for a river receiving paper mill effluents and conventional sewage. Ecol. Modelling, 58: 25–54.

Whitmore, A.P., 1991. A method for assessing the goodness of computer simulation of soil processes. J. Soil Sci., 42: 289–299.