ELSEVIER

Agricultural Systems, Vol. 54, No. 3, pp. 313–326, 1997

PII: S0308-521X(96)00077-7 0308-521X/97 $17.00+0.00

Misuse of Regression for Empirical Validation of Models

P. L. Mitchell*

Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK

(Received 24 January 1996; accepted 24 July 1996)

ABSTRACT

Regression of observations from the real system on model predictions is sometimes used for empirical validation. The arguments against this procedure are: (1) that it is a misapplication of regression; (2) that null hypothesis tests are ambiguous; (3) that regression lacks sensitivity in this context because distinguishing the points from a random cloud is rarely necessary at this stage of model development; (4) that the fitted line is irrelevant to model performance; and (5) that the assumptions for regression can be difficult to satisfy. An alternative method is given which concentrates on the deviations (prediction minus observation) and where the modeller has to define criteria for an adequate model with reference to its purpose. This method can be rigorous, objective and quantitative, and is also easy for non-modellers to understand. © *1997 Published by Elsevier Science Ltd. All rights reserved*

INTRODUCTION

Modellers generally accept that the modelling process includes a stage of validation, *pace* Oreskes *et al.* (1994) who discuss the confusion in terminology and philosophical difficulties. Validation is taken here to mean checking that the model structure or its outputs are sufficiently close to the workings or observed states of the real system. The relative emphasis on model structure or outputs depends on whether the model is primarily to explore the workings of the system and increase understanding, or to make predictions

*To whom correspondence should be addressed. Fax: (0114) 276 0159; E-mail: P.L.MITCHELL@SHEFFIELD.AC.UK

that will replace observations of the system. Beck (1987) and McCarl (1984) provide reviews of validation, concentrating on the overall rationale and on validation of model structure. The comparison of predictions from the model with observations from the real world, together with an assessment of model performance, is empirical validation. It is only a part of the whole process of validation but an important part for models that are to be applied: where the predictions are used instead of actual measurements on the real system which may be too costly or difficult to make.

Many books on modelling give little guidance on empirical validation (e.g. Burghes & Wood, 1980; Jeffers, 1982; Charles-Edwards *et al.*, 1986; Thornley & Johnson, 1990; Trenberth, 1992). It is, therefore, not surprising that modellers resort to the simplest procedure that comes to hand and seems to be suitable. I suggest that this is the reason why empirical validation is so often presented as a scatter graph of prediction and observation (e.g. Carberry & Abrecht, 1991; Clewett *et al.*, 1991; Uehara & Tsuji, 1991; Aber & Federer, 1992; Warnant *et al.*, 1994), sometimes with regression which is intended to be an objective and quantitative measure of how good the model is (e.g. Nemani & Running, 1989; Hammer & Muchow, 1991; Keating *et al.*, 1991; Parton *et al.*, 1993; Paruelo & Sala, 1995; Woodward *et al.*, 1995). Regression has been promoted for validation by Reckhow *et al.* (1990); Flavelle (1992) and Mayer *et al.* (1994) but deprecated by Harrison (1990).

The aim of this paper is to explain why regression is not appropriate for empirical validation and to outline an alternative method. Empirical validation must demonstrate to users of models that the model is adequate for its purpose. For this reason it should be objective and readily understandable without a deep knowledge of modelling or of mathematics in general. Hence, in this paper I concentrate on verbal arguments although some knowledge of elementary statistics is required.

A statistical test or a method of empirical validation can claim to be objective if all individuals using the same procedure would reach the same conclusions from a given set of data. The conclusions would not depend on the knowledge or bias of the individual but on things external to the individual. For a statistical test, the 'same procedure' means that it is agreed which test statistic to calculate and by which formula, and what the critical values are for given thresholds of significance. There is implicit agreement on the theory underlying the test (probability, distribution functions) and there has to be explicit agreement that the assumptions for the test are satisfied by the data. It is in this area that individual experience and preferences can produce different results because one person may accept, for example, more heterogeneity of variance than another who would consider a transformation necessary. There are guidelines for this, but few rigid rules (Finney, 1973; Sokal & Rohlf, 1981; Gilbert, 1989) so that each case must be supported by

argument as far as possible. Similarly for a method of empirical validation: agreement is needed on the criteria to be used. These must be stated explicitly and justified with reference to the purpose of the model. Opinions may differ on the weight to be attached to supporting arguments, but there is much less room left for personal bias, especially concealed bias, which is where subjectivity lies.

HOW REGRESSION APPEARS TO BE SUITABLE

The data for empirical validation frequently consist of pairs of predictions and observations. If the observations are collected first then the model can be run for the appropriate conditions to obtain the comparable predictions. In other cases the model predictions may be generated first and observations collected from the real system for conditions identical as far as possible to the model runs. The observations should be a set gathered specially for validation or a set kept separate from data used in model construction or calibration (estimation of values for parameters) if validation is to be an independent evaluation of the model.

Predictions and observations are plotted on a scatter graph. For the arguments set out below it makes little difference whether predictions or observations are the independent variable on the x-axis. If there was perfect agreement the points would fall on a line of perfect correspondence, 1:1, passing through the origin. This is never so in practice but a regression line can be computed from the points and its statistical significance calculated. The value of r^2, the coefficient of determination, can be examined: this lies between 0 and 1 and indicates the fraction of variation explained by the regression. Statistical tests can be carried out for whether the slope of the line differs significantly from 1·0 and the intercept from zero. This line of thought is the 'intuitive appeal' of regression (Harrison, 1990) and the steps given above are the way a biologist tends to apply the statistics learnt for analysing experimental results. Harrison (1990) and Mayer *et al.* (1994) employ a more sophisticated regression analysis, using an F-test for the simultaneous null hypotheses that the slope is 1·0 and the intercept zero.

An example is shown in Fig. 1 and the regression statistics are given in Table 1. The model is for the growth of a grass sward (A.C. Terry, personal communication) developed from the models of Sheehy *et al.* (1979, 1980). It is driven by daily weather data and predicts shoot dry weight. The observations are from six harvests during the growing season of shoot dry weight at five sites. The model was run with weather records from the sites with simulated harvests on the appropriate dates.

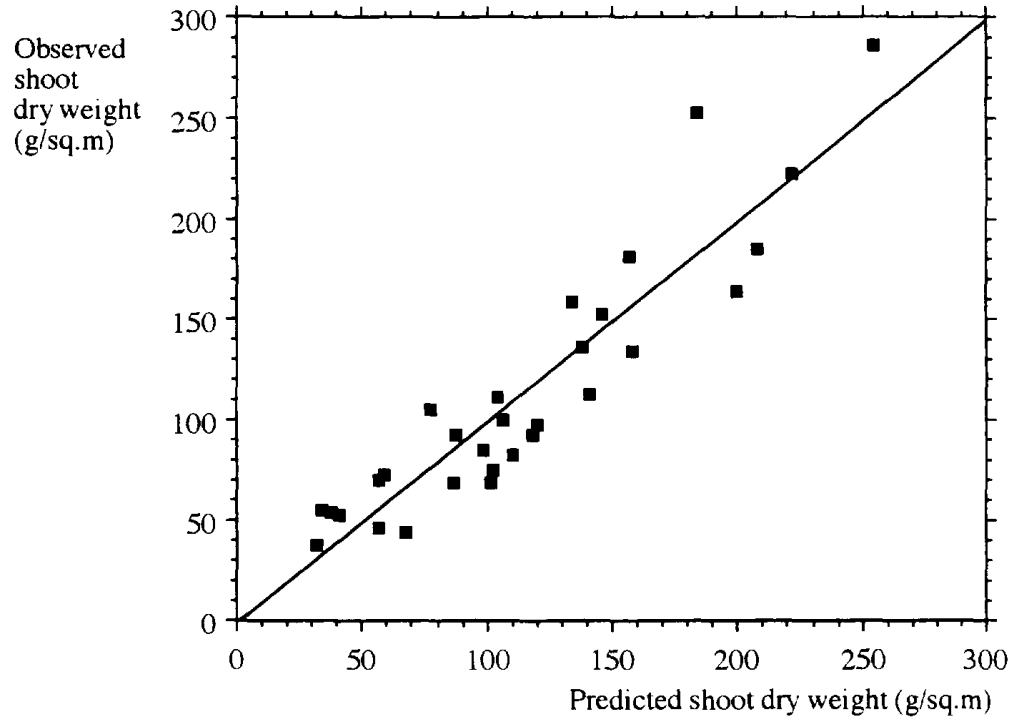

Fig. 1. Observed shoot dry weight from harvested plots of grass sward compared with predictions from a model of grass growth driven by the weather records for the harvest locations. The line is the regression of observations on predictions (details in Table 1). In this case the regression line is very close to the 1:1 line which has been omitted for clarity.

TABLE 1
Regression Statistics for the Attempted Empirical Validation Shown in Fig. 1

Regression equation: observation = −1·8430 + 1·0028 × prediction.
$n = 30$; 28 degrees of freedom.
$r^2 = 0{\cdot}85259$.
Table of analysis of variance.

	Degrees of freedom	*Sum of squares*	*Mean square*	*Calculated value of* F
Source				
Regression	1	98 687	98 687	161.95 *** ($P < 0.001$)
Residual	28	17 063	609	
Total	29	115 750		

For the tests below the critical value of t is 2·048 for $P = 0{\cdot}05$, 28 degrees of freedom.
Variance of slope = 0·0062 101.
Test on difference of slope from 1·0: $t = 0{\cdot}036$, not significant ($P > 0{\cdot}05$), 28 degrees of freedom.
95% confidence limits of slope 0·84 to 1·16.
Variance of intercept = 101·92.
Test on difference of intercept from zero: $t = 0{\cdot}183$, not significant ($P > 0{\cdot}05$), 28 degrees of freedom.
95% confidence limits of intercept −22·51 to 18·83.

WHY REGRESSION IS INAPPROPRIATE

There are five objections to the use of regression. The first two are fundamental; the other three cause difficulties in practice.

1. Misapplication of regression. The most common purpose of regression is to estimate y from x when certain assumptions are satisfied (Draper & Smith, 1981; Gilbert, 1989; Webster, 1989); a less frequent application is the estimation of the best value for the slope between two variables (functional relationship) where an ordinary least squares regression is appropriate in certain cases (Sokal & Rohlf, 1981). In the case of empirical validation there is no interest in estimating a predicted value given an observation, or vice versa, so regression is not being used for its main purpose. The fraction of variation in the y values explained by the regression (r^2) is of no relevance to validation since it is not intended to make predictions from the fitted line.
2. Ambiguity of null hypothesis tests. The test of whether the slope of the fitted line differs significantly from 1·0 can be carried out as an F-test or a t-test. In formal terms the null hypothesis is set up that the slope equals 1·0 and the test attempts to falsify this; if falsification fails then the slope is said not to differ significantly from 1·0. Unfortunately this kind of test cannot be successful unambiguously (Sokal & Rohlf, 1981, p.173; Harrison, 1990; Reckhow *et al.*, 1990) because the more scatter in the points, the greater the standard error of the slope and the smaller the computed value of the test statistic so that it is harder to reject the null hypothesis. This leads to the paradoxical result that regressions from highly scattered samples of points are more likely to have slopes not significantly different from 1·0! The test is ambiguous because falsification of the null hypothesis can fail either because the slope is really not different from 1·0 or because there is much scatter around the line. An alternative procedure, equivalent to the tests but understandable without the formal logic of statistics, is to examine the confidence limits for the slope at 95% (equivalent to $P=0{\cdot}05$) or any other value, using the appropriate value of t from tables. In Table 1 the 95% confidence limits for the slope are 0·84–1·16. The regression is very highly significant, the variance of the slope is small, and the value of t is close to its lowest possible of 1·96; visually the scatter around the line is not excessive (Fig. 1) but still the slope of the line in 95% of similar samples would lie in this sizeable range around 1·0. If the scatter was worse the range would be even larger. Exactly the same arguments apply to the test that the intercept does not differ from zero; Table 1 shows how large the 95% confidence limits of the intercept can be for a very highly significant regression.

3. Lack of sensitivity. The statistical significance of a regression is given by the F-test from analysis of variance (testing the null hypothesis that the slope is zero). A significant regression means that the points are not randomly scattered or following a curve. For a model that has reached the stage of serious empirical validation there is certain to be a good general correspondence of prediction and observation so that finding a significant regression is inevitable and therefore trivial. Experience shows that any cloud of points with a tendency to avoid two opposite corners will have a significant regression line, as the example in the Appendix demonstrates. Regression is not sensitive enough to quantify how good the line is, once past the conventional thresholds of $P = 0{\cdot}05$, $0{\cdot}01$ or $0{\cdot}001$ which are easily attained. This has been recognized by those using regression for its intended purpose. Wetz's criterion for a regression to be of practical use in predicting y from x (Draper & Smith, 1981, p.93) is that the calculated value of F in the analysis of variance should be much larger than the critical value for significance at $P = 0{\cdot}05$. The multiplier varies slightly with the number of degrees of freedom for the regression and the residual, but in general it is is 4, or 6 for a stiffer test of usefulness.
4. The fitted line is irrelevant to validation. The use of regression concentrates attention on the fitted line which is merely the best summary of a straight line relationship among the sample of points provided by pairs of predictions and observations. The fitted line is not part of the model nor is it model output and cannot be of direct relevence to model performance. The deviations, in contrast, calculated as prediction minus observation, give direct information on how far the model fails to simulate the system exactly. The deviations are the subject of the alternative method of empirical validation given below.
5. Violation of assumptions. The data used for empirical validation rarely satisfy the assumptions of linear regression. The first assumption is that the x values are known without error. This can be true if the model is deterministic and the predictions are used as x (Mayer *et al.*, 1994), as in the example in Fig. 1. The other assumptions concern the y values: they should be a random sample, independent of one another, with homogeneous variance along the x-axis and with residuals Normally distributed (Draper & Smith, 1981; Sokal & Rohlf, 1981). The assumption of independence is suspect if the observations are values from a series in time or space or are accumulated values or are autocorrelated in any other way. If the observations cover a large range then homogeneity of variance needs critical examination because larger values often have larger variability.

AN ALTERNATIVE METHOD FOR EMPIRICAL VALIDATION

This will be published in more detail elsewhere (Mitchell & Sheehy, 1997) so only a short account is given here using the data in Fig. 1. The essence of this method is that the deviations contain the important information in this set of data; that models cannot be perfect and need only be adequate so some limitation of precision has to be accepted; and that the size of the acceptable precision can be defined with reference to the purpose of the model. In this way some of the philosophical problems raised about validation can be avoided, especially that a model cannot be proved to be correct or true (Thornley & Johnson, 1990; Oreskes *et al.*, 1994). The method may appear harsh because it can so clearly expose a model as not adequate for its purpose, but this accords with the observation of Amthor & Loomis (1996) that most crop models 'fail most tests with independent data', i.e. have not been validated empirically. With such a result the model is being assessed realistically: it is not a complete failure, but needs to be developed further or exchanged for a better model. Thus, validation is seen as a stage in the continuous, cyclical process of model development and not an end point.

One purpose of the grass model mentioned above is to examine the effects of different climates and weather patterns, or amounts of controllable inputs such as water in irrigation or nitrogen fertilizer, on the shoot dry weight. This is the food available to grazing livestock or for harvest as hay or silage. In this example it is supposed that the precision (as 95% confidence limits) with which shoot weight can be measured in the field is ± 30 g/m^2. It is unreasonable to expect the model to perform as well as this so a less stringent criterion of ± 35 g/m^2 is proposed, which is still of practical use. This definition in terms of absolute precision is added to the graph of deviations (Fig. 2) as the envelope of acceptable precision in which at least 95% of points must fall if the model is to be regarded as adequate for its purpose. In this case two points lie outside the envelope so strictly speaking the model is not adequate (28/30 points = 93%). Arguably the model is borderline since in practice with 30 points the specified limit is $1\frac{1}{2}$ points outside the envelope which cannot occur with a single sample of points.

The main ideas in this method are as follows:

1. Attention is concentrated on the deviations (prediction minus observation) which are displayed graphically along the range of operation of the model. The uniformity of model performance along the range is evident. The meaning of the deviations and the graphical method are easy to understand which is important when non-modellers have to be convinced of a model's adequacy.

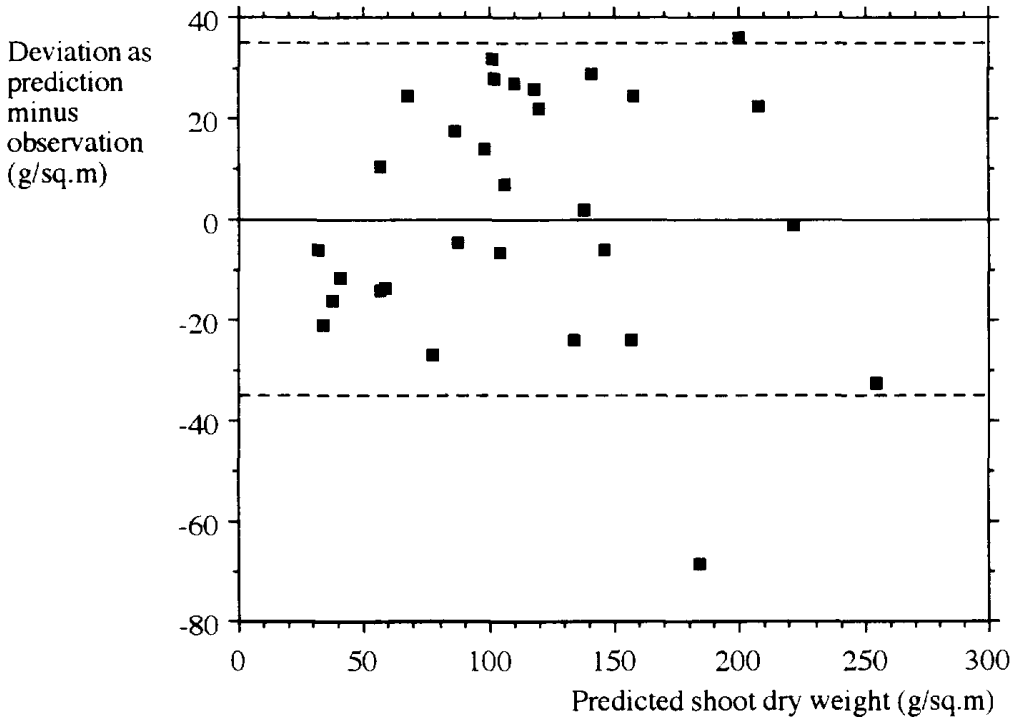

Fig. 2. The graph of deviations for empirical validation of the grass model using the data in Fig. 1. The envelope of acceptable precision shown is ± 35 g/m^2.

2. The criteria for adequacy are defined as the envelope of acceptable precision and the proportion of points that must lie within it. The envelope is drawn on the graph of deviations and delimits a region in which deviations are considered to be acceptably small. The envelope can be any shape and size, made up of absolute or relative precision, or both, at different ranges along the x-axis, symmetrical or not about the horizontal line of zero deviation. The proportion of points can be set at 0·95 by analogy with confidence limits used in statistical analysis of experiments.
3. The definition of the envelope of acceptable precision must be justified with reference to the purpose of the model. Comparison with the precision of the observations, particularly where these are routine field measurements that model predictions could replace, is probably the commonest case. Otherwise the precision of measurements or estimates of parameter values may be a suitable reference point. In some cases the precision of a model may be specified externally, for example a legislated safety standard.
4. The greatest rigour is attained if the envelope of acceptable precision and proportion of points within are defined before the observations for validation are examined. Once Fig. 2 is drawn, it is easy to imagine how

an envelope that included 95% of points could be defined by inspection, to achieve an adequate model.

5. Even if the criteria for adequacy are not defined in advance, the stated justification (section 3 above) provides some objectivity in this method. In the example shown the envelope of acceptable precision is based on the precision of field measurement but with an additional allowance (17% to obtain the round figure of ± 35 g/m^2) for the model not to be as good as measurements. The criterion has been set explicitly even if it is arguable. Similarly, the decision about adequacy of the model is debatable because of the small number of points, but the basis on which the decision can be made has been laid out clearly so that it can be reviewed by anyone.

Statistical tests on the deviations have been proposed for empirical validation. A paired *t*-test on the predictions and observations (e.g. Hassall *et al.*, 1994) is equivalent to a *t*-test that the mean deviation does not differ significantly from zero. This test suffers from the ambiguity of the null hypothesis in exactly the same way as tests for a slope of 1·0 and intercept of zero in regression. To get round this, the test can be formulated as a one-way *t*-test for the mean deviation being less than a specified value (Reckhow *et al.*, 1990) when the specification of the critical value is similar to the criterion of an envelope of acceptable precision. The various forms of *t*-test are sensitive to lack of independence in the variate (Reckhow *et al.*, 1990; Mayer *et al.*, 1994) and also require approximate Normality: assumptions that the method outlined here does not require.

DISCUSSION

The use of regression for empirical validation has been discussed by Reckhow *et al.* (1990), Flavelle (1992) and Mayer *et al.* (1994). All these workers recognize the need for an objective and quantitative method for evaluating models and claim these benefits from regression, and all acknowledge the problems in satisfying the assumptions. Flavelle (1992) emphasizes that the familiarity of regression is a particular advantage. Although regression is widely known and readily available in computer packages, it is poorly understood judging from its frequent misapplication (Webster, 1989).

Harrison (1990) concluded that regression should not be used for validation because of the difficulty in meeting the assumptions and the ambiguity when the null hypothesis cannot be rejected. Mayer *et al.* (1994) took up this issue, but did not fully answer all of Harrison's objections. They demonstrated by Monte-Carlo simulation that the *F*-test for regression was a reliable method

for distinguishing valid models from others; that is to say, models where the predictions and observations do have a linear relationship with slope of 1·0 and intercept of zero from those having no such relationship. This is of limited use since most empirical validation is concerned with models developed well past the initial stage at which the performance is so poor that the validation data do not give a significant regression.

Reckhow *et al.* (1990) tackled the ambiguity of null hypothesis tests by proposing one-tailed *t*-tests where the estimated regression coefficient is less or greater than a pre-determined criterion, not just different. Instead of testing the null hypothesis that the slope is 1·0 they set a criterion that the slope should lie in a defined range around 1·0. However, this poses the question of whether the limits should be 0·95–1·05 or 0·9–1·1 or any other pair of values. Although Flavelle (1992) interprets the slope of the line in a general way in terms of model bias, there is no direct, quantitative relationship of the slope of the fitted line to the performance of the model. Thus, there is no objective way to select limits for the slope that are relevant to model performance.

In conclusion, regression is inappropriate for empirical validation, except perhaps in the very early stages of evaluation of competing models envisaged by Mayer *et al.* (1994). (Their title claims great generality: 'an appropriate overall test of model validity', but I believe that 'initial' instead of 'overall' would be more realistic.) Even in these limited circumstances I would argue that choice among competing models should be made after structural (conceptual) validation and that empirical validation—comparison of predictions and observations—is premature. The alternative method proposed (Mitchell & Sheehy, 1997) does not depend on satisfying the assumptions necessary for statistical tests, and can be made rigorous, objective and quantitative. It can be made objective in the sense that the criteria are explicit and relevant to the purpose of the model; if the criteria are agreed then application of the method by anyone will produce the same result. The method is flexible and can be applied to different sorts of data: spot comparisons, time courses, or spatial sequences. In addition, it is easy to understand which is vital if non-modellers are to be convinced of the power and utility of modelling.

ACKNOWLEDGEMENTS

Financial support was from awards to Professor F. I. Woodward through the European Community MEDALUS (Mediterranean Desertification and Land Use) programme and through the Natural Environment Research Council TIGER (Terrestrial Initiative in Global Environmental Research) programme, and from the International Rice Research Institute as funds to attend the symposia Application of Systems Simulation in Rice Research

(SARP) and the Second International Symposium on Systems Approaches for Agricultural Development (SAAD2) at I.R.R.I., December 1995. I am grateful to Professor F. I. Woodward, Dr J. E. Sheehy and participants at the symposia for encouragement in this work, and to an anonymous referee for helpful comments.

REFERENCES

Aber, J. D. & Federer, C. A. (1992). A generalized, lumped-parameter model of photosynthesis, evapotranspiration and net primary production in temperate and boreal forest ecosystems. *Oecologia* **92**, 463–474.

Amthor, J. S. & Loomis, R. S. (1996). Integrating knowledge of crop responses to elevated CO_2 and temperature with mechanistic simulation models: model components and research needs. In *Carbon Dioxide and Terrestrial Ecosystems*, eds G. W. Koch & H. A Mooney. Academic Press, London, pp. 317–345.

Beck, M. B. (1987). Water quality modeling: a review of the analysis of uncertainty. *Wat. Resources Res.* **23**, 1393–1442.

Burghes, D. N. & Wood, A. D. (1980). *Mathematical Models in the Social, Management and Life Sciences*. Ellis Horwood, Chichester.

Carberry, P. S. & Abrecht, D. G. (1991). Tailoring crop models to the semiarid tropics. In *Climate Risk in Crop Production: Models and Management for the Semiarid Tropics and Subtropics*, eds R. C. Muchow & J. A. Bellamy. CAB International, Wallingford, pp. 157–182.

Charles-Edwards, D. A., Doley, D. & Rimmington, G. M. (1986). *Modelling Plant Growth and Development*. Academic Press, London.

Clewett, J. F., Howden, S. M., McKeon, G. M. & Rose, C. W. (1991). Optimising farm dam irrigation in response to climatic risk. In *Climate Risk in Crop Production: Models and Management for the Semiarid Tropics and Subtropics*, eds R. C. Muchow & J. A. Bellamy. CAB International, Wallingford, pp. 307–328.

Draper, N. R. & Smith, H. (1981). *Applied Regression Analysis*, 2nd edn. Wiley, New York.

Finney, D. J. (1973). Transformation of observations for statistical analysis. *Cotton Growing Review* **50**, 1–14.

Flavelle, P. (1992). A quantitative measure of model validation and its potential use for regulatory purposes. *Adv. Wat. Resources* **15**, 5–13.

Gilbert, N. (1989). *Biometrical Interpretation*, 2nd edn. Oxford University Press, Oxford.

Hammer, G. L. & Muchow, R. C. (1991). Quantifying climatic risk to sorghum in Australia's semiarid tropics and subtropics: model development and simulation. In *Climate Risk in Crop Production: Models and Management for the Semiarid Tropics and Subtropics*, eds R. C. Muchow & J. A. Bellamy. CAB International, Wallingford, pp. 205–232.

Harrison, S. R. (1990). Regression of a model on real-system output: an invalid test of model validity. *Agric. Systems* **34**, 183–190.

Hassall, R. B., MacMillan, D. C. & Miller, H. G. (1994). Predicting Sitka spruce yields in the Buchan area of north-east Scotland. *Forestry* **67**, 219–235.

Jeffers, J. N. R. (1982). *Modelling*. Chapman & Hall, London.
Keating, B. A., Godwin, D. C. & Watiki, J. M. (1991). Optimising nitrogen inputs in response to climatic risk. In *Climate Risk in Crop Production: Models and Management for the Semiarid Tropics and Subtropics*, eds R. L. Muchow & J. A. Bellamy. CAB International, Wallingford, pp. 329–358.
McCarl, B. A. (1984). Model validation: an overview with some emphasis on risk models. *Rev. Marketing Agric. Econ.* **52**, 153–173.
Mayer, D. G., Stuart, M. A. & Swain, A. J. (1994). Regression of real-world data on model output: an appropriate overall test of validity. *Agric. Systems* **45**, 93–104.
Mitchell, P. L. & Sheehy, J. E. (1997). Comparison of predictions and observations to assess model performance: a method of empirical validation. In *Applications of System Approaches at the Field Level*. eds M. J. Kropff, P. S. Teng, P. K. Aggarwal, J. Bouma, B. A. M. Bouman, J. W. Jones & H. H. van Laar. Kluwer, Dordrecht, The Netherlands, pp. 437–451.
Nemani, R. R. & Running, S. W. (1989). Testing a theoretical climate–soil–leaf area hydrologic equilibrium of forests using satellite data and ecosystem simulation. *Agric. Forest Meteorol.* **44**, 245–260.
Oreskes, N., Shrader-Frechette, K. & Belitz, K. (1994). Verification, validation and confirmation of numerical models in the earth sciences. *Science* **263**, 641–646.
Parton, W. J., Scurlock, J. M. O., Ojima, D. S., Gilmanov, T. G., Scholes, R. J., Schimel, D. S., Kirchner, T., Menaut, J.-C., Seastedt, T., Garcia Moya, E, Kamnalrut, A. & Kinyamario, J. I. (1993). Observations and modelling of biomass and soil organic matter dynamics for the grassland biome worldwide. *Global Biogeochem. Cycles* **7**, 785–809.
Paruelo, J. M. & Sala, O. E. (1995). Water losses in the Patagonian steppe: a modelling approach. *Ecology* **76**, 510–520.
Reckhow, K. H., Clements, J. T. & Dodds, R. C. (1990). Statistical evaluation of mechanistic water-quality models. *J. Environ. Engng* **116**, 250–268.
Sheehy, J. E., Cobby, J. M. & Ryle, G. J. A. (1979). The growth of perennial ryegrass: a model. *Ann. Botany* **43**, 335–354.
Sheehy, J. E., Cobby, J. M. & Ryle, G. J. A. (1980). The use of a model to investigate the influence of some environmental factors on the growth of perennial ryegrass. *Ann. Botany* **46**, 343–365.
Sokal, R. R. & Rohlf, F. J. (1981). *Biometry: the Principles and Practice of Statistics in Biological Research*, 2nd edn. Freeman, New York.
Thornley, J. H. M. & Johnson, I. R. (1990). *Plant and Crop Modelling*. Oxford University Press, Oxford.
Trenberth, K. E. (editor) (1992). *Climate System Modeling*. Cambridge University Press, Cambridge.
Uehara, G. & Tsuji, G. T. (1991). Progress in crop modelling in the IBSNAT Project. In *Climate Risk in Crop Production: Models and Management for the Semiarid Tropics and Subtropics*, eds R. C. Muchow & J. A. Bellamy. CAB International, Wallingford, pp. 143–156.
Warnant, P., François, L., Strivay, D. & Gérard, J.-C. (1994). CARAIB: a global model of terrestrial biological productivity. *Global Biogeochem. Cycles* **8**, 255–270.
Webster, R. (1989). Is regression what you really want? *Soil Use and Management* **5**, 47–53.
Woodward, F. I., Smith, T. M. & Emanuel, W. R. (1995). A global land primary productivity and phytogeography model. *Global Biogeochem. Cycles* **9**, 471–490.

APPENDIX

This demonstration of how easy it is to obtain significant regressions is not a mathematical proof but I believe that it is a valid way of summarizing practical experience with regressions. Here the regression is tested with the most familiar null hypothesis that the slope is zero. From a table of random numbers, 100 were selected with values between 0 and 99. The second 50 were aligned against the first 50 to create random points when plotted in Fig. 3(a). As expected, a regression of all 50 points is not significant. Pairs of points were removed successively from the initial set, each pair consisting of the farthest point in the top left and bottom right corners. After removing three pairs a significant regression first occurs ($P < 0{\cdot}05$, $r^2 = 0{\cdot}11$, $n = 44$; Fig. 3(b)). Removal of one more pair increases the significance to $P < 0{\cdot}01$ ($r^2 = 0{\cdot}17$, $n = 42$; Fig. 3(c)). When two further pairs are removed a still more significant regression can be obtained ($P < 0{\cdot}001$, $r^2 = 0{\cdot}30$, $n = 38$; Fig. 3(d)). The 38 points in Fig. 3(d) constitute a broad band remaining after six points in each of two opposite corners have been removed from the original random 50. This broad band boasts a very highly significant regression line, although the line explains only 30% of the variation.

In empirical validation the data rarely exhibit as much scatter as the points in Fig. 3(d) but the regression has already used up the practical scale of significance. That is to say, the regression has passed the three conventional levels of significance $P < 0{\cdot}05$, 0·01 and 0·001. Quantitative comparisons of the significance of regression lines can hardly be made once all the lines are significant at $P < 0{\cdot}001$, as is generally the case in empirical validation.

NOTE ADDED TO PROOF

Kleijnen *et al.* (1997) give a methematical proof that the regression of predictions on observations is an incorrect method of empirical validation. They propose an alternative in which the deviations are regressed on the sums of observation and prediction, and a simultaneous F-test is used on the joint hypothesis that for this line both intercept and slope are zero. If the hypothesis is rejected, the model is not valid. The assumptions are that observations and predictions are positively correlated (well-founded—see third objection above) and that the sets of observations and predictions are each Normally and independently distributed

Kleijnen, J. P. C., Bettonvil, B. & Groenendaal, W. (1997). Validation of trace-driven simulation models: a novel regression test. *Management Science* (in press).

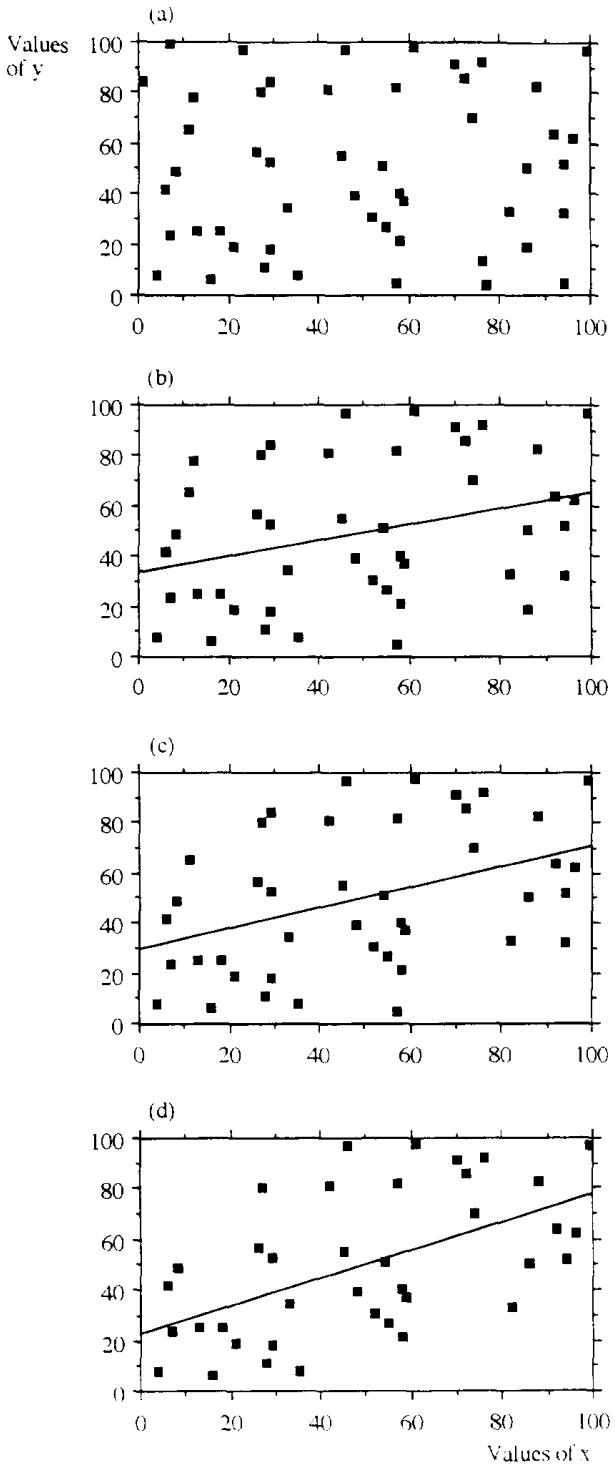

Fig. 3. Fifty random points (a) and subsets created by removing successively pairs of points from top left and bottom right corners. The subsets shown are those where the regression line first exceeds a significance level: $n=44$, $P<0{\cdot}05$ in (b); $n=42$, $P<0{\cdot}01$ in (c); and $n=38$, $P<0{\cdot}001$ in (d).