Journal of Soil Science, 1991, **42**, 289–299

A method for assessing the goodness of computer simulation of soil processes

A. P. WHITMORE
AFRC Institute of Arable Crops Research, Rothamsted Experimental Station, Harpenden, Hertfordshire AL5 2JQ, UK

SUMMARY

Any satisfactory computer simulation model of a soil process must match actual behaviour in the laboratory or field; a model can be evaluated by how well it does so. This paper describes a method for assessing models using anion diffusion and nitrate leaching as examples. The method partitions the sum of squares of the differences between measurement and simulation into two components, one calculated from the differences between the simulation and the mean of replicate measurements (the 'lack of fit'), and the other calculated from the variance within each set of replicate measurements (the 'pure error'). If the former is not significantly larger than the latter than the data present no grounds for rejecting the model. Where a model simulates the change in a process with time the method can also take account of how experimental error in the initial measurements affects the goodness of fit of the simulation of subsequent measurements.

The method is particularly valuable where it is difficult or costly to take many replicate measurements, such as often happens in soil science or agriculture; nonetheless, some replicates must be taken.

INTRODUCTION

Advances in computing power have led to a great expansion in computer modelling. Some models aim at understanding a process and prove their worth with the insights they provide. Others aim to predict one or more components of a process and so require verification to confirm that their predictions are reliable. Modellers need reliable techniques to help them choose the best parameters for their models, and to help them assess the model's worth. As Addiscott & Wagenet (1985) noted when writing about leaching from soils, few authors have published sound methods for evaluating how well models fit or predict measured data.

One obvious method is to plot simulations against measurements on a graph and leave the reader to judge for himself. The better the model, the closer the graph will be to a straight line with a slope of unity passing through the origin; this is sometimes known as the line of perfect agreement (or 1:1 line). This qualitative presentation does not allow the model user to assess 'improvement' to his model objectively, nor does it allow ready comparisons between different authors' models simulating the same data. Addiscott & Whitmore (1987) discussed several ways of quantifying the discrepancy between model and data. They concluded that any one method used alone might be misleading, but several methods used together could summarize satisfactorily the closeness of a model's estimates and measurements. They considered:

(i) the product moment correlation coefficient (r) between measurement and simulation
(ii) the mean difference (M) between measurement and simulation

$$M = \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i) \qquad (1)$$

where y_i is the ith measurement, x_i its simulation and N the number of such pairs.

They also considered the dispersion of $y_i - x_i$, as have Richter *et al.* (1985). If 90% of the simulations are within some arbitrary, but important, range of the measurements, then the model may be satisfactory. This becomes meaningful in practice; if, for example, a farmer cannot apply fertilizer to crops more precisely than to the nearest 10 or 20 kg ha^{-1}, it is pointless to try to predict the optimum rate more precisely than this.

In the statistical analysis of designed experiments the difference or deviation (d) between a measurement and its expected value is known as the residual. In evaluating a computer model it is expected that the simulations should agree with the measurements, so the residuals may be calculated as the difference between measurement and simulation ($y_i - x_i$). Residuals can be positive or negative; in an unbiased model their sum will tend to zero. Greenwood *et al.* (1985) compared the size of the sum of the squares of the residuals with the total sum of squares in the data about their mean. Expressed as a percentage this is analogous to the per cent variation in the data accounted for by the model; however, it can exceed 100%. This quantifies neatly the comparison between models but takes no account of replicated measurements, except to include all replicates in both sums of squares. Some replicated measurements are bound to contain less experimental error than others, and it is sound technique to give these more weight when choosing parameters for, or testing, a model.

If sufficient replicates were taken for each measurement in an experiment, it is possible to use Student's t-test to see whether a simulation is within the experimental error in the measurements (Snedecor & Cochran, 1980). The test is carried out by comparing each simulation (x) with the mean measurement for that simulation $\bar{y}$. The t-statistic is calculated as

$$t = \frac{(\bar{y} - x)}{SE} = \frac{\bar{d}}{SE} \tag{2}$$

where SE is the standard error of the mean measurement, $\bar{y}$, and $\bar{d}$ is the mean deviation. Sutherland *et al.* (1986) used the t-statistic to assess the closeness of each simulation to a set of replicated measurements. They recorded the size and sign of each deviation and ranked these against the measurements. The model is biased if there is a significant correlation between rank order and deviation. The t-statistic is unreliable with very few degrees of freedom; for example, with two replicates t can reach 12.7 before the model must be rejected, or 4.3 for three replicates. These are fairly easy tests for a model to pass. However, it is often costly or time-consuming to take more than two or three replicates and a compromise must be sought.

Addiscott & Bland (1987) have discussed the dangers that disregarding experimental error in input data can have on simulations. One model (Addiscott & Whitmore 1987), that simulates the change in soil mineral-N content over winter, is initiated with a measurement of the mineral-N content in autumn from which it predicts the mineral-N content of the soil in spring. It does this by taking daily account of the weather in the intervening period. It may be unreliable to initiate such a model with a single measurement in autumn because measurements of soil mineral-N in autumn are just as subject to spatial variation as any taken in spring. The model should be assessed on its ability to simulate the change within each replicated block of a field experiment. Ideally the change within any one block with time will be the same as within another; the extent to which this is not true is an additional source of experimental error.

In this paper two related problems are addressed and it is demonstrated how published and novel techniques may be used to overcome them. First, it is shown how to examine the random and systematic deviations of simulations from measurement and, based on the relative sizes of these, a statistic is produced. This statistic should be minimized when choosing optimum parameters for a model. In statistical parlance this is fitting a model to data. It can also be used to assess how well the finished model predicts data independently of that used to develop it. The same statistic also quantifies how well different models estimate the same data, allowing a user to choose between them. Second, it is shown how to use the statistic with any model that predicts the change in a variable with time. Initial and final measurements are equally subject to experimental error: the statistic can take account of this and so be used to obtain a more reliable assessment of a model or its parameters.

THE METHOD

Consider Fig. 1a and c. In these the same data are plotted in two different ways for a model which can be expressed as

$$y \simeq x = f(T) \tag{3}$$

where y are the measurements and T is an independent variable such as time. The equation $x = f(T)$ then represents the output of a model. Both graphs display the data well, although, where x is a continuous function of T, 1a is usually preferred to 1c. Now consider Fig. 1b. This is similar to Fig. 1a, and the model could be said to simulate both sets of data equally well. However, there is a difference in the ways in which the model does *not* fit the data. In Fig. 1a, a shift of origin could enable the model to simulate each experiment perfectly; that is to say, an extra input variable expressing the difference between the experiments could account for the systematic failure of the model to predict each experimental data point. In Fig. 1b, the differences between model and measurement are about the same size as those in Fig. 1a, but now the discrepancy is all random error or 'noise' in the measurements. Although Student's t-statistic is not a useful test of agreement with two replicates, if the errors in all the measurements were pooled and the deviations of the model from their means also pooled, then comparison of error and deviation for all the data together would be much more reliable with additional degrees of freedom: that is essentially how the method works.

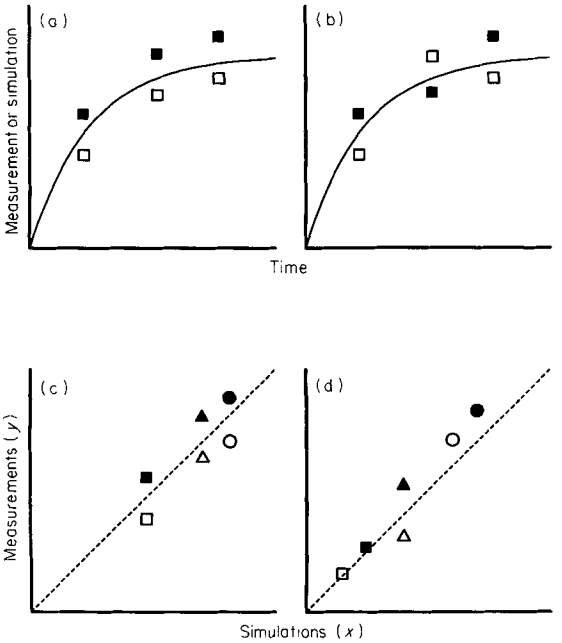


Fig. 1. Relationships between duplicated (open *vs* closed symbols) measurements (y) and simulations (x) where: in the case of (a) and (b) x is a continuous function of a third variable, say time, and for (c) and (d) there is no simple continuous interrelationship between the simulations (x).

All the discrepancy between model and data, whether random or systematic, is in the residuals (d). The sum of the squares of the residuals can be partitioned into two other sums of squares: that due to pure error (i.e. random variation) and that due to lack of fit (i.e. systematic variation). The sum of squares due to pure error is the sum of the squares of the differences between measurements and their mean within each set of replicate experiments. The sum of squares due to lack of fit can be

obtained by subtraction of the error from the residual sum of squares, or computed directly as shown below. It is a measure of the average deviation of the model from data in each experiment. The formula for calculating the sum of squares due to lack of fit is derived in the appendix. Where *RSS* is the residual, *SSE* the error and *LOFIT* the sum of squares attributable to lack of fit then

$$RSS = \sum_{j=1}^{N} \sum_{i=1}^{n_j} d_{ij}^2 = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (y_{ij} - x_j)^2 \tag{4}$$

$$SSE = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (d_{ij} - \bar{d}_j)^2 = \sum_{j=1}^{N} \sum_{i=1}^{n_j} \left((y_{ij} - x_j) - (\bar{y}_j - x_j) \right)^2 \tag{5}$$

$$LOFIT = \sum_{j=1}^{N} n_j \bar{d}_j^2 = \sum_{j=1}^{N} n_j (\bar{y}_j - x_j)^2 \tag{6}$$

where N is the number of experiments, n_j the number of replicates within each, x_j the simulation for the jth experiment and y_{ij} the ith measurement in the jth experiment. The quantity $\bar{y}_j$ is the mean of the measurements in the jth experiment, d_{ij} is the deviation $(y_{ij} - x_j)$, and $\bar{d}_j$ the mean deviation is $(\bar{y}_j - x_j)$.

Dividing each sum of squares by its associated number of degrees of freedom gives a mean square (which is a variance). The relative sizes of the mean squares due to error and lack of fit can be compared using the variance ratio, or F-test (cf. Teng, 1981), and where the lack of fit is significantly greater than error the model could almost certainly be improved. The number of degrees of freedom in the residual is

$$\sum_{j=1}^{N} n_j;$$

in the error

$$\sum_{j=1}^{N} (n_j - 1);$$

and in the mean squares due to lack of fit (obtained by difference) N.

Where a model simulates the change in a variable (x) that must be measured at the start of computer simulations, the method can refine the estimates of lack of fit and error. Fig. 1d shows three possible outcomes from duplicated measurements; to simulate these the model needs to be run twice, initiated with each duplicate in turn. In set 1 (□, ■) the model simulates the data perfectly and the experimental error is zero, in set 2 (△, ▲) the model is not biased, and it simulates the data well given the experimental error. In set 3 (○, ●), however, the random experimental error is tiny just as in set 1, because the model has estimated the same change in each duplicate; but it significantly and systematically underestimates them both. With the notation above, the equations for the various sums of squares where both initial and final measurements were replicated now become

$$RSS = \sum_{j=1}^{N} \sum_{i=1}^{n_j} d_{ij}^2 = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (y_{ij} - x_{ij})^2 \tag{7}$$

$$SSE = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (d_{ij} - \bar{d}_j)^2 = \sum_{j=1}^{N} \sum_{i=1}^{n_j} \left((y_{ij} - x_{ij}) - (\bar{y}_j - \bar{x}_j) \right)^2 \tag{8}$$

$$LOFIT = \sum_{j=1}^{N} n_j \bar{d}_j^2 = \sum_{j=1}^{N} n_j (\bar{y}_j - \bar{x}_j)^2 \tag{9}$$

where $\bar{x}_j$ is the mean of the n_j replicate simulations for the jth experiment and x_{ij} is the ith simulation in the jth experiment. Equation (9) is derived in the appendix.

The way in which the calculations work becomes difficult to represent graphically with more than two replicates as in Fig. 1d, but the statistics will obviously cope with three or more replicates.

EVALUATION

Using a computer simulation model of a continuous process

Addiscott's (1982) model simulates the diffusion of anions within cubic or other regularly shaped objects. He tested his model by measuring the diffusion of bromide ions out of different sizes of chalk cube. Diffusion is supposed to take place between imaginary concentric volumes within the cube, and also between the outermost volume and the surrounding solution according to Fick's first law. In the limit, calculus might give an exact answer, but the program solves the equations numerically for a finite number of concentric volumes. The model requires the volume and initial concentration of anions in the cube (assumed the same in each concentric volume); in addition, the porosity and the tortuosity within a cube must be estimated. Once a cube is immersed in distilled water, gentle stirring begins and continues until an equilibrium concentration of anion is achieved in solution. Expressing the concentration C in solution at time T as a proportion of this equilibrium concentration (C_{eq}) gives a curve of the kind shown in Fig. 1a. Addiscott's (1982) actual data, Fig. 2, are reproduced with permission from the author. The model was tested in three different experiments by putting either 20 mm, 40 mm or 75 mm chalk cubes, previously saturated with bromide solution, into distilled water. The increase in bromide ion concentration in the surrounding solution was then measured with time. Fig. 2a, b and c show C/C_{eq} plotted against time for bromide ion in solution. Table 1a shows the mean squares attributable to the residual, lack of fit and pure error. Two separate experimental runs were made for each size of chalk cube; measurements made at the same time in the two experimental runs were treated as duplicates. Measurements were also used as duplicates, where their simulated values remained almost the same for a very small difference in time T. Table 1a shows that the model simulations taken as a whole do not differ significantly from the measurements made of diffusion of bromide out of the 20 mm chalk cube, but are just significant for the 40 mm cube. However, the lack of fit is very significantly larger than error for the 75 mm cube size and the bias at the beginning ($T < 1000$ min) is clear in Fig. 2c. Table 1 also has a statistic that shows what proportion the error and lack of fit together are of the total variation about the mean of the data (Greenwood *et al.*, 1985). In all cases they are rather small.

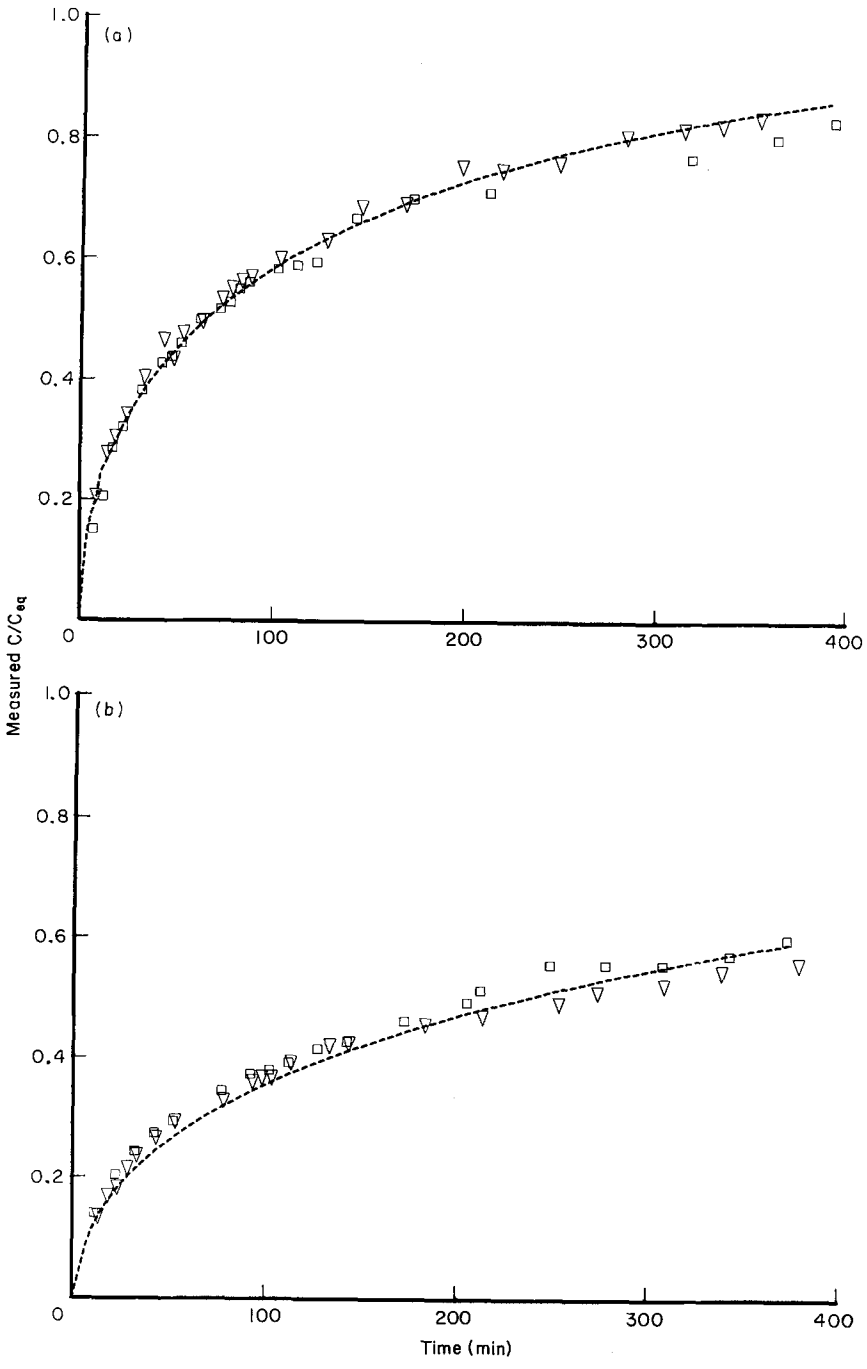
Table 1. Statistics to evaluate the method for assessing: (a) model simulations of the diffusion of bromide out of three sizes of chalk cube, (b) simulations of soil mineral-N in spring

	[a]*RSS*	[a]*MSE*	[a]*MSLOFIT*	[a]Ratio	[a]% Variance
(a)					
[b]20 mm	0.02178	0.0002404	0.0005557	2.312	1.41
[b]40 mm	0.01657	0.0001449	0.0005214	3.600	2.46
[b]75 mm	0.03619	0.0000482	0.001703	35.35	2.43
(b)					
	5498	92.4	205.0	2.22	27.3

[a]*RSS* = residual sum of squares; *MSE* = mean square due to pure error; *MSLOFIT* = mean square due to lack of fit; Ratio = *MSLOFIT*/*MSE*;

% variance $= 100\left((\mathrm{RSS}/\sum_{j=1}^{N}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2\right)$ after Greenwood *et al.* (1985); [b]sizes of chalk cubes.

Addiscott (1982) suggested that the initial distribution of bromide within the 75 mm cubes might not have been as uniform as in the 20 mm and 40 mm cubes. Such lack of uniformity violates assumptions within the model. Assumptions may need careful examination in the light of experimental or model results. The method assesses how closely the model simulates the data but it cannot distinguish, in this instance, between failure of the model and experimental error in the input data.



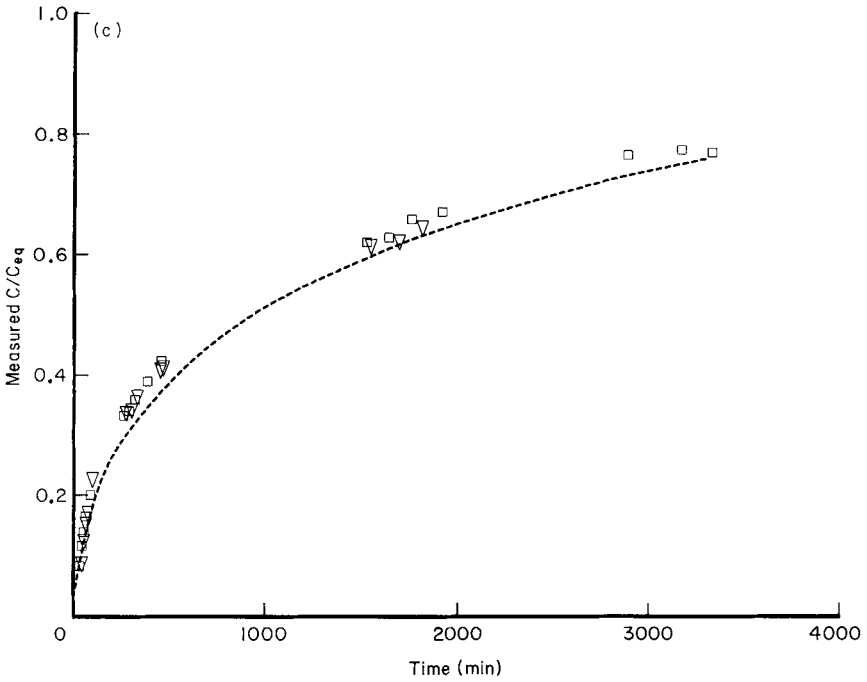


Fig. 2. Duplicate (□ and ▽) simulations made by the model of Addiscott (1982) of the diffusion of bromide ions with time out of chalk cubes of sizes: (a) 20 mm, (b) 40 mm, and (c) 75 mm.

This becomes possible only where the number of separate experiments is large and each is replicated, as is usually the case when simulating discontinuous data.

Using a model to simulate data from unrelated experiments

Where a model simulates results from many experiments the relationship between measurement and simulation is discontinuous. Addiscott & Whitmore (1987) describe an example of this use of a model. Their model estimates the amount of mineral-N in soil in spring at the time farmers apply N fertilizer to winter wheat. It determines the amounts of nitrate-N leached, ammonium-N released from soil organic matter and nitrified to nitrate-N, and the N taken up by the crop each day. The inputs to the model are daily rainfall, evaporation and soil temperature, the water-holding capacity and texture of the soil, the mineral-N in the soil at the start of the computer simulation, the soil moisture deficit at the end of August, the sowing date of the crop, and the recent cropping history of the field. The model must be assessed with different experiments from different years and in different parts of the country because only one or two sets of replicated experiments are likely to be made during a growing season and because the model is intended to be used throughout the country. Diagrams similar to these in Fig. 1c and d provide the best graphical representation of the performance of this kind of model.

Addiscott & Whitmore (1987) modelled the change in mineral-N content of the soil over the winter period starting with a measurement made in autumn. They assessed their simulations against results from many experiments, but only those experiments where duplicate measurements were made in both autumn and spring are used here. In Fig. 3 all the measurements made in spring (that is both duplicate measurements) are plotted against simulations. There is no elegant way of representing the connection between each pair of duplicates graphically as in Fig. 1d, which shows different symbols for three experiments only. Table 1b shows the mean squares attributable to error, lack of fit and the residual sum of squares. The variance ratio is significant but not large for

the 20 sets of data. The residual variation (between model and measurement) is about 27% of the variation in the measurements about their mean; this suggests that the model is sound and that the range against which it has been tested is fairly large compared with experimental error and lack of fit.

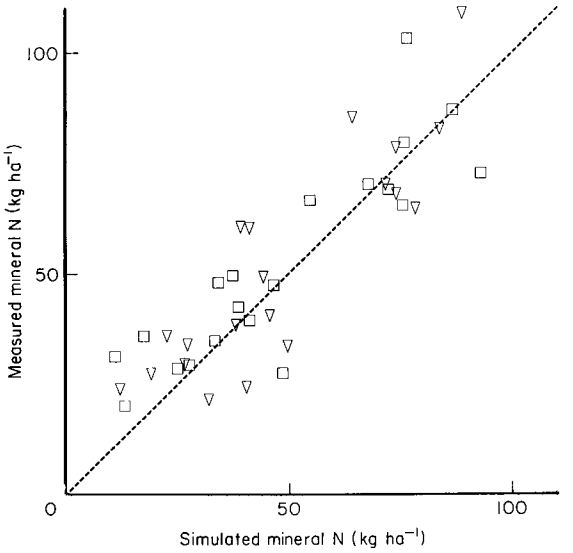


Fig. 3. Simulations of the amounts of mineral N in soil in spring made by the model of Addiscott & Whitmore (1987) compared with duplicate measurements (□ and ▽).

DISCUSSION

The design of an experiment contributes much to its success. A poorly designed sampling arrangement can lead to the rejection of a good model or the acceptance of a poor one. In any scheme for collecting data resources are limited, which sets bounds on the number of measurements taken. However, it is advisable that each measurement should be replicated. It is equally sensible to test the model fully throughout its intended range of application, keeping replication to two or three. This range may be a time period for a continuous model or the whole country for one that estimates soil mineral N. In this way the numbers of degrees of freedom of pure error and lack of fit will be similar. Poor estimates of a model's success may result if either of the pure error sum of squares or the sum of squares due to lack of fit contains most of the degrees of freedom. If the error term contains most of the degrees of freedom then the model might be tested best by comparing the individual simulations with means of the measurements using a *t*-test as described by Sutherland *et al.* (1986).

If the aim is to find parameters for a model by fitting it to data, then it is unwise to neglect lack of fit which warns of bias. If the aim is to assess a model then Fig. 2c and Table 1 illustrate a very important point: no predictive model can ever be perfect. There will always be small differences between measurements and simulation. A given level of significance can always be reached by taking more measurements; and so users should not put too much reliance on the significance (in the statistical sense) of the lack of fit. According to the criterion of Greenwood *et al.* (1985) the model used to describe the loss of anions from a 75 mm chalk cube was a good one—it explained more than 97% of the variation in the data during an experiment lasting over 4000 min. Analysis of lack of fit warns us that the model may be less good if used solely to predict the loss of anion after say 500 min.

In pooling variances or testing their relative magnitude it is assumed that although the means of sets of replicates from the various experiments differ, the variances are similar and the data

independent and normally distributed. For instance, if most replicates differ from their means by 10 units (min, or kg N ha^{-1} say) then a single one which differs by 100 or even 1000 units may invalidate the analysis. Measurements that appear to be widely different from the rest should be investigated separately in case they bias the results. The data in Fig. 3 and Table 1b illustrate this point. Some of the experimental blocks were waterlogged over winter (R. J. Darby, personal communication), which is not a fair test of a model that assumes the soil is freely draining. If such extraneous variation is commonplace, as appears to be the case in Table 1b, then it may indicate that the model is missing some important factor. By examining the individual sets of residuals it should be possible to identify the experiments where the model fails, and elucidate the reasons and any deficiencies such as the need to model waterlogging.

CONCLUSIONS

To choose parameters for a computer model or assess how well it predicts independent data I suggest the following procedures. In all cases plot the data on a graph.

(i) Where none or few of the measurements were replicated

Choose the best parameters for a model by minimizing the sum of squares of the deviations (Greenwood *et al.*, 1985). To evaluate a single model compute the product moment correlation coefficient (r) together with the mean difference (M). Specify the acceptable error in the prediction at the outset. The model may be checked to see what fraction of its predictions fall within this range. Models are best compared with a single statistic, for example Greenwood's. Compare models giving advice against the acceptable error.

(ii) Where most or all of the measurements were replicated

Partition the sum of squares of the deviations between simulation and measurement into the components due to lack of fit and to pure error. Choose parameters for a model that minimize the lack of fit; reduce it to zero if possible. If lack of fit is significantly larger than error overall, then examine the individual experiments. Where lack of fit is generally much larger than error, this implies that the model or its parameters are poor. If the lack of fit is greater in just a few experiments, inspect the data; though poor data would normally inflate the error as well. You are more likely to have omitted some important feature common to these experiments from the model; if so you must decide whether to include this feature or restrict your use of the model.

Evaluate a model or compare different models with the ratio of the mean square lack of fit to the mean square error. Keep in mind that only a very good model indeed will have a statistically non-significant result with more than 10 or 20 data; it may be more meaningful to test models giving advice against the acceptable error.

(iii) For models that simulate changes in a property and for which both initial and final measurements of this property were replicated

Compare error and lack of fit as above. Lack of fit is now a more robust indicator of failings in the model. Minimize lack of fit to find the optimum parameter for a model, examine lack of fit carefully for signs of systematic bias in the model. Evaluate or compare models as in (ii) above.

The author's FORTRAN computer program calculates the sums of squares due to error and lack of fit, and copies can be supplied to interested readers.

ACKNOWLEDGEMENTS

I am grateful to T. M. Addiscott, who made available a model and data for testing and who, along with R. Webster, K. W. Coleman, C. J. Dyer and an unknown referee, made valuable comments on the manuscript. I am also grateful to G. J. S. Ross for helpful discussions on modelling.

REFERENCES

ADDISCOTT, T.M. 1982. Simulating diffusion within soil aggregates: a simple model for cubic and other regularly shaped aggregates. *Journal of Soil Science* **33,** 37–45.

ADDISCOTT, T.M. & BLAND, G.J. 1987. Nitrate leaching models and soil heterogeneity. In *Nitrogen Efficiency in Agricultural Soils* (eds D.S. Jenkinson & K.A. Smith), pp. 394–408. Elsevier Applied Science, London.

ADDISCOTT, T.M. & WAGENET, R.J. 1985. Concepts of solute leaching in soils: a review of modelling approaches. *Journal of Soil Science* **36,** 411–424.

ADDISCOTT, T.M. & WHITMORE, A.P. 1987. Computer simulation of changes in soil mineral nitrogen and crop nitrogen during autumn, winter and spring. *Journal of Agricultural Science, Cambridge* **109,** 141–157.

GREENWOOD, D.J., NEETESON, J.J. & DRAYCOTT, A. 1985. Response of potatoes to N fertilizer: dynamic model. *Plant and Soil* **85,** 185–203.

RICHTER, J., NORDMEYER, H. & KERSEBAUM, C.K. 1985. Simulation of nitrogen regime in loess soils in the winter half-year: comparison between field measurements and simulations. *Plant and Soil* **83,** 419–431.

SNEDECOR, G.W. & COCHRAN, W.G. 1980. *Statistical Methods*, 7th edn. Iowa State University Press, Ames, IW.

SUTHERLAND, R.A., WRIGHT, C.C., VERSTRAETEN, L.M.J. & GREENWOOD, D.J. 1986. The deficiency of the 'economic optimum' application for evaluating models which predict crop yield response to nitrogen fertilizer. *Fertiliser Research* **10,** 251–262.

TENG, P.S. 1981. Validation of computer models of plant disease epidemics: A review of philosophy and methodology. *Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz* **88,** 49–63.

(Received 5 April 1990; accepted 31 October 1990)

APPENDIX

(i) Derivations without including the error in input data

$$RSS = \sum_{j=1}^{N} \sum_{i=1}^{n_j} d_{ij}^2 = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (y_{ij} - x_j)^2 \tag{A1}$$

$$\text{and } SSE = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (d_{ij} - \bar{d}_{ij})^2 = \sum_{j=1}^{N} \sum_{i=1}^{n_j} \left((y_{ij} - x_j) - (\bar{y}_j - x_j) \right)^2 \tag{A2}$$

$$\text{then } RSS - SSE = \sum_{j=1}^{N} \left(\sum_{i=1}^{n_j} (y_{ij} - x_j)^2 - \sum_{i=1}^{n_j} \left((y_{ij} - x_j) - (\bar{y}_j - x_j) \right)^2 \right) \tag{A3}$$

$$= \sum_{j=1}^{N} \left(2 \sum_{i=1}^{n_j} (y_{ij} - x_j)(\bar{y}_j - x_j) - \sum_{i=1}^{n_j} (\bar{y}_j - x_j)^2 \right) \tag{A4}$$

$$= \sum_{j=1}^{N} (2n_j\bar{y}_j^2 - 4n_j x_j \bar{y}_j + 2n_j x_j^2 - n_j\bar{y}_j^2 + 2n_j x_j \bar{y}_j - n_j x_j^2) \tag{A5}$$

$$= \sum_{j=1}^{N} (n_j\bar{y}_j^2 - 2n_j x_j \bar{y}_j + n_j x_j^2) \tag{A6}$$

$$= \sum_{j=1}^{N} n_j (\bar{y}_j - x_j)^2 \tag{A7}$$

$$\text{and } LOFIT = RSS - SSE \tag{A8}$$

(ii) Derivations including error in input data

$$= \sum_{j=1}^{N} \sum_{i=1}^{n} d_{ij}^2 = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (y_{ij} - x_{ij})^2 \tag{A9}$$

$$SSE = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (d_{ij} - \bar{d}_{ij})^2 = \sum_{j=1}^{N} \sum_{i=1}^{n_j} \left((y_{ij} - x_{ij}) - (\bar{y}_j - \bar{x}_j) \right)^2 \tag{A10}$$

$$RSS\text{-}SSE = \sum_{j=1}^{N} \sum_{i=1}^{n_j} \left((y_{ij} - x_{ij})^2 - ((y_{ij} - x_{ij}) - (\bar{y}_j - \bar{x}_j))^2 \right) \tag{A11}$$

$$= \sum_{j=1}^{N} \sum_{i=1}^{n_j} \left(2(y_{ij} - x_{ij})(\bar{y}_j - \bar{x}_j) - (\bar{y}_j - \bar{x}_j)^2 \right) \tag{A12}$$

$$= \sum_{j=1}^{N} (2n_j\bar{y}_j^2 - 4n_j\bar{x}_j\bar{y}_j + 2n_j\bar{x}_j^2 - n_j\bar{y}_j^2 + 2n_j\bar{x}_j\bar{y}_j - n_j\bar{x}^2) \tag{A13}$$

$$= \sum_{j=1}^{N} (n_j\bar{y}_j^2 - 2n_j\bar{x}_j\bar{y}_j + n_j\bar{x}_j^2) \tag{A14}$$

$$= \sum_{j=1}^{N} n_j(\bar{y}_j - \bar{x}_j)^2 \tag{A15}$$

and $LOFIT = RSS - SSE$ (A16)