SOME CONTROVERSIES ON PREDICTIVE AND NON-PREDICTIVE VALIDATION STRATEGIES FOR DYNAMIC, SOCIO-ECONOMIC MODELS

M. Sommer

Universität Bielefeld, Fak. f. Pädagogik, D-4800 Bielefeld, FRG

Starting out with the "classical" system dynamics validation concept, which favours predictive instead of non-predictive procedures (section 2), we will investigate the crucial objections, which Forrester and his co-workers have recently launched against statistical tests as model validation procedures (section 3). The critical appreciation of these research activities results in a plea for accepting rather than neglecting the informations yielded by non-predictive tests as an important step within a multi-stage validation concept (section 4).

1. INTRODUCTION

Validation of dynamic, socio-economic models has long been one of the most controversial issues between followers of a more data-based methodology - mainly econometricians - and those of a more concept-based philosophy - mainly system dynamicists [1,2]. In the last few years some attempts have emerged to assess the degree of compatibility of econometrics and system dynamics and to examine at least four possible future relations between these two important modelling approaches: dominance of one approach and finally elimination of the other [3], convergence into a single and broader methodology [4,5,6], passive coexistence in different ecological niches [1] and different forms of active cooperation [7,8,9,10,11]. Although model validation is just one aspect within a much wider range of characteristics - see [12] for a comprehensive treatment of the subject - it has been the central feature of a sometimes eclectic discussion between both parties long before the just mentioned efforts for systematic inter-paradigmatic comparisons were started. These controversies may well be traced back to the 1962 discussion between Forrester, Holt and Howard [13], i.e. one year after the publication of "Industrial Dynamics" [14].

2. THE "CLASSICAL" SYSTEM DYNAMICS VALIDATION THEORY

(a) Forrester's view

The main source of the SD-validation theory is Chapter 13 "Judging Model Validity" in [14], which is still considered obliging by most system dynamicists (see e.g. [15]). Following Forrester's views validity is not an absolute and "truth"-searching but rather a relative and purpose-dependent concept. Since the dominant purpose of SD-modelling is the improvement of a real system's dynamic behaviour, a model's final validity rests upon the success of system redesign recommendations as results of model experiments. Two problems arise with final validity: 1. since final validity can only be assessed after system redesign it cannot guide the redesign recommendations themselves, 2. it is very hard to examine which part of behavioral improvement in the real system is actually to be attributed to the proposed policy

change [15]. Therefore the researcher has to confine himself with an evaluation of the model's interim validity, which exists of two components: 1. non-predictive validation (a) of the set of endogenous variables necessary for describing the relevant behaviour modes (system boundary), (b) of the interconnections between variables including time-lags and functional forms (specification analysis), (c) and least important the validity of chosen parameter values; 2. predictive validation of the correspondence of model and real system behaviour by reasonable crisis tests, and examination of the model's ability to reproduce trouble symptoms, periodicities, time-phase relationships and transition characteristics observable in the real system. Because of the dominant purpose of real-system improvement by model-guided policy change, behaviour correspondence is only then not a necessary but also sufficient prerequisite for final validity, if it is achieved by a structurally valid model.

(b) Consenting and dissenting views

Model Purpose. EC as well as SD recognizes that model validity heavily depends on model purpose which may easily be underpinned by a quote from a multi-authored, paradigmatic article on econometric model evaluation. "In the current state of our knowledge and analytical needs, to concentrate our attention solely on proving or disproving the 'truth' of an econometric model is to choose an activity virtually quaranteed to supress the major benefits which can flow from the proper use of econometric models. Having constructed the best models of which we are capable, we ought to concern ourselves with whether or not particular models can be considered to be reliable tools for particular uses, regardless of the strict faithfulness of their specification. In this context, 'validation' becomes a problem-dependent or decision-dependent process, differing from case to case as the proposed use of the model under consideration changes. Thus a particular model may be validated for one purpose and not for another." [16].

In spite of very similar statements by SD-authors [14,15] one should be careful not stretch the unison to far, because "model purpose" appearently has a double meaning as can be demonstrated by the following passage. "Thus a model which accurately predicts the employment effects of alternative tax policies may be considered 'successful' even if its prediction of the composition of GNP is poor by the standards for other uses of a model." [16]. SD-followers accord to the relevance of a specific purpose the model should serve, but disagree on the more general aspect of the model purpose: while EC favours accurate predictions of future system states as a desirable goal of knowledge to be gained from modelling, SD holds for prediction of global behaviour characteristics in order to achieve system improvement. This dissent on the proper model purpose is of overwhelming but oftenly underrated importance not only for understanding differences in model structures (e.g. the SD-verdict against and the EC-allowence for exogenous variables) but also for distinctions in the way, predictive validations are performed.

Predictive validation. Forrester has given broad scope on the necessity to distinguish between prediction of behaviour characteristics and the prediction of future system states and argues that only the former is able to contribute to a model's interim validity [14]. This position has been questioned by many critics, e.g. Apel: "A model, where variables reproduce the dynamic behaviour of a real system, only shows that the model already hits the system characteristics. But why employ quantitative modelling, if one rests content with this?" [17]. Actually objections like these raise doubt against the feasability of Forrester's modelling goals (system improvement) and implicitly call for the econometric goal of accurate predictions, but they are not sound criticisms of Forrester's validation by behaviour mode predictions as they claim to be. In fact it can be demonstrated [12] that Forrester's insistence on predictive validation by behaviour correspondence tests is quite defendable, provided that one accepts his view of model purposes. On the other hand it is equally plausible that a model's capability to produce dynamic behaviour similar to that of the real system is not sufficient, if accurate prediction of ex ante values of endogenous variables at certain points of time is the postulated goal

of the modelling effort. We believe that the center point of the validation controversies should not be the alternative between behaviour prediction and future state prediction within the realm of predictive validation itself - because the choice between the two so much depends on the chosen purpose - but should rather concentrate on questions about the relation between predictive and non-predictive validation.

Non-predictive-Validation. We had already touched the crucial point that behaviour correspondence can only be relevant for final validity, if it is accomplished by a structurally valid model. Avoiding to be caught in a vicious circle obviously requires means of assessing the structural and parameter correspondence between a model and reality independently of model behaviour tests. Forrester himself admits that the latter tests are "meaningful only because we believe *independently* that the causal relationships of the actual system are represented in the mechanisms of the model. An endless variety of model details having no similarity to the actual system could be assembled that would create the Model II curve." [14]. A carefull comparison of section 4.7 on "Sources of Information for Constructing Models" and section 13.4 in the validation chapter of [14] substantiates that Forrester recommends to use the same information for validating parameters and specification assumptions (structure) that have already been incorporated within the prior formulation of the model. "In the design and justification of a model, we need to call upon the full variety of knowledge that is available about the system. Most of our knowledge is in the experience and the minds of people who have observed and worked with the system. Much information is in the descriptive literature. Only occasionally will there be numerical and statistical evidence sufficient to settle important model-building questions." [14] (see also [18,19,20]). Since using the same information twice cannot yield any genuine insights, this strategy must be regarded as a spurious validation. On another occasion Forrester did even go further arguing that only poor models present insurmountable questions of validation, in other words: "good" models do not require validation, "bad" models cannot be exposed to validation [21]. The whole validation topic - including Forrester's own treatment - is thus in danger of erosion and might in the end be regarded as a spurious problem. But this may not be the last word, because the question remains unanswered how one knows which models are the "good" ones and which belong to the "bad" species.

Forrester's reluctance to employ statistical tests in the stage of parameter and specification evaluation obviously leads us back to his analogous opinion in "Industrial Dynamics" that statistical methods for parameter estimation are superfluous. It would indeed make no sense to evaluate the validity of parameters, which have been estimated in an informal ad hoc fashion, by statistical test tools. Due to limited space we have to sustain from appraising the SD-theory of ad hoc parameter estimation (see [12] sec. 4.5. for a detailed discussion), which draws heavily on two assumptions: 1. insensitivity of model behaviour to most parameter values wherefore their exact values need not be known anyway and 2. direct observability of parameters from the real system.

With respect to parameter observability we have to diagnose a fundamental dissent between SD and EC: On the scale, which ranks data and a-priori-information used in the process of parameter value determination, we find as extreme cases SD-models at one end, which use only a-priori-information, and EC-models at the other end relying only on data information. [22] But there seems to be a contradiction in the SD argumentation against formal parameter estimation: while claiming that parameters do not have to be estimated because they are observable, variables are held to be frequently unobservable which in turn prohibits statistical parameter estimation because of lack of data on variables. At least for socio-economic models on a macro-level the official statistics provide a vast amount of time-series and cross-section data on variables but almost none on parameters. This is implicitly acknowledged by Lehmann, whose SD-model of the Federal Republic of Germany [23] partially draws on time-series data for variables but not for parameters. Although he recognizes that EC could complement SD in the areas of estimation and validation, he does not make any use of these offers within his own model.

The consequence of this discussion is rather simple: the proposition of the "classical" SD-validation theory, which regarded statistical methods for estimating and evaluating parameters and equation specifications as superfluous is not defendable. This leaves a logical gap in the SD-parameter determination philosophy as well as in the SD-concept of interim validity, causing its pretended multi-stage validation procedure to shrink into the single stage of predictive validation of model behaviour.

3. NEW SD-OBJECTIONS AGAINST NON-PREDICTIVE VALIDATION: SOME CRITICAL REMARKS

Probably as a by-product of the extensive work on an SD-National Model of the U.S.A., the increasing econometric challenge to the SD-methodology has encouraged the reoccupation with the statistical testing of parameters and specifications - in line with research on econometric estimation of SD-models. "Although the literature of regression analysis and econometrics dominates the social sciences in describing the use of data for relating real life to models, much new light can be shed on the proper and possible uses of statistical methods by experiments such as Peter Senge is now conducting at MIT (not yet published)." [24] "The laboratory tests indicate that the generalized least-squares data analysis can give not only major errors in the estimates of parameters but also misleading indications from the internal measures of validity." [25]. These statements must be considered as a remarkable shift in the SD-validation theory: while in "Industrial Dynamics" statistical tests were assumed to be an acceptable, though mostly useless and only exceptionally necessary instrument for independently supporting faith in the parameter values and model structure, they are now judged as even dangerous, because their internal validity criteria are supposed to provide wrong inferences.

(a) The methodology of the estimation-validation experiments

The methodology of Senge's experiments [26] is well known from Monte Carlo studies which have been performed in EC to assess the small sample proporties of various estimators when certain "classical" assumptions of the linear regression model are violated [27,28,29]. Senge used a linear-in-the-parameters version of Forrester's Market Growth Model [30] to produce synthetic data instead of real world data for the parameter estimation. These data can be corrupted by errors in variables to take account of sampling and measurement errors. Furthermore it is possible to study the effects of misspecifications of structural equations and econometric hypothesis about the residuals on the parameter estimates and test statistics.

(b) The main results and some critical comments

Ideal conditions: Using error-free data and the same model specification for parameter estimation that has generated the data base, OLS yields very satisfactory parameter estimates for the Market Growth Model [26]. Columus (2) and (3) and of Table 1 confirm that the OLS-parameters are hardly biased and differ from zero at a 10 %-level of significance.

Errors in variables: acceptance of inaccurate parameters. A 10 % random error of measurement causes inaccurate GLS-estimates for many parameters. Nevertheless Senge's main objection against econometric parameter estimation and testing, which he derives from this experiment, is not directed against the inaccuracies but towards the fact, that all parameters except two (K 15 und K 16, see column 4) are significant. "Hence estimation results relying on statistical significance measures would lead to the acceptance of estimates which are, in fact, quite inaccurate." [26].

There are two problems with Senge's interpretations of his results. The first one is mentioned by himself. "Bias occurs in the estimates... because measurement errors violate a key assumption made in both OLS and GLS estimation. Both techniques assume that the error process is uncorrelated with each of the explanatory variables. This assumption is satisfied in the ideal case, but not when the explanatory variables

Table 1: Parameter Estimates of Forrester's Market Growth Model

(1)	(2)	(3)	(4)	(5)
Parameters	True Values	Ideal Conditions (OLS)	Measurement Errors (GLS)	Misspecification (GLS
K1	475.0	486.8	252.8	199.3
K2	-61.5	-61.13	-31.6	-44.5
K3	-0.6178	-0.6686	-0.2942	-
K4	0.1324	0.144	0.066	-
K5	-0.00975	-0.01066	-0.00586	-
K6	-	-	-	0.133
K8	0.6178	0.6317	0.0876	
K9	-0.1324	-0.1335	-0.0246	
K10	0.00975	0.00962	0.00227	
K11	-1.0	-1.029	-0.1024	Same
K12	0.0003	0.0003288	0.000511	as in
K13	-0.05	-0.0575	-0.098	Run 3
K14	-0.0698	-0.0752	-0.0345	
K15	0.1244	0.1366	0.0364 *	
K16	-0.08138	-0.08983	-0.0133 *	
K17	0.02704	0.02887	0.01049	
Simulation	Run 1	Run 2	Run 3	Run 4

* parameters not significant at a 10 % confidence level

are measured imperfectly." [26]. We think that OLS and GLS should not be blamed for being unable to solve problems they were not designed to tackle. We rather believe that an estimation technique, adequate for errors in variables conditions, should have been chosen (see e.g. [28,29,31]). Peterson [32] has proofed on the same model, that employing an adequate estimation technique renders excellent parameters.

Senge's second critique against the acceptance of inaccurate parameters is even more problematic. Following the usual rule of thumb, a t-statistic greater than 2 implies significance of the parameter and only tells that the estimated parameter certainly differs from zero at the chosen confidence level - it does not give an indication if there is a significant bias between the actual and true value. It would have been preferable, if Senge had also applied a t-test for this latter hypothesis, which is possible under experimental conditions where the true parameter value is known. So far there is no reason, why the estimated parameters should not differ significantly from zero only because they are biased. Looking at Senge's results we would rather come up with an almost contradictory conclusion: the t-values greater than 2 indicate, that the respective variables can be regarded to inhibit a good deal of explanatory power for the dependent variable of the particular equation in spite of biased parameter values due to an inadequate estimation technique.

Acceptance of a misspecified functional form. In another experiment Senge changed the true nonlinear dependence of the delivery rate on the production capacity into a linear function.
True delivery rate equation (see Table 1, Column 2):

$$DR = \left[- 0.6178 \cdot (BL/PC) + 0.1324 \cdot (BL/PC)^2 - 0.00975 \cdot (BL/PC)^3 \right] \cdot PC \qquad (1)$$

Estimated linear equation [1] (see Table 1, Column 5):

$$DR = \underset{(t=2,8)}{0.133} \cdot PC \qquad (2)$$

Although it is true that "the erroneous capacity utilization estimate is statistically significant, thereby giving the model-builder no warning of the consequences of the structural misspecification" [26], we would again like to draw attention to the other and positive side of the same coin: the estimate still confirms that the production capacity is a relevant variable for the explanation of deliveries; non-significance would not necessarily have led to the idea that something might be wrong with the way, PC enters the DR-equation, but could as well have suggested to drop PC at all. On the other hand: the researcher is free to try a non-linear formulation as well as linear one, especially when his a-priori-information raises doubts about a constant capital utilization factor.

Misleading test statistics: where do they lead to? Section 3 started with Forrester's thesis that internal measures of validity might be misleading. At a first glance Senge's experiments seemed to support this statement. But besides the so far exposed objections the question has to be brought up, why the outcome of Senge's econometric estimation of Forrester's Market Growth Model should be hazardous for system dynamicists. Since they are not interested in obtaining as accurate as possible estimates of individual parameters like econometricians are, a bias in parameter estimates is not per se a problem. The bias only becomes critical if it questions the main SD purpose: behaviour mode prediction. Only in this sense would it be justified to speak of "misleading indications from internal measures of validity." Since it is well known that "optimal proporties of the individual estimates of the coefficients are not a necessary prerequisite for the good predictive performance of a model" [29] it should be interesting to supplement Senge's estimation experiments with an analysis of their impact on the behaviour modes - otherwise the heart of the SD-validation theory.

Figures 1 through 4 provide the simulations with the parameter estimates of the above discussed experiments (see again Table 1). The reader may ascertain himself that the general dynamic characteristics are not destroyed in runs 2 to 4 compared to the reference run 1. It is easy to imagine that the correspondence could well be improved with some of the above mentioned refinements. All in all the simulations do not support the negative conclusions, Forrester and Senge have drawn from their investigations.

4. CONCLUSION

We feel that the simulation results do not contradict the statistical validity tests but rather corrobarate them. Furthermore it should not be neglected that unlike in an experimental set-up the true parameter values as well as some aspects of the structural specification are unknown or at least rather uncertain in actual model-building. This assigns econometric procedures an important role within a sound con-

(1) Actually, the delivery rate equation was estimated as a part of the backlog equation. We will not comment on the possible pitfalls of such a procedure in this paper.

cept of multi-stage validation, where different evaluation techniques make independent contributions. Thus there seems to be a good chance that the apparent contradictions between predictive and non-predictive validation might be cleared not by neglecting but rather by accepting the latter. These chances are improved when the non-predictive validation is performed with a careful combination of economic, statistical and econometric criteria [29].

LIST OF SYMBOLS

BL = Backlog

DDRM = Delivery delay recognized by market

DR = Delivery rate

EC = Econometrics

GLS = Generalized least squares estimation

OB = Orders booked

OLS = Ordinary least squares estimation

PC = Production Capacity

S = Salesmen

SD = System Dynamics

SE = Sales effectiveness

REFERENCES

1 Meadows, D.H. 'The Unavoidable A Priori', in J. Randers and L.K. Ervik (eds.), The System Dynamics Method, pp. 161-240. Oslo, 1977

2 Harbordt, S. Computersimulation in den Sozialwissenschaften, Vol. 1. Reinbek, 1974

3 Roberts, E.B. 'On Modelling', Technological Forecasting and Social Change, 1976, pp. 231-238

4 Apel, H. et.al. Ökonomische Aspekte des Umweltproblems. Frankfurt-New York, 1978

5 Gaonkar, R. 'Comparative Analysis of System Dynamics Modeling and Econometric Modeling', in M.H. Hamza (ed.), Proceedings of the International Symposium Simulation '77, pp. 394-400. Anaheim - Calgary - Zürich, 1977

6 Zwicker, E. Simulation und Analyse dynamischer Systeme in den Wirtschafts- und Sozialwissenschaften. Berlin-New York, 1980

7 Chen, K. 'On the Choice and Linkage of Large Scale Forecasting Models', Technological Forecasting and Social Change, 1976, pp. 27-33

8 Meadows, D.L. 'Selecting Among Competing Models of a Specific Social System', in D.L. Meadows (ed.), Methodological Aspects of Social System Simulation, pp. 108-127. Hanover, New Hampshire, 1974

9 Randers, J. 'System Dynamics: A Tool for Broad Policy Analysis', in J. Randers and L.K. Errik (eds.), The System Dynamics Method, pp. 13-30. Oslo, 1977

10 Zahn, E. Probleme bei Simulation sozio-ökonomischer Systeme und die Möglichkeiten zu ihrer Beherrschung. Stuttgart, 1976 (unpublished paper)

11 Hoschka, P. 'Formen und Anwendungen von Modellverknüpfungen und ihre DV-Unterstützung', in B. Schips and H. Schmidt (eds.), Verknüpfung sozio-ökonomischer Modelle, pp. 62-78. Frankfurt-New York, 1980

12 Sommer, M. System Dynamics und Makroökonometrie. Bern-Stuttgart, 1981

13 Forrester, J.W. 'Managerial Decision Making', in M. Greenberger (ed.), Computers and the World of the Future, pp. 36-91. Cambridge, Massachusetts, 1962

14 Forrester, J.W. Industrial Dynamics. New York-London, 1961

15 Coyle, R.G. Management System Dynamics. London-New York-Sydney-Toronto, 1977

16 Dhrymes, P.J. et.al. 'Criteria for Evaluation of Econometric Models', Annals of Economic and Social Measurement, 1972, pp. 291-324

17 Apel, H. Simulation sozioökonomischer Zusammenhänge. Ph.D. dissertation Frankfurt, 1976

18 Forrester, N.B. The Life Cycle of Economic Development. Cambridge, Massachusetts 1973

19 Forrester, J.W. Urban Dynamics. Cambridge, Massachusetts, 1969

20 Meadows, D.L. 'Dynamic Systems Modeling', in N. Hawkes (ed.), International Seminar on Trends in Mathematical modeling, pp. 60-77. Berlin-Heidelberg-New York, 1973

21 Forrester, J.W. 'Industrial Dynamics - A Response to Ansoff and Slevin', Management Science, 1968, pp. 601-618

22 Schleicher, S. 'Erweiterung der Ökonometrie', in G. Bruckmann (ed.), Langfristige Prognosen, pp. 245-254. Würzburg-Wien 1977

23 Lehmann, G. Wirtschaftswachstum im Gleichgewicht. Stuttgart, 1975

24 Forrester, J.W. 'Educational Implications of Responses to System Dynamics Models', in C.W. Churchmann and R.O. Mason (eds.), World Modeling: A Dialogue, pp. 27-35. Amsterdam-Oxford 1976

25 Forrester, J.W. 'A National Model For Understanding Social and Economic Change', Simulation Today, 1975, pp. 125-132

26 Senge, P.M. 'Statistical Estimation of Feedback Models', Simulation, 1977, pp. 177-184

27 Goldberger, A.S. Econometric Theory. New York-London-Sydney, 1964

28 Johnston, J. Econometric Methods. 2nd edn. Tokyo 1972

29 Koutsoyiannis, A. Theory of Econometrics. 2nd edn. London-Basingstoke, 1977

30 Forrester, J.W. 'Market Growth as Influenced by Capital Investment', Industrial Management Review, 1968, pp. 83-105

31 Maddala, G.S. Econometrics. New York, 1977

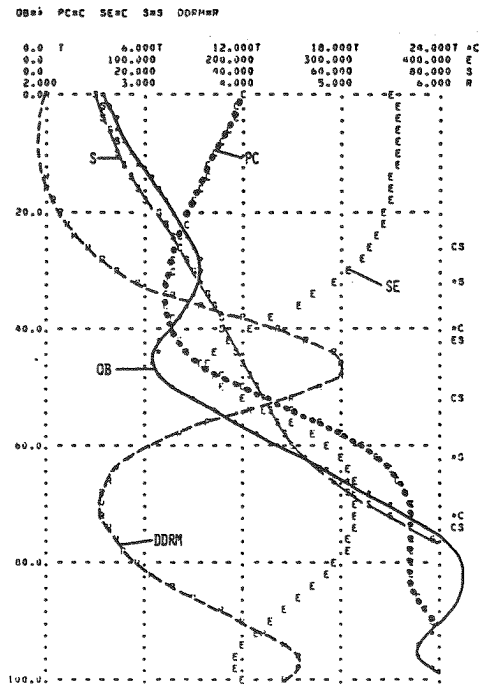

Fig. 1 Reference run

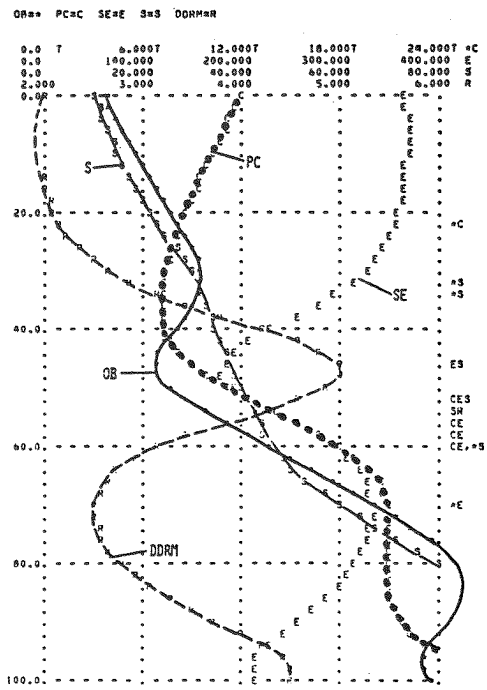

Fig. 2 Ideal Conditions

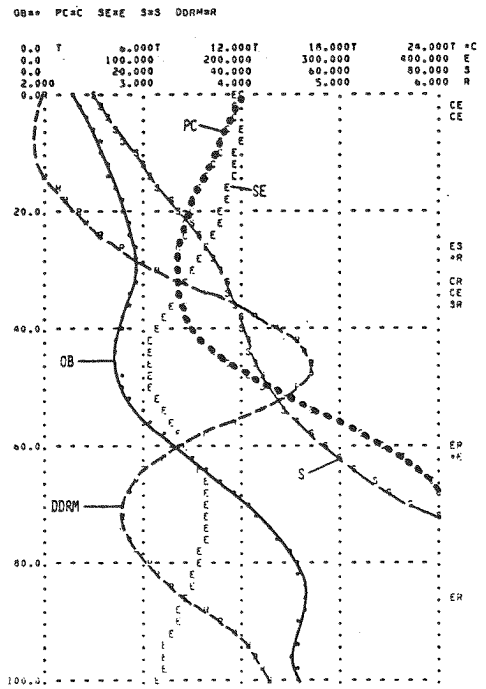

Fig. 3 Errors in variables

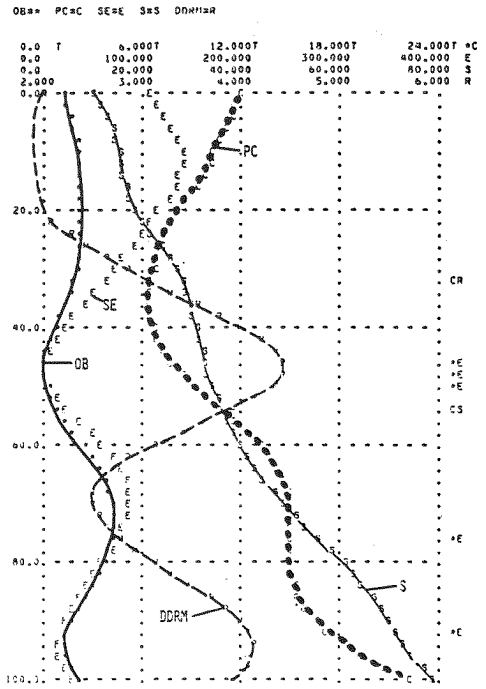

Fig. 4 Misspecification

UKSC 81

Proceedings of the 1981

UKSC CONFERENCE ON COMPUTER SIMULATION

Harrogate May 1981

Westbury House