ANALYTICAL AND EMPIRICAL VALIDATION OF DYNAMIC THERMAL BUILDING MODELS

Allen E & Bloomfield D (Building Research Establishment, UK),
Bowman N & Lomas K (Leicester Polytechnic, UK),
Allen J & Whittle J (Nottingham University, UK),
Irving A (Rutherford Appleton Laboratory, UK)

ABSTRACT - This paper describes the validation methodology adopted by a group of co-operating institutions in the United Kingdom. The need for such a methodology is discussed together with an assessment of the current situation. The role of inter-model comparisons, analytical tests and empirical validation is discussed. A recent comparison between models has attempted to eliminate some of the confusing factors usually accompanying such studies. Substantial differences arising from the basic algorithms are shown to exist. A set of analytical tests supplementing those developed by SERI is described briefly. The problem of **uncertainty** arising from measurement error in the context of empirical validation and the role of sensitivity analyses in quantifying this is discussed.

INTRODUCTION

In a recent review paper(1) it was noted that over 300 techniques for evaluating the thermal performance of buildings were in use in the USA. These encompass a wide range of complexity, from very simple manual methods to extremely large computer programs (eg DOE 2.1). They span not only a wide range in the explicit ability to model details of the physical processes actually occurring, but also in the range of design problems which can be addressed and in the degree of ease with which they can be used. To the user, the most obvious differences between methods are the ones that receive most publicity by their developers - the features that they are claimed to be able to deal with, the user interface and the set-up and computation time involved. This paper addresses another aspect, one that is frequently ignored, or for which unsubstantiated claims are often made - validation. The word itself is often misunderstood and certainly implies different things to different people. It is often interpreted to mean a once and for all check of the absolute accuracy of a program. For the case of thermal performance of buildings the number of parameters, (including such essentially unknown quantities as user behaviour) that may be varied is effectively infinite and it would be quite impossible to test all feasible combinations even if the 'correct' answers were known. Most calculation methods are claimed by their developers to be 'validated'. Usually this takes the form of a comparison of program results against some measured building performance, (eg energy) - 'empirical validation'. There are many problems with such a procedure, some of which are discussed later. It is important to realise at the outset that these 'validation' claims can not be complete. At best, a very limited range of buildings, operating conditions and output quantities can be compared. In actual use for different conditions (eg different climate, more solar gain, heavier weight structure, intermittent heating) the method may still be in error. This paper describes the work of a group of UK Institutions - Building Research Establishment (BRE), Leicester Polytechnic (LP), Nottingham University (NU), and Rutherford Appleton Laboratory (RAL), collaborating in this field. The main aim is to develop a set of verification tests that can be used by both program developers and users to examine the adequacy of any model and its component parts. This should enable more positive guidance to be given on the conditions for which they are adequate, the probable magnitude of errors arising from their use, and the level of modelling detail needed for a specific application.

The paper is divided into three sections dealing first with comparisons between different models. The problems arising from some previous studies are described and the results of a recent exercise designed to avoid them are presented. Secondly, the role that analytical tests can play in establishing the errors associated with individual algorithms is discussed. Finally, the paper deals with the consequences of measurement uncertainties in the context of empirical validation.

INTER-MODEL COMPARISONS

Apart from attempts to compare predictions against measured data, a number of previous exercises have been performed in which the results of different models have been compared, usually for a fictitious building, (inter-model comparisons). A study conducted under the International Energy Agency Annex, Annex I(2) led to the conclusion:

> 'In order to define a building and/or system in sufficient detail such that analysts need make no assumptions about input data, an incredible amount of detail has to be provided, which is not realistic in the design situation. Consequently, differences arising from interpretations of the specification are liable to produce significant differences in predicted energy consumption, irrespective of the quality of the computer program.'

In another study(3) 25 users (22 of whom were consultants) used a single large computer program and the predicted heating energy consumption varied from +106% to -46% of the mean value.

One of the most important contributing factors the inconclusiveness of these results is that documentation and the input / output structure of the models is usually such that subjective judgments have to be made by the user. It is therefore rarely possible to 'validate' a model; more usually it is the combination of model + user + documentation + building description that is assessed.

BRE has carried out an exercise designed to eliminate some of these variables and in so doing to shed light on the variations in predictions of a number of methods currently in use in the UK(4). Both large dynamic simulation models (ESP, SERIRES, TAS) and simpler methods (CIBS admittance, RIBA calculator, BREDEM) were included. For this study, a single user prepared a very detailed building specification (based upon 'typical' UK conditions) and also conducted all the runs. This allowed much greater confidence to be placed in their equivalence, since every possible attempt was made to ensure that each method was solving the same problem, within the restrictions inherent in each method. Fig 1 shows the predicted annual heat loads for a typical masonry attached-house in London, both with and without glass fibre insulation in the roof space. These calculations assume that the boiler is switched off via a time clock between the hours of 2300 and 0700 h, and is under thermostat control for the rest of the time. This is very common practice in the UK. A range in dictions of nearly 2:1 is obtained for the andard' (insulated) house, which for such a well controlled comparison is a matter for concern. Even more worrying is the range of 5:1 obtained for the predicted difference in additional heating requirements for the uninsulated case. Examination of other design alternatives using these models shows that, if a designer were trying to make a decision between, eg insulating the external walls and installing double glazing (on the basis of energy consumption alone), he would be led to a different conclusion by the use of ESP or TAS, as opposed to by the use of other methods. The ranking of design options from this exercise is dependent on the model used and demonstrates the need for careful consideration to be given to validation. It is equally apparent from the description of this exercise so far that there is no possibility of establishing the 'correct' value for heating requirements. The most that can be deduced is that significant errors can be made by using some of these methods.

Past experience has shown that it is not possible to proceed very far in explaining the discrepancies in predictions without understanding the workings of the models at the component level, ie the individual algorithms. The comparison of predictions by different methods does however form a very useful part of a validation methodology provided it is carried out in a well-controlled and a detailed breakdown of gains and losses is obtained. It can help to suggest the algorithms that need to be examined more thoroughly and to identify the presence of errors in different algorithms which cancel each other out and might not otherwise be observed.

The most useful inter-model comparisons result from the use of a set of very simple buildings, designed to progressively introduce more and more complex features in such a way that the point where significant divergence is obtained identifies the algorithm responsible. BRE is currently designing such a set of test data.

ANALYTICAL TESTS

Analytical tests look at the basic heat transfer processes which are common to all thermal models used in the building energy analysis field, namely conduction, convection and radiation. Whilst it is relatively easy to find analytical solutions to many of the simpler heat transfer problems when they are considered in isolation, eg Carlslaw & Jaeger(5), the development of these solutions into tests for the examination of building energy analysis codes is not as straightforward as it might at first appear.

Consideration needs to be given to the difficulties which may be encountered when implementing the tests on real models. Such tests may be used either during model development or retrospectively. Their use during model development would be under the supervision of the model authors and would typically be used for testing a new algorithm or checking that coding changes have been successfully implemented. Their implementation in this situation should not be difficult. Retrospective testing is more often performed by people other than the program authors and it is in this situation that problems are likely to arise. Changes to program coding may be needed and resources for performing the tests may well be limited. Coding changes may be necessary under the following circumstances:

(a) the model may not be able to simulate exactly the conditions described in the test specifications - eg external surface convective heat transfer may be treated as varying with wind speed and direction while the test requires a constant surface coefficient;

(b) the program incorporates checks on the input parameters or derived quantities to prevent excursions outside the 'real world' boundary, eg the maximum solar flux on a surface may be restricted to the value of the solar constant while the test requires larger values in order to increase accuracy;

(c) the required output quantities are not provided by the program. Implementing these changes will be impossible if the program is only available as object code. Where it

is possible to examine the source code the task is still not easy as the documentation provided rarely contains sufficient information; in many cases the programming expertise required to perform such changes is simply not available.

The Solar Energy Research Institute (SERI) has produced a series of analytical tests which can be executed using entire thermal prediction models. They consist of a number of tests in which, effectively, a single wall is subjected to step changes in temperature or heat flow. Both steady state and dynamic response is examined. These have proved successful in identifying major errors in the conduction algorithm. BRE has been using these tests on SERI-RES and has found them useful in providing information on the level of detail necessary in modelling mass walls. The results so far seem to show that greater accuracy in the zone temperature is achieved by cruder (ie fewer nodes) modelling. This will be investigated further. SERI concluded that the satisfactory execution of these tests has not proved to be a sufficient guarantee of the adequacy of the whole model. The BRE group has proposed a further set of tests designed to complement the SERI ones. In these, exact solutions have been used, rather than the approximate ones employed by SERI. A different range of building properties and a wider range of boundary conditions has been explored and the type of input excitation widened to include, eg sinusoidal variations with a 24 hour period. This is, after the steady state performance, the most fundamental type of input in terms of importance to building thermal performance. Ramp inputs are also considered, allowing a better approximation to the variations occurring in real buildings. Tests have also been derived for a two zone system. This allows investigation of an area where many codes make approximations. These tests are currently being applied to the programs ESP (UK) and SERI-RES and their usefulness for more general use will then be assessed.

In addition to the tests so far described which are conducted using the whole model, the BRE/SERC group is devising tests for other algorithms which can be carried out external to the programs (eg long wave radiation exchange, convection, solar processes). In order for this approach to be useful it is essential to have good program documentation and access to the source code. This is often not possible, even at the level of finding out what the basis of the major algorithms are. A questionnaire has been devised to elicit the theoretical basis of a model and has, so far, been completed for the programs ESP, SERI-RES and HTB2 (another UK finite difference model).

A set of tests for internal long wave radiation has been devised in which the effects of different limiting room geometries and surface emissivities are examined. A number of commonly used methods have so far been examined and large errors in the radiation exchange have been found under some circumstances, (eg DEROB, ESP).

Although the tests described so far are quite simple in that one main algorithm is tested, those employing the whole model do simultaneously test many of the program features (including input and output). In real situations, many algorithms are being used at any one time and the interactions between them can be very important.

One of the difficulties with the application of analytical tests for individual algorithms lies with the interpretation of their significance. The experienced program user may feel that he can by intuition alone draw conclusions as to how satisfactory a particular test result is. However, in the authors' view, particularly because of the complicated interactions between algorithms, it is important to devise a rational basis for determining the significance of errors arising from these tests. It is suggested that the following approach should be adopted:

(a) establish the parameters to be treated as fixed inputs for the algorithm;

(b) quantify the range of their values occurring in practice;

(c) determine the exact or most accurate implementation of this algorithm from literature reviews, etc;

(d) determine under what conditions this is applicable by reference to original sources (eg for a convection process, examine the original experimental data upon which convection coefficients are based);

(e) determine the 'best' input-output relationship using (c) within the limits of (d) for the range of building structures and operating conditions of interest;

(f) perform similar tests for the algorithm being examined and compare the results;

(g) establish the significance of the resulting errors by reference to whole model simulations conducted using the most accurate and detailed models available so that the relative significance of, eg an error in the dynamic modelling of a conduction process can be assessed in the light of its overall importance for, eg annual energy consumption.

The use of a set of simplified buildings to investigate the latter is currently being explored.

EMPIRICAL VALIDATION

In principle, empirical validation can provide an absolute test for a model and, unlike analytical verification, it is not limited to simple buildings. The attraction of the technique means that it has

been widely used. The LP team has responsibility for the assessment of this technique and they have recently completed a review which identified over 130 comparisons of actual building performance with predictions made by dynamic thermal models. The fundamental difference between it and other techniques is that it involves experimentation, with all its attendant problems. Such questions as - what, where, how often and how accurately shall I measure? - must be considered.

Judkoff has categorised sources of error as either internal or external. The former arise from inaccuracies in the physical modelling and numerical solution techniques and from coding errors. External error sources arise from the gathering of input data for the model, their transfer to the model and in the measured data.

Much can be learned from a careful examination of the empirical validation undertaken to date, and LP have based their criteria for selecting useful data sets on the experience gained in so doing.

Empirical validation has been performed most often by the model developers and the published studies usually record 'good agreement'. However the accuracy and completeness of the building description and measured data often leaves much to be desired. Parameters to which the model may be quite sensitive are often not measured. Plausible values have therefore to be chosen and, if these do not lead to predictions which match the measured data, new values may be selected. Under such circumstances it is more truthful to state that the program is capable of reproducing observed building performance with appropriately chosen input values, rather than to claim that the model can predict the response of a given building.

The greatest uncertainty is introduced by building occupants. An American study(6) found variations of 40:1 due to occupant effects. Following a review of over 24 studies of occupied buildings, comprising about 100 simulations, performed by a variety of users with 18 models, each using anything from 1 to 243 buildings, Wagner concluded that:

> 'the availability of accurate and sufficiently complete input data, especially on occupant behaviour, limits the ability of even detailed models to accurately predict energy use, in some cases severely so'.

The magnitude of the user effect can easily be understood if one thinks of such user action as opening windows, adjusting thermostat settings, altering the position of shading devices, etc.

Perhaps the second largest problem found with existing datasets is the total absence of some important measurement, eg air infiltration. It is hard to make a reasonable estimate of such parameters, so that claims about the predictive accuracy of models based upon comparison with such datasets should be treated with great scepticism. The effect of other missing data, eg climatic data, depends both on the parameter in question and on the structure under consideration. It should be assessed for each case individually.

Apart from missing data, the use of standard data for, eg material properties has been shown to give rise to large errors. For a single storey ranch house in Colorado errors in auxiliary load prediction of approximately 60% due to incorrect wall conductance values, were reported in (7).

The main conclusions drawn by the LP group from their review of past work are:

(a) numerous sources of error may exist in the data input to models; these propagate through the model, leading to uncertainty in the predicted values;

(b) the presence of external errors means that, in most investigations to date, no conclusive evidence of internal errors can be produced;

(c) the absence of a clear methodology has led many empirical investigations into difficulties, eg inadequate and inaccurate data has been used for input variables; the building selected has not been suitable;

(d) more thought needs to be directed towards what parameters should be both measured and predicted and to how the comparison should be made;

(e) only the highest quality building construction and data gathering techniques can hope to produce conclusive evidence of internal errors;

(f) it is difficult, expensive and time-consuming to obtain the high quality data needed for validating models.

Comparison against experimental data for the full range of buildings, climates and other boundary conditions that could be obtained in practice is clearly impractical. Cohen(8) has suggestsd that, instead, a model's predictions should be tested against a statistically significant set of buildings and climates. In view of (f) above, it is extremely unlikely that it will be possible to apply sufficient tests to achieve complete statistical significance. The LP group are conducting an examination of extant data sets throughout the world in order to select a set of high quality datasets that could be useful in testing the principal algorithms currently employed in dynamic simulation models. The philosophy adopted in selecting datasets is similar to that employed with the other validation techniques, ie to devise tests which progress from simple to more complex situations. In view of the conclusions cited above, this argues, in the first

instance, for the selection of datasets for unoccupied buildings with full, Class A monitoring with construction and operational features which can be explicitly modelled. A full physical description of the building and of its operation and control schedule must exist. Climatic data must have been obtained at the site and measurements must have been taken over short time intervals, (one hour or less). These criteria have so far been applied to some 200 datasets and the results will be published shortly.

Insufficient attention has been paid to date to the assessment of the adequacy of experimental datasets for the purposes of validation. In addition to eliminating datasets which do not contain measurements for all relevant parameters, where user effects are too large or where the building is insufficiently described, consideration should also be given to the consequences of inaccuracies due to the measuring process itself.

This can be explored by conducting sensitivity analyses using one or more models. Sensitivity studies have usually been used to investigate the consequences of variations in the value of one single parameter on some chosen output variable, eg heat/energy consumption, peak temperature, etc, (differential sensitivity analysis). In practice many input parameters will be subject to some uncertainty and it would be more realistic to investigate the consequent range of possible values of the output parameter of interest by allowing variations in all the input parameters simultaneously and choosing coincident values by reference to their probability distributions, (stochastic sensitivity analysis). It is important to clearly specify the purpose of the study in advance and select an appropriate output parameter, realistic input perturbations and a statistical measure for agreement between perturbed and unperturbed outputs. In one study conducted by RAL the standard deviations of the thermal properties of the walls of a test cell were estimated as: conductivity 3%, density 4%, specific heat 3% of the mean value. The effect on predicted air temperature for a 3 month winter period was estimated for fluctuations by four standard deviations in these parameters, varied both singly and simultaneously. Illustrative results are plotted in Fig 2 for one day. Two confidence intervals, corresponding to a probability of 20:1 that the temperature would lie within this bound, have been calculated. One (3.5C) has been derived from an approximate method of combining the results of the single parameter variation simulations in quadrature. This assumes, inter alia, that the variations in each input parameter are uncorrelated. The second confidence interval (2.0C) has been derived from a more detailed stochastic method which has employed a Monte-Carlo method to allow the generation of simultaneous variations in the three material properties. For this simple example, it can be seen that by using values of material properties in which we are 95% certain, we can only be sure of the predicted air temperatures to within approximately 3C (5F). These studies are only at a very preliminary stage and the procedures for conducting and interpreting particularly the stochastic sensitivity tests need to be developed further. The results so far do, however, illustrate that when measured values are subject to some error, their use as fixed input values to a simulation program can give rise to uncertainties in the output variable(s), considerably in excess of its measurement error. Great care should therefore be exercised in claiming that comparisons of predicted with measured data have established validity.

In addition to the above way of using sensitivity analyses to assess whether a measured data set is suitable for use in validation, they can also be of great value in identifying situations where a particular modelling assumption is of importance. This can influence the selection of suitable datasets. In one study conducted by Leicester Polytechnic a test cell typical of many existing facilities, eg Los Alamos, was studied. The sensitivity of predicted heating load and air temperature to the distribution of shortwave radiation amongst the internal surfaces was tested using a program capable of detailed solar mapping calculations. For the situation where a significantly heavier weight construction was used for the back wall (as is found in the Peterborough test cell currently being monitored for the UK Department of Energy) a maximum difference in peak temperature of 8C (ie 28C compared to 20C) was predicted when the solar was directed to the floor instead of the back wall. This suggests that if a program which employs simple fixed assumptions about the distribution of solar radiation is to be tested using this dataset, this effect would swamp many other errors. It would therefore be advisable to select an alternative dataset, with similar weight walls, so that the correctness of other algorithms could be checked first.

Sensitivity analyses can, of course, be extremely valuable in assessing the adequacy of pure modelling assumptions. Another study conducted on the same test cell investigated the number of timesteps per hour required for the program ESP (a large UK program) to produce consistent results. For this fairly lightweight cell a difference in predicted peak air temperature of 2C was found between the 1 and 4 timesteps/hour simulations. Virtually no difference was observed as the number of timesteps was increased further.

CONCLUSIONS

This paper has presented a summary of the approach to validation adopted by a group of co-operating UK institutions co-ordinated by BRE.

The main findings to date are given here:

1. A review of previous validation work on methods for predicting the thermal performance of buildings has shown a concentration on the comparison of measured and predicted performance - empirical validation. This technique can do no more than test a very small subset of possible situations in which a user may wish to employ such models, even then only if comprehensive and reliable data is available.

There is, therefore, a need to employ other techniques in order to extend the range of situations covered.

2. Errors can arise for a large number of reasons. They can be categorised as either 'internal', due to the modelling process itself, or 'external' which are more under user control. Comparisons between models have also been conducted quite frequently in the past and have usually suffered from a number of limitations, mainly arising from inadequacies in the specification of the problem, compounded by the differing interpretations made by different users. A model can not easily be tested, only the combination of user and model, so that the word 'validation' ceases to have quite the absolute connotation usually associated with it.

3. Some of the results of an inter-model comparison study designed to eliminate the usual sources of confusion, are reported and substantial differences in predicted results due to internal errors are found. Inter-model comparisons can form a useful part of a set of testing techniques, provided that care is taken in their design.

4. The usefulness of 'analytical tests' (ie simple input excitations for which exact results can be calculated) is discussed and a distinction drawn between those that can be carried out on an entire 'whole' model and those that can only be carried out on a specific algorithm separately from the program itself. It is concluded that both types are of value and preliminary results are presented. Although useful, these tests do not at the moment give sufficient attention to the significance of the reported errors in terms of the everyday application of models. This is an area which needs to be developed and the approach to be adopted by the BRE/SERC group is described.

5. Empirical validation is seen as, perhaps, the obvious test that a user would wish to see performed in order to convince him of the usefulness of the model he is using. What is not normally given sufficient attention is the fact that the measured values themselves do not represent 'truth'. They too contain uncertainties due to measurement inaccuracy. In order to predict energy consumption, models have to make many assumptions about the probable value of 'input' variables, eg material properties, thermostat set points. If the uncertainties in these input variables are propagated through a model, a large uncertainty in the predicted energy consumption results, due solely to this cause. The results from a simulation model should not therefore be seen as purely deterministic, rather one should be speaking of the probability of the energy consumption lying within a certain confidence interval. Comparisons of predicted and measured results should properly be seen in this light.

6. Techniques for allowing this to be done are being developed and preliminary results shown here demonstrate that for a typical test cell, the 95% confidence interval for internal air temperature resulting from uncertainties in material properties alone can be as high as 3 C (5F). This is far in excess of the measurement error associated with air temperature itself.

7. A literature survey has been conducted and has so far identified some 200 datasets potentially suitable for empirical validation. A set of simple selection criteria to establish their usefulness has been prepared and, together with sensitivity analyses, will be applied to these datasets.

REFERENCES

1. KUSUDA, T. Review of current calculation procedures for building energy analysis. US Nat. Bureau of Standards, Rep. NBSIR80-2068, for the US DoE, 60 pp, (1980).

2. Comparison of load determination methodologies for building energy analysis programs, International Energy Agency Report,(December 1979).

3. JONES, L. The analyst as a factor in the prediction of energy consumption, Proc. 2nd. Int. CIB Symposium on Energy Conservation in the built environment, Danish Building Res. Inst., Copenhagen, pp 313-321, (1979).

4. ALLEN, E & BLOOMFIELD, D. Improving confidence in thermal calculation procedures, Proc. CLIMA 2000, Copenhagen, (1985).

5. CARLSLAW, H & JAEGER, J. Conduction of heat in solids, OUP, (1959).

6. WAGNER, B. Comparison of predicted and measured energy use in occupied buildings, ASHRAE Trans., 90,2, (1984).

7. JUDKOFF, R, et al. Empirical validation using data from the SERI Class 'A' Validation House, Proc. Annual meeting of American Section of ISES, Minneapolis, Mn, USA, 6, pp 705-710, (1983).

8. COHEN, T. Statistical problems in design technique validation, SERI Report RR-721-377, (1980).

ACKNOWLEDGEMENTS

The work has been conducted by a group of institutions as described in the paper. The BRE contribution forms part of the research programme of the Building Research Establishment of the Department of the Environment and this paper is published by permission of the Director.

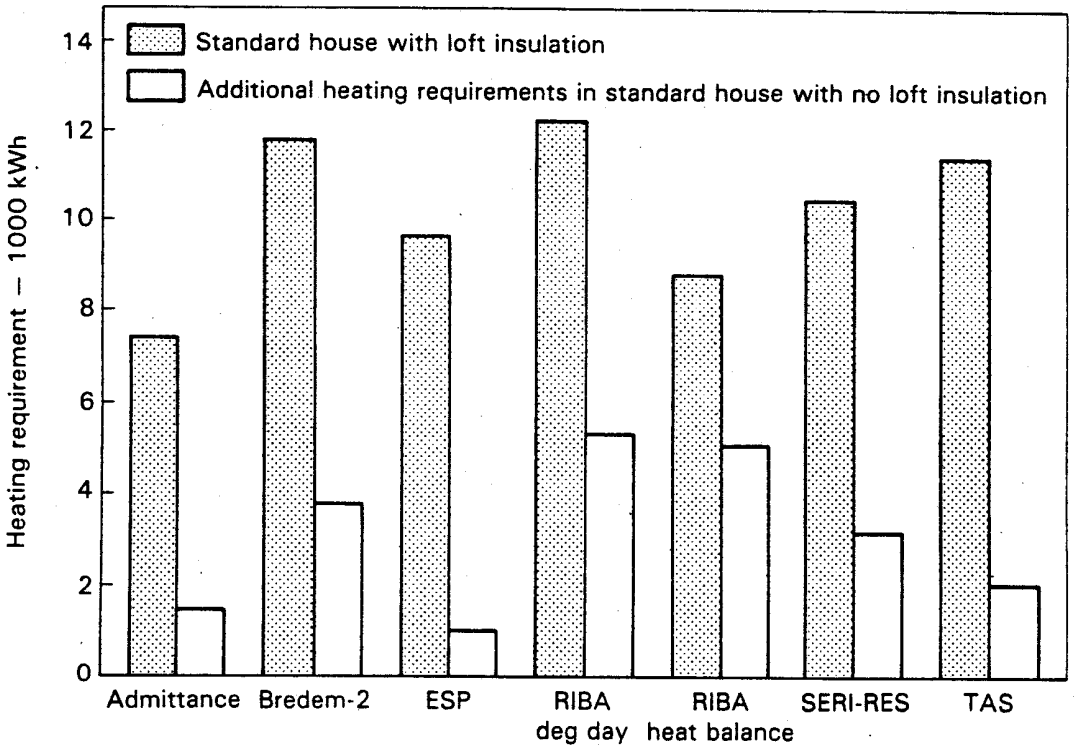

Fig. 1 Predicted annual heating requirements

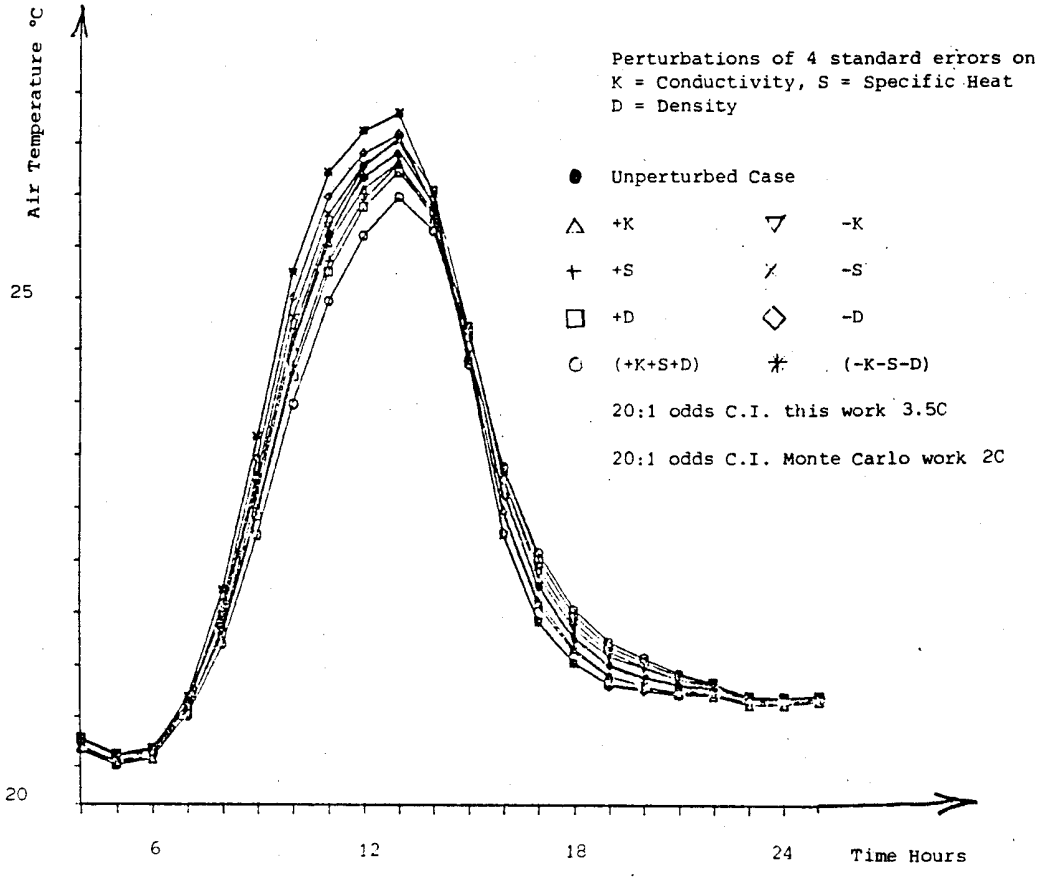

Fig. 2 Variation in air temperature for a typical day. Peterborough Test Cell modelled by, ESP. Timestep = 7.5 minutes. Kew, October-December 1967.