ELSEVIER

Energy and Buildings 26 (1997) 253–275

ENERGY AND BUILDINGS

Empirical validation of building energy simulation programs

K.J. Lomas [a,*], H. Eppel [a], C.J. Martin [b], D.P. Bloomfield [c]

[a] *Institute of Energy and Sustainable Development, De Montfort University, Leicester, LE1 9BH, UK*
[b] *Energy Monitoring Company Ltd., 3 Chapel Court, Newport Pagnell, MK16 0EW, UK*
[c] *Building Research Establishment, Garston, Watford, WD2 7JR, UK*

Received 3 January 1996; revised 5 January 1997

Abstract

The largest-ever exercise to validate dynamic thermal simulation programs (DSPs) of buildings has recently been completed. It involved 25 program/user combinations from Europe, the USA and Australia, and included both commercial and public domain programs. Predictions were produced for three single-zone test rooms in the UK. These had either a single-glazed or double-glazed south-facing window, or no window at all. In one 10-day period the rooms were intermittently heated and in another 10-day period they were unheated. The predictions of heating energy demands and air temperatures were compared. The observed interprogram variability was highly likely to be due to inherent differences between the DSPs, rather than the way they were used. Predictions of the difference in performance of two rooms were no more consistent than predictions of the absolute performance of a single room. By comparing the predictions with the measurements and taking due account of experimental uncertainty, the DSPs that are likely to contain significant internal errors are distinguished from those which, in these tests, performed much better. The likely sources of internal error are discussed. It is recommended that empirical validation exercises should consist of an initial blind phase in which program users are unaware of the actual measured performance of the building, and then an open phase in which the measurements are made available. The work has produced five empirical validation benchmarks, which have significant practical benefits for program users, vendors and potential purchasers. There is considerable scope for improving the predictive ability of DSPs and so suggestions for further work are made. © 1997 Building Research Establishment. Published by Elsevier Science S.A.

Keywords: Dynamic thermal simulation programs; Heating energy demands; Air temperatures; Predicted values; Measured values; Empirical validation exercises

1. Introduction

The need for reliable techniques to predict the thermal performance of buildings was prompted by the oil embargo of the mid-70s. More recently, concerns about the global environment and the quality of the environment within buildings have sustained this need. These factors prompted the development of dynamic simulation programs (DSPs) of buildings. These programs seek to predict the time-varying energy demands, internal temperatures and heat fluxes in complex multi-zone buildings subject to real weather and operating conditions. DSPs are now extremely powerful and affordable, and are becoming widely used to assist in the design of new or refurbished buildings.

Two issues which influence the extent to which DSPs are used concern their validity and usability. These aspects are closely related and both were studied in International Energy Agency (IEA) Energy Conservation in Buildings and Community Systems (BCS) Annex 21. Usability was addressed in Subtask A (program documentation), Subtask B (formalizing and documenting building performance assessment methods) and Subtask D (design support environments).

This paper focuses on the empirical validation work undertaken by a group formed by combining: IEA BCS Annex 21, Subtask C; and IEA Solar Heating and Cooling (SHC) Task 12, Subtask B. The work was directed by the UK Building Research Establishment (BRE) and managed by the Institute of Energy and Sustainable Development at De Montfort University (DMU), UK, and the Energy Monitoring Company (EMC), UK. It began in November 1989 and was completed with the production of the final report [1] in September, 1994.

The empirical validation work complemented the other validation activities of the IEA Annex 21/Task Group 12. These activities resulted in the production of a set of Building Energy Simulation Tests (BESTEST), based on intermodel comparisons [2,3]. These tests centred on domestic-scale buildings and were structured so that reasons for poor pro-

[*] Corresponding author.

0378-7788/97/$17.00 © 1997 Building Research Establishment. Published by Elsevier Science S.A.
PII S0378-7788(97)00007-8

gram predictions could be diagnosed. Other tests based on intermodel comparisons relate to commercial buildings [4]. The predictions obtained in the empirical validation work are compared with results obtained in these other (intermodel comparisons) exercises later in this paper.

At present, few properly documented whole program validation benchmarks exist [5] and even fewer have been tested on a wide range of programs. Given this situation, it was decided that the aims of the empirical validation work should be to:

(a) develop well-documented and well-tested empirical validation benchmarks for DSPs;

(b) provide a 'snap-shot' of the ability of DSPs to predict the performance of a few simple buildings under conditions reflecting those which exist when they are used to model real buildings; and

(c) devise and test a strategy for developing empirical validation benchmarks.

It was not feasible to try to discover why programs performed well or why they performed badly—other tests e.g. analytical tests [6], sensitivity analysis [7], algorithmic substitution, or intermodel diagnostic tests [2,3], can do this better. However, some general observations pertinent to this matter are made.

This paper describes the selection of the data sets, the organization and management of the empirical validation exercise, the participating programs and their users, and the predictions obtained. The interprogram variations are quantified and compared with the variations obtained in the complementary BESTEST [2,3] and commercial building [4] studies. Comparisons between the predictions of the DSPs and the field measurements are undertaken taking due account of the uncertainties in the test room descriptions and the measurements. The way in which these results should be interpreted is discussed and future validation needs are outlined.

The work discussed in this paper, and the interpretation o it, reflects the views of the participants in IEA Annex 21, Task 12. Their final report [1] was approved by the Executiv Committees which together represent some 20 countries.

2. Selection of data sets

It is extremely difficult to collect data sets of sufficientl high quality that they can be used for empirical validation The vast majority of previous empirical validation exercise have been unsuccessful because the measured data contained a few critical, but easily identifiable, limitations [8]. In recognition of this, the IEA group felt it important to conduct a thorough review and assessment of available data sets prior to making a commitment to undertake empirical validation. The following recommendations were laid down to guide the review.

(i) The data set(s) must fulfil nine criteria which define high quality data (Table 1);

(ii) the data must be available for use both within the IEA project and for subsequent use by others;

(iii) ideally, the site from which the data were collected should be active;

(iv) the actual measured performance of the buildings should not be widely known so that programs could be tested 'blind', i.e. without program users knowing what the measured performance of the building was (see Section 5).

A previous world-wide survey [8] provided the basis for the review. However, to ensure that this survey had not overlooked any data sets, and to capture any recently collected data, a questionnaire was distributed to principal researchers in the validation field. Only one additional facility capable of producing high quality data was revealed [9], but these data were not immediately available.

Therefore, the 72 data sets identified by the previous survey as being both available and of high quality, were examined

Table 1
Criteria for classifying data sets (based on Lomas [8])

A: Preliminary acceptance criteria which data sets must fulfill to be of value for validing any dynamic thermal program. Data sets that pass all three criteria are termed 'Acceptable Data Sets'.

Criterion 1 Structures must not include operative active solar space heating or cooling systems.

Criterion 2 The weather data must have been collected at the site of the building.

Criterion 3 The measured building performance data, and the weather data, must be available at hourly, or more frequent intervals.

B. Data sets which fulfill three additional criteria are termed 'Useful Data Sets'.

Criterion 4 All three major elements of the weather, air temperature, wind speed, and the direct and diffuse components of solar radiation, must be measured at the site of the building for the whole comparison period.

Criterion 5 The structure must be unoccupied, it must not contain passive solar features which cannot be explicitly modelled and each zone in the building must have independent heating and/or cooling plant and controls.

Criterion 6 Measured infiltration and, where appropriate, interzonal air flow rates, must be available for the whole comparison period.

C. Data sets which also pass three further criteria have been termed 'High Quality Data Set'.

Criterion 7 The structure must not contain features, or environmental control systems, which cannot be modelled explicitly by any of the programs being validated.

Criterion 8 The data medium must be of a type which is readily usable, and close liaison with the monitoring institution must be possible.

Criterion 9 Data for sites which have never produced data for model validation work, or data which, due to external errors, has introduced unacceptable uncertainty into previous validation work, must not be included.

further. These data were from: the Polytechnic of Central London test cells in Peterborough, UK [10]; the British Gas cell in Cranfield, UK [11–14]; the PASSYS cells in Glasgow, UK [15,16]; the Energy Monitoring Company (EMC) test rooms [17–24]; and the National Bureau of Standards (now the National Institute for Science and Technology) passive solar test facility in Gaithersburg near Washington, DC, USA [25,26].

Except for data from the EMC test rooms, all these data sets failed to meet one or other of the above recommendations. The EMC rooms had produced 48 high quality data sets, so a diverse range of validation tests could, in principle, be developed. After further close scrutiny, the IEA participants felt that empirical validation should be undertaken using some of the data from these rooms.

The review highlighted three important points.

(a) Whilst data sets may be classified as high quality by the nine criteria in Table 1, they may still be unsuitable for a particular empirical validation exercise. These criteria should therefore be seen as the minimum requirements for data which are to be used for validating DSPs.

(b) There is a need to collect and archive further very high quality data sets for validating a wide range of thermal programs.

(c) Very little research is being undertaken to assess the absolute accuracy of thermal programs of buildings.

Although some high quality data sets have been produced since the review, e.g. PASSYS cells, the broad conclusions from this review are still valid.

3. Description of the data sets

The EMC test rooms occupy an unobstructed site at Cranfield airfield. They were originally built to be a good compromise between the needs for realism and experimental accuracy. The rooms (Fig. 1) were built in pairs, separated by a heavily insulated party wall. The outer shells were of stud-frame construction covered by plasterboard, and they had concrete slabs on the floor (Figs. 2 and 3). The monitored spaces were well insulated and extremely well sealed to reduce infiltration to less than 0.05 air changes per hour. A roof space was less well insulated and poorly sealed so infiltration occurred. A well-insulated ceiling limited the heat flow between this space and the monitored room below.

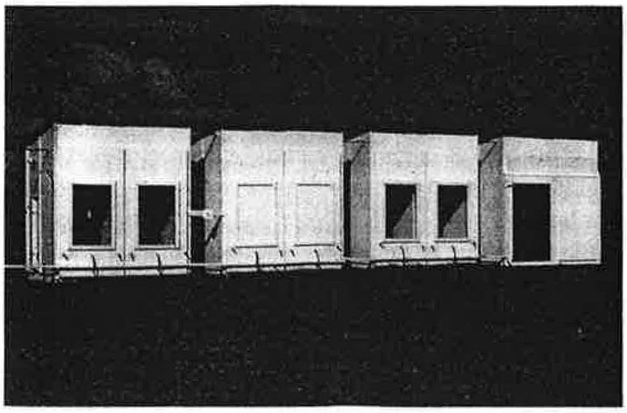
Fig. 1. External view of the test rooms.

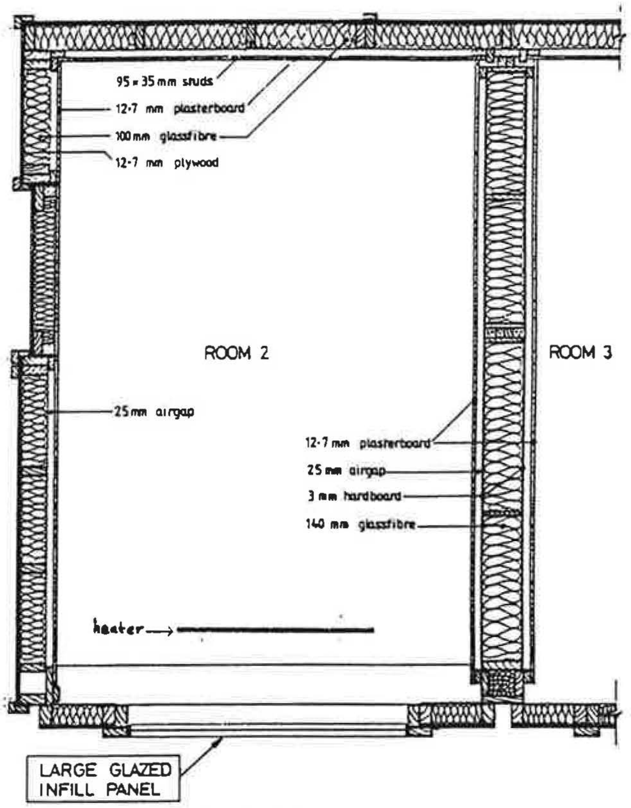

Fig. 2. Plan view of a test room showing construction.

The IEA work focused on three rooms (Rooms 1, 3 and 5) and each was monitored for two 10-day periods (Table 2). In the first period the rooms were intermittently heated by an oil-filled panel radiator which produced both radiant and convective heat and had a nominal peak power output of 680 W (see Fig. 2), in the second period the rooms were unheated (free-floating). In both periods, Room 3 housed an opaque panel in the south-facing front wall whereas the other two rooms had (different) glazed façades (Table 2).

The rooms are representative of typical lightweight rooms in UK houses in terms of the level of insulation, the amount of thermal mass and the window-to-floor area ratio. They stress the solar gain and fabric heat loss processes because they have very low infiltration rates, a large surface area-to-volume ratio, and no incidental internal heat gains.

The climate data typically required by DSPs were collected (Figs. 4 and 5) and the thermo-physical properties of the construction materials were defined. The key parameters measured in each room were the hourly heating energy consumption and the air temperatures at three levels. Also recorded were the total hourly external south-facing vertical solar irradiance; the temperatures of the inside surfaces of the floor, back wall and ceiling; and the roof space and floor void temperatures. Data acquisition is described more fully elsewhere [1].

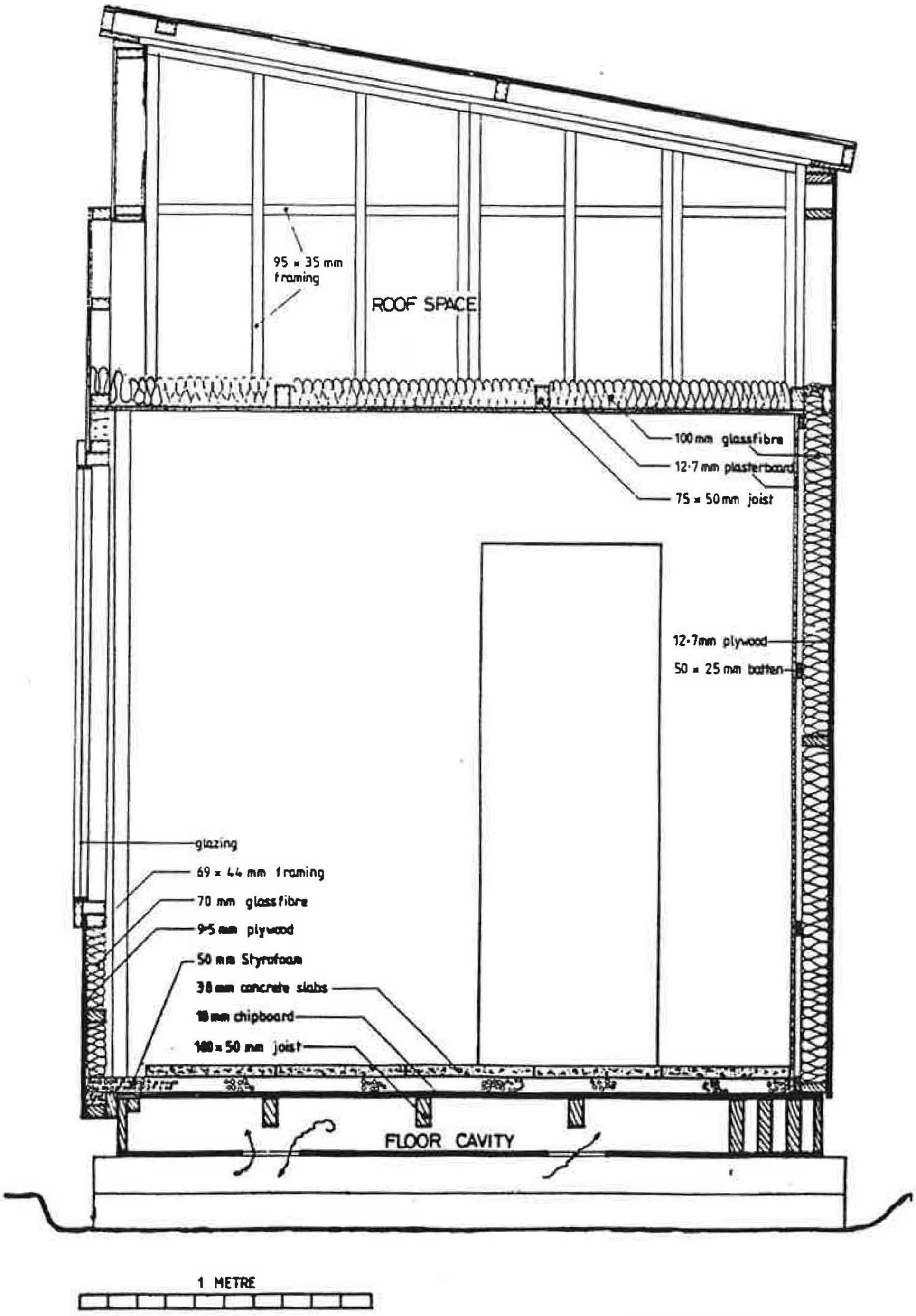

Fig. 3. Section through a test room showing construction (heater omitted).

4. Description of participants and programs

Initially, the only modellers involved were six of the IEA Annex 21/Task 12 participants who between them planned to run nine DSPs. However, it was desirable to gauge the performance of as many programs as possible so others were invited to participate. This attracted an additional 14 institutions and private companies (Table 3). Most were either skilled users or the authors, vendors or support offices for the programs.

In total, 25 results sets were eventually obtained from 17 genuinely different programs (the remaining results were from alternative versions of some of these); one program, WG6TC, was only applicable to temperature predictions in the unheated rooms. There were 13 commercial programs, 10 research programs and 2 programs which were still being

Table 2
Synoptic description of the data from the EMC test rooms

Prediction period	Room	Glazing type	Glazing area	Heating	Set point
May 24–30, 1990	1	Double	1.5 m^2	None	n/a
	3	Opaque		None	n/a
	5	Single	1.5 m^2	None	n/a
October 20–26, 1987	1	Double	1.5 m^2	06:00–18:00	30°C
	3	Opaque		06:00–18:00	30°C
	5	Double [a]	0.75 m^2	06:00–18:00	30°C

[a] Predictions made for 1.5 m^2 of single glazing.

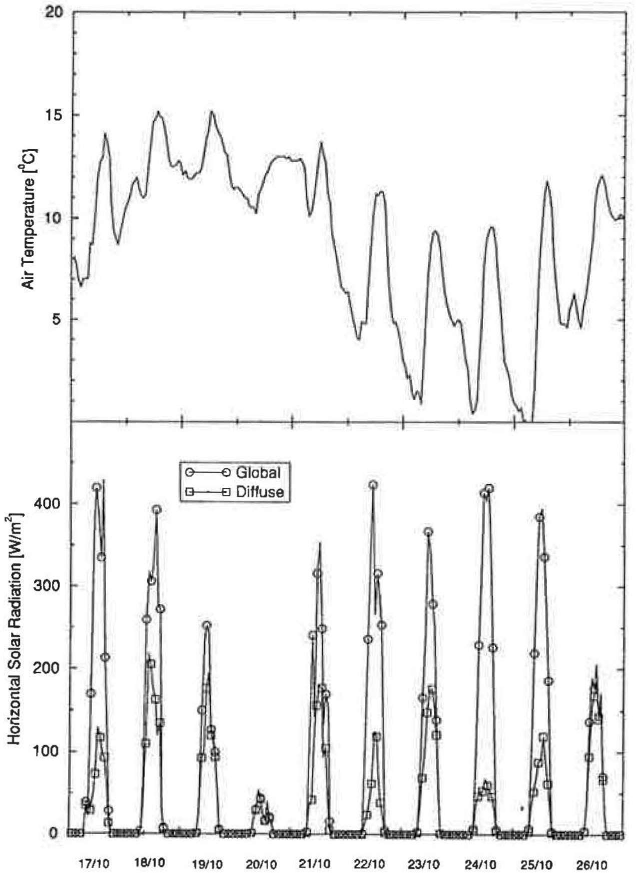

Fig. 4. Recorded climate data for the October period (wind speed and direction not shown, relative humidity not measured).

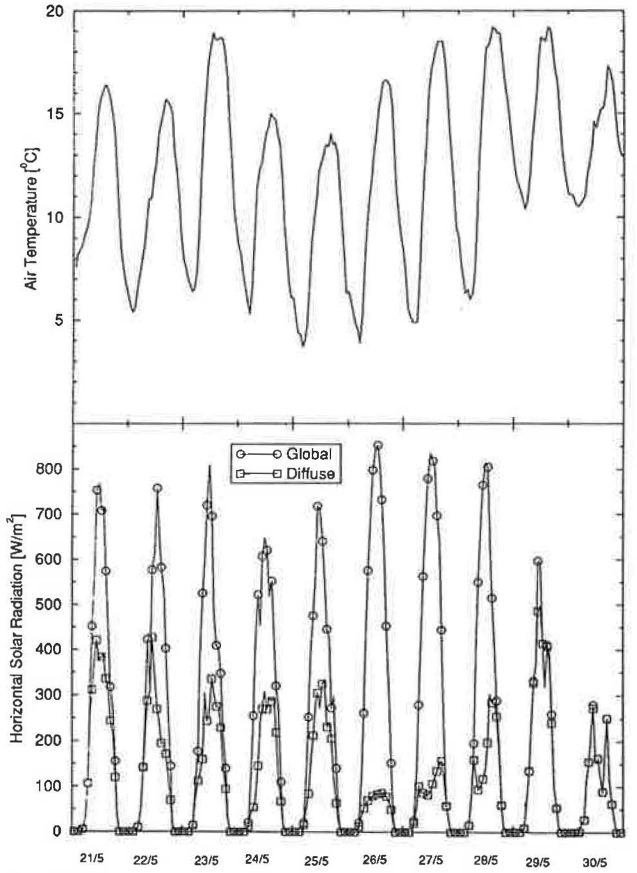

Fig. 5. Recorded climate data for the May period (wind speed and direction not shown, relative humidity not shown).

developed. The programs originated in 6 different European countries, USA and Australia. A program description questionnaire indicated that they employed a diverse range of algorithms for the key heat transfer processes (Table 4).

This was the most extensive empirical exercise ever undertaken, and it was encouraging to see so many program users and authors willing to participate. All the participants were free to discontinue their involvement at any time, but, having begun, and seen the prediction task being set for them, they all chose to continue. None of the participants expressed concern that the task was inappropriate to their program, and none of the participants objected to their results being published.

5. Operation of the validation exercise

5.1. Project management and phasing

A review of the work conducted in IEA Task VIII [27] led to important messages about how to conduct empirical validation.

(I) Strong centralized project management should be responsible for:

(i) ensuring that the agreed methodology and time-scales are followed;

(ii) interfacing between the data collection team and the modellers to ensure that the same information is available to

Table 3
List of programs and participants

Program	Version	Country of origin	Operating Institution
APACHE	6.5.2	UK	Facet Ltd., UK [a]
BLAST	3lvl143	USA	Colorado State University (CSU), USA [a]
BLAST	3.0lvl203 [d]	USA	Politecnico di Torino, Italy [b]
CHEETAH	1.2	Australia	CSIRO, Australia [a]
CLIM2000	1.1	France	Electricité de France (EdF) [a]
DEROB	LTH	USA	Lund Institute of Technology, Sweden [b]
DOE	2.1E	USA	Lawrence Berkeley Laboratory (LBL), USA [a]
ENERGY2	1.0	UK	Arup R&D, UK [a]
ESP	6.18a	UK	De Montfort University (DMU), Leicester, UK [b]
ESP-r	7.7a	UK	Energy Simulation Research Unit (ESRU), Univ. of Strathclyde, UK [a]
ESP+	2.1	UK	DMU Leicester [c]/ASL Sterling [a], UK
HTB2	1.2	UK	Fachhochschule für Technik (FHT), Stuttgart, Germany [b]
HTB2	1.10	UK	University of Wales College of Cardiff (UWCC), UK [a]
SERI-RES	1.2	USA	Building Research Establishment (BRE), UK [b]
SUNCODE	5.7	USA	Ecotope, USA [a]
S3PAS	2.0	Spain	Escuela Superiore Ingenieros Industriales, Sevilla, Spain [a]
TASE	3.0	Finland	Tampere University of Technology, Finland [a]
TAS	7.54	UK	DMU Leicester [c]/Environmental Design Solutions Ltd (EDSL) [a], UK
TRNSYS	13.1	USA	University of Wisconsin, Madison, USA [a]
TRNSYS	12	USA	BRE, UK [b]
TRNSYS	13	USA	BRE, UK [b]
TRNSYS	13.1	USA	Vrije Universiteit Brussel, Belgium [b]
TSBI3	2.0	Denmark	Danish Building Research Institute (SBI) [a]
WG6TC	1992	Italy	Institute di Fisica Technica, Udine, Italy [a]
3TC	1.0	UK	Facet Ltd., UK [a]

[a] Program authors/vendors/support office.
[b] Experienced users of program.
[c] DMU ran program on behalf of vendors, input files checked by vendors in Phase 1, vendors ran programs in Phase 2.
[d] 3.0lvl193 used in Phase 1.

all modellers and that this information is accurate and consistent; and

(iii) analysing the results.

(II) The initial predictions should be made 'blind', that is, program users should not be aware of the actual measurements. This mimics the conditions which prevail when DSPs are used in a real building design context.

These messages led to a work programme that had two distinct phases, a blind phase and an open phase.

5.2. The blind phase

In Phase 1, all the predictions were made blind, i.e. without any knowledge of what the actual performance of the EMC test rooms was. A detailed 'Empirical Validation Package' was distributed to each participant that consisted of a site handbook and a validation guidebook.

The site handbook gave a full description of the test rooms, and the nominal values for all site data, geometrical parameters (volumes, surface areas etc.), thermo-physical properties (density, specific heat capacity, conductivity), surface properties, glazing transmission properties, infiltration rates, and heater and control characteristics. As far as possible, the nominal geometrical and thermo-physical properties were either measured directly by the EMC, or calculated by the EMC, or based on measured values supplied by material manufacturers. Some nominal values had to be either estimated or standard (handbook) values had to be used. The program users were expected to use these nominal values in their simulations.

It was the policy of the exercise not to supply ad-hoc derived parameters (e.g. U-values, surface heat transfer coefficients). Rather, the program users had to choose appropriate values as necessary for their program. This would be the situation if the program were used in a real building design context. The uncertainty associated with these program-specific parameters was accounted for by sensitivity analyses.

The only important deviation from this policy was that an estimate was made of the extra heat loss through edge and corner constructions. This was done because all the participating programs assumed that heat flow through surfaces was one-dimensional. In reality this is not so, and in the test rooms, one-dimensional heat flow assumptions would cause an underestimate of the total heat loss coefficient (W/K) of about 10%. To relieve participants of the labour of accounting for these the extra losses, they were calculated and the conductivity of some wooden elements was amended (increased) to account for them.

Rigorous quality assurance procedures were adopted by the EMC (see Ref. [1]) to ensure, to within the defined experimental errors, that the rooms conformed in all significant respects to the descriptions provided to the program users.

Table 4
Features of the participating programs

		ESPv6.18a (DMU, UK)	ESP-Rv7.7a (ESRU, UK)	ESP+v2.1 (DMU, UK)	SERI-RESv1.2 (BRE, UK)	SUNCODEv5.7 (Ecotope, US)	TRNSYSv12 (BRE, UK)	TRNSYSv13 (BRE, UK)	TRNSYSv13.1 (Brussel, B)	TRNSYSv13.1 (UWISC, US)	TASEv3.0 (Tampere, FIN)	BLASTv3.0 (Torino, I)	BLASTv3Ml143 (CSU, US)	S3PASv2.0 (Sevilla, E)	DEROBvth (Lund, S)	CLIM2000v1.1 (EDF, F)	HTB2v1.2 (FHT, GER)	HTB2v1.10 (UWCC, UK)	APACHEv6.5.2 (Facet, UK)	3TCv1.0 (Facet, UK)	CHEETAHv15.2 (CSIRO, AUS)	ENERGY2v1.0 (Arup, UK)	TASv7.54 (DMU, UK)	DOE2.1E (LBL, US)	TSBI3v2.0 (DBRI, DK)	WG6TCv1992 (Udine, I)
Program type	Public domain	x	x		x						x			x	x	x	x	x							x	
	Commercial			x		x	x	x	x	x		x	x						x		x		x	x		
	Prototype																			x		x				x
Solution Method	Response factor						x	x	x	x	x	x	x	x							x		x	x		
	Implicit fin diff														x				x						x	x
	Explicit fin diff				x	x											x	x				x				
	Other	x	x	x												x				x						
Window model	Fixed U-value				x	x					x								x	x	x	x			x	
	Variable U-value						x	x	x	x					x											
	TMC *	x	x	x								x	x	x		x	x	x					x	x		x
Internal heat transfer coeff	Fixed				x	x	x	x	x	x				x					x	x	x	x		x	x	
	Varying	x	x	x							x	x	x		x	x	x	x					x			x
Air cavity model	Fixed resistance	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Varying																									
External longwave loss	Explicity modelled	x	x	x							x	x	x	x		x	x	x	x	x	x	x	x	x		
	Not modelled/fixed				x	x	x	x	x	x					x										x	x
Diffuse sky model	Isotropic					x	x				x	x	x	x	x	x	x	x			x		x			x
	Anisotropic	x	x	x	x			x	x	x									x	x		x		x	x	
Internal solar distribution	To floor	x		x												x					x					
	To various surfaces		x		x	x	x	x	x	x	x	x	x	x	x		x	x	x	x		x	x	x	x	x
Weather conversion needed	No						x	x	x	x		x	x		x	x									x	
	Cloud cover creation										x						x	x			x	x		x		
	Hour centering	x	x	x	x														x	x	x		x			
	Other					x								x												x
Heater model	Pure convective				x	x	x	x	x	x											x	x		x		
	Convective and rad	x		x													x	x					x		x	
	Detailed		x																x	x						
	Unknown										x	x	x	x	x	x										

* Transparent Multi-layer Construction

The site handbook also contained a diskette of the recorded weather data (except the vertical irradiance measurement). These data were presented to the modellers as hourly averages centred both on the hour and on the 1/2 hour—to suit the needs of different programs. The internal measurements and vertical irradiances were retained by the EMC until Phase 1 was complete.

The validation guidebook gave a description of the simulations to be undertaken, guidance on how to proceed, and the format in which the results were to be presented.

Each participant was required to predict hourly values for all three rooms for the last 7 days of the 10-day collection period. (Data from the first 3 days was provided so that modellers could pre-condition their programs.) The required hourly values were:

(i) the room air temperature (both periods);

(ii) the temperatures of the floor, back wall and ceiling (both periods);

(iii) the heating energy (power) consumption (October period); and in addition

(iv) the south-facing global solar irradiance.

Given the large number of participants, this produced a considerable amount of data. These have all been retained and archived at DMU.

The validation guidebook asked the program users "to model each of the three rooms in as much detail as the simulation program will allow". To help achieve this, a direct 'hotline' to DMU was established in order that all the modellers could immediately resolve any uncertainties that they encountered. All details of the construction, operation, monitoring and experimental quality control were available to modellers. All enquiries and the responses were logged. A timetable was defined in order to give each modeller a similar opportunity to produce the best possible results from their program.

To ensure that all the modellers had access to all the relevant information, newssheets were circulated. These listed all the current participants, the relevant hotline enquiries made, and the responses given. The newssheet exposed a number of subtle modelling aspects which some of the participants had clearly not appreciated—it therefore proved to be a useful learning tool. In all, 7 newssheets were produced during the 12 months of the blind phase.

A small number of errors was revealed in the site handbook and the validation guide. The most significant was that during the October period, Room 5 was specified as containing 1.5 m^2 of single-glazing, whereas in fact it had 0.75 m^2 of double-glazing. The predictions in this case could not, therefore, be compared with measured values. However, they were still used in the intermodel comparison. A weakness in the measurements was the missing relative humidity data for the October period.

Each modeller submitted the first set of results to the DMU team along with the input files which they had created and a completed questionnaire describing the key features of their program.

The onus was on the modellers to conduct appropriate quality assurance checks; however, the DMU team also inspected the input files to try and detect any obvious errors. They fed the results of this inspection back to the modellers. Many of the errors were minor but in two or three cases serious modelling errors had been made. Careful data checking by program users is essential to ensure the quality of DSP predictions. Following the feedback, most participants sent a second set of Phase 1 results and the new input files to the DMU team.

It was not possible for the authors/vendors of TAS (EDSL) and ESP+ (ASL Sterling) to participate in quite the way that has been already described. Because EDSL had previously worked with the EMC using data from the test rooms, they could not work blind, and ASL Sterling did not have sufficient resources to participate. To ensure that these important UK programs were included, DMU produced the first set of input files and the first set of results for each program. These were then sent to the vendors so that they could check the files and correct any errors. The two vendors then sent a second set of results and input files to the DMU team (in the case of ESP+, no hourly results files were obtained). Thereafter these two programs were treated in the same way as the other 23 program/user combinations.

Throughout this phase, the measured building performance data were retained by the EMC, not even the DMU team had access to the data. Thus the exercise was truly 'blind', and the hotline responses, newssheet entries, and personal feedback to modellers could not, even unwittingly, bias the results. The two-stage data checking process sought to minimize the likelihood of errors in the input data.

All participants submitted their results, along with the program input files, before Phase 2 work began. The 25 sets of results obtained at the end of Phase 1 were analysed and reported to the participants.

5.3. *The open phase*

The primary purpose of Phase 2 was to give all participants the opportunity to explore the reasons for any divergence between the predictions of their programs and the measurements. To facilitate this, all program users were sent a diskette containing all the measured data and estimates of the uncertainties (external errors) in the building description and climate data.

As in Phase 1, newssheets were circulated to keep all the participants informed of progress, any future plans, and the emerging results. Seven further newssheets, making 14 in all, were produced during this phase. The participants were asked to provide a three-page model user's report explaining their investigations. They were invited to:

(i) explore the reasons for any divergence between the predictions and the measurements;

(ii) undertake sensitivity analyses with their own programs;

(iii) provide a new set of predictions where modifications to the room descriptions or the program had been made;

(iv) comment on the IEA empirical validation exercise; and

(v) provide further descriptive information about their program.

It was hoped that these reports would: (a) permit improvements to be made to the validation package; (b) help to formulate recommendations about the conduct of future validation exercises; and (c) help direct future program development work by highlighting perceived areas of weakness in the current generation of DSPs. Reports were obtained from 11 participants.

New input files and results were produced for all 6 rooms for: BLASTv3.0lvl203 (which is a different version from that which the Turin group used in Phase 1); HTB2v1.2; SERIRESv1.2; SUNCODEv5.7; and TSBI3v2.0. The vendors of TASv7.54 produced results for the double-glazed and opaque buildings only. For the unheated, free-floating buildings, new results were obtained from WG6TCv1992.

These new results were generated for a number of legitimate reasons (Table 5); usually because the room model could be improved, but also because mistakes had been made, or program errors had been detected. (The DMU team were able to compare the input files generated in Phase 1 with the new files submitted in this Phase.) There was no evidence of attempts to manoeuvre an improved fit between the predictions and the known measurements.

Table 5
Modifications to modelling procedures to produce the Phase 2 results

Program	Modifications
WG6TCv1992 (Udine, I)	coding error was corrected [a]
TSBI3v2.0 (SBI, DK)	incorrect heater schedule had been used (heater off on last day for opaque case). This was corrected [b] horizon altitude was incorrectly modelled [c] direct normal and diffuse horizontal irradiance data preferred [c]
TASv7.54 (DMU/EDSL, UK)	construction details refined [c] shading effect by neighbouring test room modelled [c] more detailed modelling of interior solar distribution [c] infiltration rates of roofspace and floorspace were adjusted until their air temperatures matched the measured temperatures [c] internal clock adjusted by 1/2 hour for consistency with measured data [c]
HTB2v1.2 (FHT, GER)	solar calculation routine was originally in error. This was corrected [a] timing convention in error for May period. This was corrected [c]
BLASTv3.0lvl203 (Torino, I)	modified program version used [a] ceiling insulation had been omitted. This was added [b] the partition wall had been modelled as an external wall. This was corrected [b] adjacent cell shading was not modelled [c] roof absorptivity was in error. This was corrected [b]
SUNCODEv5.7 (Ecotope, US)	building orientation was incorrectly specified (as 9° east of south rather than 9° west). This was corrected [b]
SERI-RESv1.2 (BRE, UK)	inconsistency in climate data was corrected [c]

[a] program improvement.
[b] input error correction.
[c] modelling improvement.

6. Methods of analysis

Data management was simplified because all the participants were asked to produce their hourly predictions in the same consistent fashion. The computer-based statistical analysis and visualization program PV-Wave [28] eased the analysis and data presentation process. Nevertheless, the large amount of data precludes a detailed presentation of all the results so the analyses focused on:

(i) the total 7-day heating energy consumption during the October period;

(ii) the maximum and minimum air temperatures in all 3 rooms during both 7-day periods; and

(iii) the total south-facing vertical global solar irradiance for both 7-day periods.

By studying the south-facing solar irradiance it was possible to obtain an insight into the accuracy of the sky models used by the DSPs and the extent to which errors in predictions might be due to any inaccuracies in them.

Programs are often used to predict the changes in building performance that will occur as a result of design alterations, or to predict differences between the thermal performance of a proposed design and a 'reference' building. Therefore analyses were also undertaken for:

(iv) the difference in the total 7-day heating energy consumption between the opaque room and the 1.5 m^2 double-glazed room (October period);

(v) the difference in peak air temperature between the double-glazed room and the opaque room (for the May period); and

(vi) the differences in peak air temperature between the double-glazed room and single-glazed room (in the May period).

All these parameters (i to vi) are illustrated by simple bar charts which show the average result (for all the programs) and the variation in the predictions, i.e. the interprogram variability. This is represented by the upper and lower bounds within which 95% of predictions would be expected to lie if the results were normally distributed (i.e. 1.96 times the calculated standard deviation). In practice, since only 24 (or 25) results are available, the upper and lower bounds are very close to the actual maximum and minimum predicted values, i.e. the difference between the bounds is virtually the same size as the range in the predictions.

The bar charts also show the measured values surrounded by an uncertainty band (shaded) which represents the magnitude of the errors in the measurements and program input data. The band is such that there is only about a 1% chance that the predictions could lie above the upper edge or below the lower edge by chance. In other words, there is about a 99% chance that the predictions which fall outside the band are different from the measurements (for example due to internal errors in the program or to mistakes by the program user). The procedure used to calculate the error band is described in the Appendix, in which all the salient predicted measured and calculated values are tabulated.

To give more insight into the underlying reasons for any difference between the predictions themselves and between the predictions and the measurements, simple graphical comparisons of hourly values are presented for one typical day in the October period and one typical day in the May period. Simple statistical measures of the discrepancy between the measurements and predictions over the whole 7-day period were derived [1] but complex statistical techniques, such as time-series analysis, cross-correlation analysis [29], or qualifying power spectrum tests [30] were not employed. This was partly because the duration of the exercise precluded it and partly because sufficient insight was gained without them.

The simple statistical measures of the differences between the predictions of parameters (i) to (iii) changed very little between Phases 1 and 2, even though new Phase 2 results were received for 7 programs. (This was because, in general, the results for these seven programs did not change by much and because small increases in predictions by one program were compensated for by the reduced predictions of another.)

In this paper, the predictions available after Phase 2 are analysed. However, the results produced during Phase 2 are distinguished from the majority, which were produced in Phase 1, in the bar charts by shading, in tables by italics and in graphs by an asterisk.

7. Intermodel comparison

This section compares the predictions of the programs only, comparisons between the predictions and the measurements are made in Section 8. By separating the comparisons in this way, it is possible to describe the interprogram variability without the complications introduced by experimentation—in interprogram comparisons these are irrelevant. Readers should, however, avoid the temptation to assume that programs which differ from the majority are, in some sense, in error. The divergent predictions may be due to a special modelling feature which renders the program superior to the majority. The comparison of predicted values with those measured illuminates this matter.

7.1. Results for heated rooms

The overall variability in the predictions of energy consumption (Fig. 6) was of similar magnitude for all three rooms. The range in predictions, expressed as a percentage of the mean value, was about 40% for both the opaque and double-glazed rooms, and 51% for the single-glazed room.

In all three rooms, DEROBvlth, TRNSYSv13.1 and ESP + v2.1 produced predictions which were invariably lower than those produced by the other programs. In fact, all four TRNSYS programs and all three ESP programs always produced predictions towards the lower end of the range. No single program consistently predicted values that were higher than those from the others.

All the programs correctly predicted that the double-glazed room would consume less energy than the opaque room (Fig. 7). However, the predicted energy savings varied by a factor of about 3, i.e. from 13% for HTB2v1.10 to 40% for ESP + v2.1. This result does not support the contention that programs can predict the differences in energy demands of two buildings more reliably than their absolute energy demands (the absolute energy demands varied by a factor of about 1.7 in both rooms).

Thermal simulation programs are often called upon to predict the peak internal temperatures in buildings because these

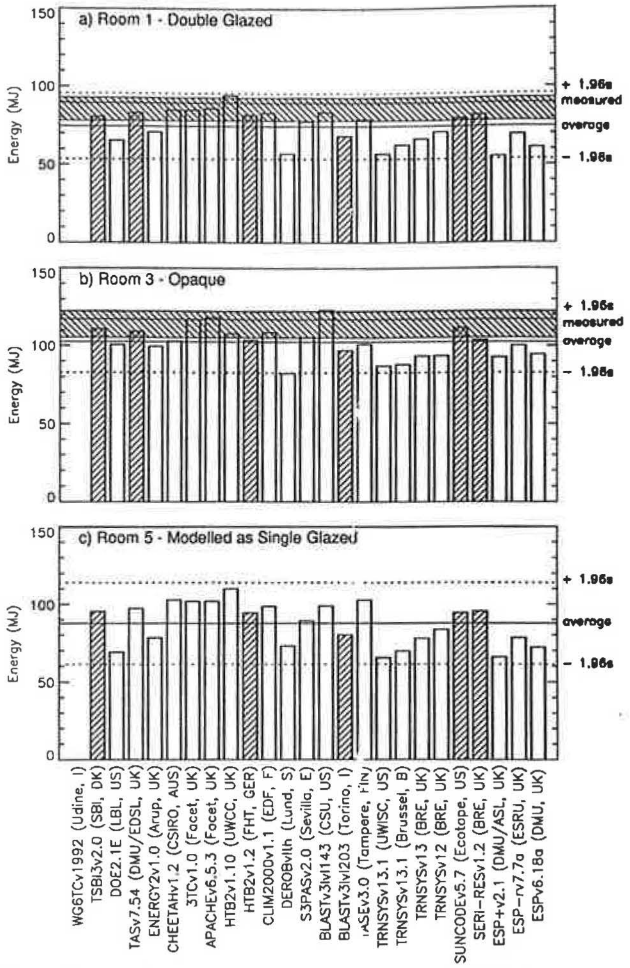

Fig. 6. Measured and predicted total heating energy demands in the opaque and double-glazed rooms during the October period.

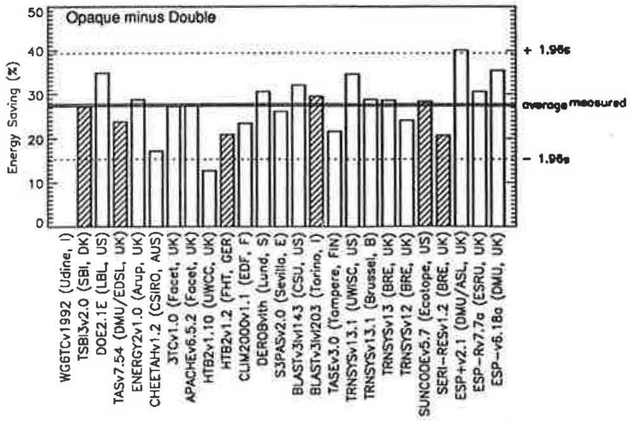

Fig. 7. Measured and predicted difference between the total heating energy demand of the double-glazed room and the corresponding demand for the opaque room.

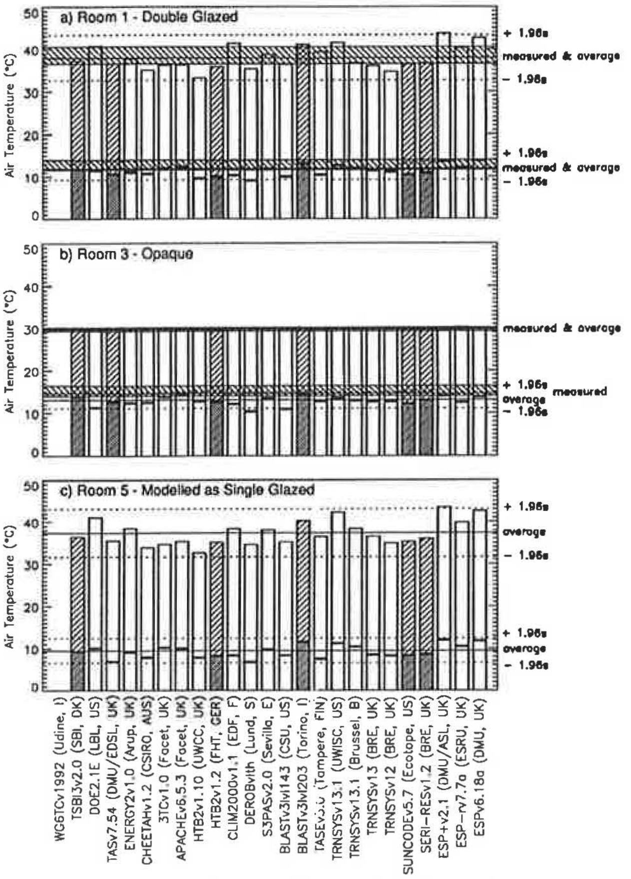

Fig. 8. Measured and predicted maximum and minimum air temperatures in the opaque and double-glazed rooms during the October period.

influence the thermal comfort of occupants and give an indication of whether mechanical (air conditioning) equipment is likely to be needed. In the glazed rooms, the predicted peak air temperatures varied by 11°C, i.e. from 3°C to 14°C above the (30°C) set point (Fig. 8). In both rooms, TRNSYSv13.1 and ESP + v2.1 produced predictions which were invariably higher than those produced by the other programs. This is the corollary of their lower energy consumption predictions (Fig. 6) and suggests that these programs are assuming an overall net heat loss which is lower than that assumed by the other programs.

The programs predicted minimum air temperatures (which occurred during the nighttime unheated period) in all three heated rooms, which varied by about 5°C (Fig. 8).

Although the absolute hourly predictions of heating energy demand and internal air temperature depended on the room being studied, there were consistent features in the relative predictions of the programs. It can be seen (Fig. 9(c)) that DEROBvlth predicted peak heating power demands immediately after start-up, which were less than the capacity of the heater; all the other programs predict maximum power output. DEROB also predicted a more rapid rise in the room air temperature.

There was often considerable variation in the hourly heating energy demand predictions. As an illustration, the demands at hour 18 varied from about 700 to 1200 kJ (Fig. 9(c)).

7.2. Results for free-floating rooms

In the glazed rooms, the peak air temperatures predicted by WG6TCv1992 and DEROBvlth were higher than the values predicted by the other programs (Fig. 10). The hourly temperature results for these two programs were also noticeably different from the results for the other programs (Fig. 11(b)) although the WG6TCv1992 results were much closer to those obtained by the other programs than was the case in Phase 1. At that stage, WG6TCv1992 had predicted a peak temperature in the double-glazed rooms which was about 22°C higher than any other prediction and a minimum temperature which was 5.7°C less than that predicted by any other program.

The 'improvements' arise because a coding error was found in the program and corrected (and not because the input data was changed). The search for the error was prompted by the poor Phase 1 performance in this exercise, the error had not been revealed by previous tests. This illustrates the usefulness of the empirical validation work.

As already noted, the reliability of peak air temperature predictions is of particular importance in the context of real building design applications. In the double-glazed room, the predicted values varied by 8.6°C, from 26.4°C for HTB2v1.10 to 35.0°C for DEROBvlth. In the single-glazed room their range was 7.0°C (ignoring the results for DEROBvlth the ranges become 6.3°C and 6.2°C, respectively). In a building design context, predictions accurate to within a degree or two may be sought.

All the programs predicted that the peak temperatures in the double-glazed room would exceed those in the opaque room (Fig. 12). However, predicted differences in peak temperatures varied from 10.2°C (for HTB2v1.10) up to 17.2°C (for DEROBvlth). Even ignoring the DEROB result, the range was from 10.2°C to 15.8°C. This range, 5.6°C, is only marginally less than the range in the absolute peak temperature predictions for the glazed rooms (6.3°C and 6.2°C). The result does little to support the frequently-espoused view that programs are more consistent when predicting the differences between the temperatures in two buildings than they are at predicting the absolute temperatures in a single building.

Most programs predicted that the peak temperatures in the single-glazed room would exceed those in the double-glazed room; however, CLIM2000 and TASEv3.0 were notable in that they predicted the reverse trend (Fig. 12).

7.3. Results for south-facing vertical solar irradiance

The average of the south-facing vertical solar irradiance predictions was similar for both periods, a total of 76.2 MJ in the October period and 80.0 MJ in the May period. However, the predictions were much more variable in the October period, ranging from 84.1 MJ for TRNSYSv13.1 (Brussels), [redacted] [redacted]

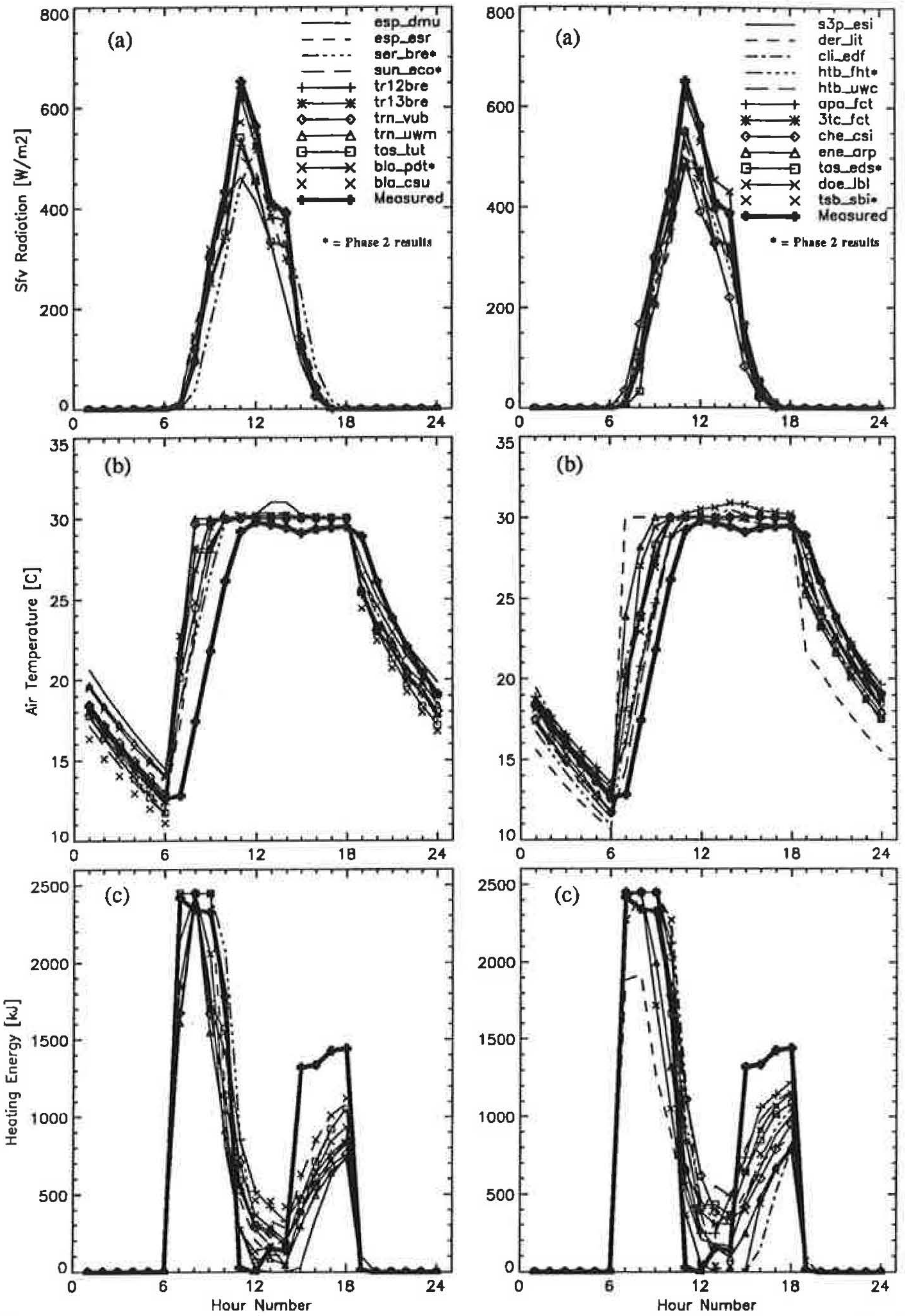

Fig. 9. Measured and predicted hourly energy demands, air temperatures and south-facing vertical solar irradiances in the double-glazed room on 23 October.

down to 67.0 MJ for HTB2v1.10 (Fig. 13). This range is 22% of the average value. Thus, even before the complex thermal interactions which take place in the room are considered, there is roughly a 22% difference in the potential solar radiation energy flux. Both TRNSYS (the highest predictor) and HTB2v1.10 (the lowest predictor) use an isotropic sky model. Considering only those models with anisotropic sky models, the range in the predictions reduces to about 10% (from 71.9 MJ for Cheetah to 79.1 MJ for S3PAS).

The total range (22%) is roughly half the total variability observed in the total heating energy demand predictions for October. A plot of the predicted energy demands of the double-glazed room versus the predicted south-facing solar irradiances did not, however, reveal any clear causal relationship.

7.4. Comparison with results from other IEA studies

To gain some reassurance that the variations in the predictions were not due to peculiarities in the test rooms, inadequacies in the room descriptions, or mistakes by the program users, the results were compared (as far as was possible) with those obtained in the two other intermodel studies being undertaken by the IEA Annex 21/Task 12 group.

Nine of the programs which participated in this empirical validation exercise had also been used in the parallel Building Energy Simulation Tests (BESTEST) [2,3]. The BESTEST work used hypothetical domestic-scale buildings, located in Denver, USA, heated by an ideal warm air system. One BESTEST building had a modest amount of thermal mass and south-facing double-glazing. The annual heating energy demand predictions for this building were compared with those produced for the heated double-glazed EMC room.

Seven of the programs had also produced results for a hypothetical commercial building [4]. This was a three-zone building, also located at Denver, with a modest amount of thermal mass, consisting of 'offices' facing north and south with a corridor between. The offices were heated by an ideal warm air system in winter and cooled by ventilation in summer. The predicted annual energy demands for the double-glazed south-facing office were compared with the predicted energy demands for the double-glazed room.

There were many differences between the EMC experiments and the two hypothetical buildings, and so the absolute magnitude of the predictions differed considerably. However, the rank order of program predictions was broadly consistent (see Ref. [1] for details). In all the buildings, there was a tendency for versions of ESP to predict lower energy demands than the other programs.

8. Comparison of measured and predicted values

8.1. Results for heated rooms

Eight programs predicted energy demands within the error band for the opaque room, and 11 were within the error band for the double-glazed room (Fig. 6). Six programs (3TCv1.0, APACHEv6.5.3, CLIM2000v1.1, SUNCODE-v5.7, TASv7.54 and TSBI3v2.0) produced results inside the error band for both rooms. More programs produced predictions that were inside the error band for the double-glazed room than for the opaque room—which, on the face of it, seems to be an easier modelling problem.

The measured reduction in the energy demand due to replacing the opaque facade with a double-glazed window was 24% (Fig. 7). The predictions of fourteen programs were within 5% of the measured energy savings and three programs were within 1%—Tasv7.54 and CLIM2000v1.1 predicted the energy demands of both rooms well, whereas TRNSYSv12 underpredicted the energy demands of both rooms in a consistent fashion. Five programs, DOEv2.1E, HTB2v1.10 TRNSYSv13.1 (UWISC), ESP+v2.1 and ESPv6.18a, predicted energy savings which differed from the measured savings by more than 10%.

In the double-glazed room, the solar gains drove the temperatures above the 30°C set point (Fig. 8). Seven of the programs, TSBI3v2.0, ENERGY2v1.0, S3PASv2.0, TASEv3.0, TRNSYSv13.1 (Brussels), SERI-RESv1.2 and ESP-rv7.7a, predicted values within the error band. Two programs, ESP+v2.1 and ESPv6.18a predicted values more than 4°C above the measurements and HTB2v1.10 predicted over 4°C below the measurements.

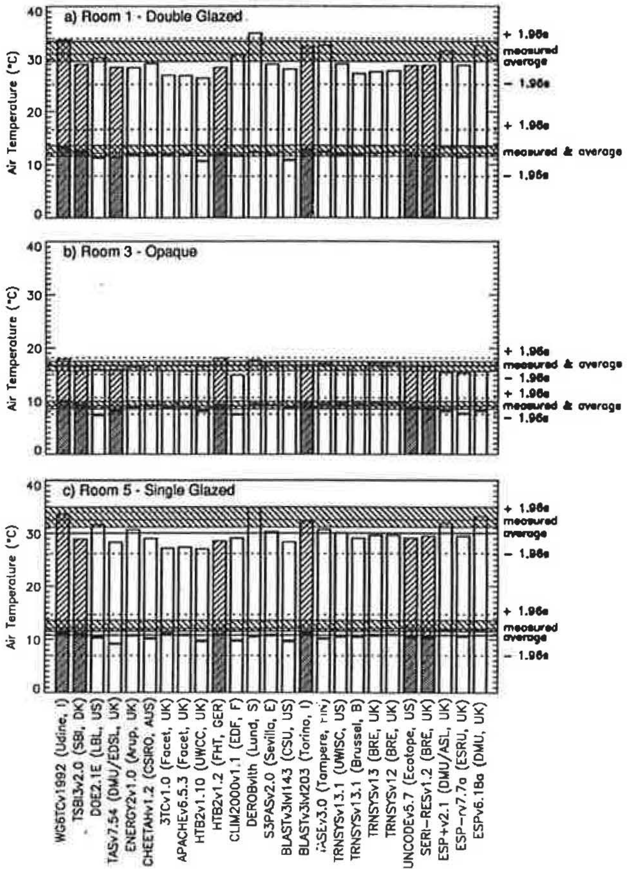

Fig. 10. Measured and predicted maximum and minimum air temperatures in each of the three rooms during the May period.

There was a general tendency towards underpredicting the minimum temperatures (Fig. 8). Ten programs predicted inside the error band for the double-glazed room and five programs, TSBI3v2.0, 3TCv1.0, APACHEv6.5.2, BLAST-v3.0lvl203 and ESP+v2.1, produced predictions inside the error bands for both rooms.

One program, TSBI3v2.0, predicted maximum and minimum temperatures and energy demands which were always inside the error bands.

All the DSPs predicted a more rapid rise in the air temperature at the start of the heating period than that which was actually measured (e.g. Fig. 9(b)). Most of the DSPs also predicted a faster decrease in the air temperature at the end of the heating period; DEROBvlth exhibited the most extreme behaviour of this type. Overall the air temperature predictions of APACHEv6.5.3, HTB2v1.10 and HTBv1.2 were the closest to the measurements. These results typify those obtained on other days and in the opaque room.

Because the programs predicted a rapid rise in the air temperature, the set point was reached earlier (in the case of DEROBv1th 3 h earlier) than was in fact the case (Fig. 9(b)). As a result, the predicted power output from the heater tended to decline much more rapidly than the measured

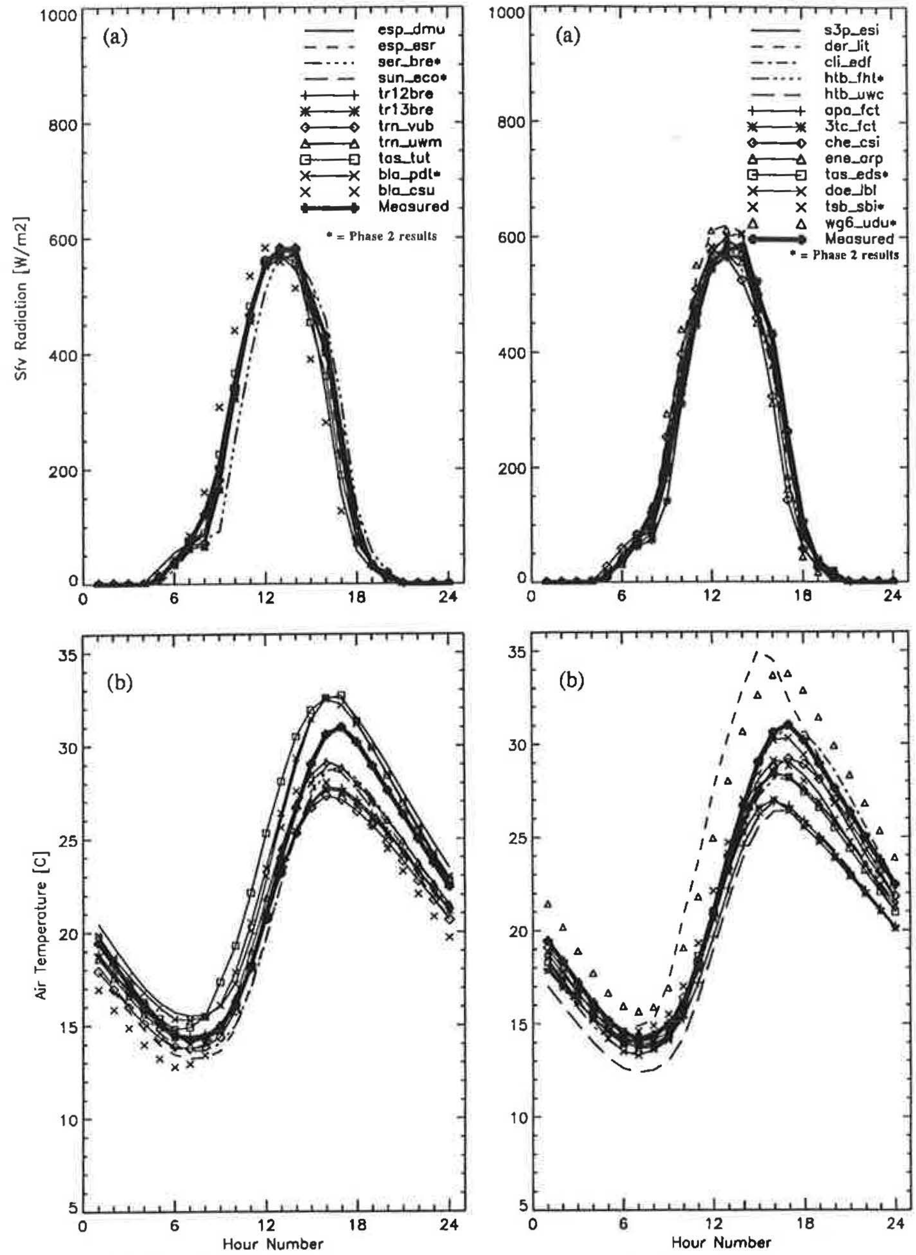

Fig. 11. Measured and predicted hourly air temperatures and south-facing vertical solar irradiances in the double-glazed room on 27 May.

power output (Fig. 9(c)). Such differences could be caused, in part, by the assumption in some programs that the room heater and its controls are ideal. In fact, the system has some thermal inertia which produced a time-delayed response [1].

For both the opaque and double-glazed rooms, the closest overall agreement between the hourly measured and predicted heating energy values were produced by 3TCv1.0, APACHEv6.5.3, BLASTv3.0lvl203, SUNCODEv5.7 and TASv7.54 (Fig. 9(c)). For the opaque room, TSBI3v2.0 also performed well. For both rooms, DEROBvlth and TRNSYSv13.1 (UWISC and Brussels) showed marked differences from the measurements. During the last four hours of the heating period, all the programs predict much lower power demands than was actually measured (Fig. 9(c)).

8.2. Results for the free-floating rooms

The prediction of internal temperatures in a free floating opaque 'box' is perhaps one of the easiest real-world prediction problems with which a DSP could be faced. Overall 15 programs predicted both maximum and minimum temperatures which were within the error bands (Fig. 10(b)). Four programs, CLIM2000v1.1 and the three ESP versions, produced temperature predictions which differed from the measurements noticeably more than did the other programs (Fig. 10(b)).

Four programs, DOEv2.1E, BLASTv3.0, ESP+v2.1 and ESPv6.18a, produced predictions of peak air temperature inside the error band for both glazed rooms. (Fig. 10(a) and

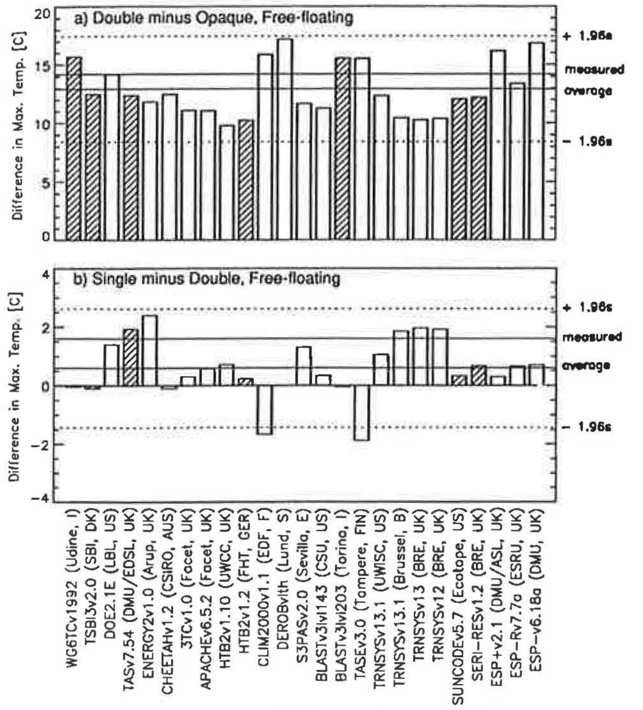

Fig. 12. Measured and predicted differencesbetween: (a) the peak air temperatures in the double-glazed room and the corresponding values in the opaque room; (b) between the peak air temperatures in the single-glazed room and the corresponding values in the double-glazed room.

10(c)), but 17 of the 25 predictions lay below the lower error bound for both rooms with 3TCv1.0, APACHEv6.5.2 and HTB2v1.2 being over 4°C below the measurements (Fig. 10(a), 10(c) and Appendix).

The generally low peak air temperature predictions reflect the underlying tendency to predict low hourly values, this is illustrated by Fig. 11(b). WC6TCv1992 and DEROBvlth are notable in that they produced hourly values (Fig. 11(b)) that are well above the measurements.

All the programs correctly predict that a change from double-glazing to the opaque panel would result in a reduction of the peak temperature (Fig. 12(a)). Eighteen programs predicted reductions which were within 3°C of the measured reduction (14.2°C) with four programs, ESP-rv7.7a, DOE2.1E, and both BLAST programs, predicting within 1.5°C of the measurement.

All but two programs correctly predicted that the peak temperature in the single-glazed room would exceed that in the double-glazed room (Fig. 12(b)). Eighteen programs predicted results which were within 1.5°C of the measured temperature difference (of 1.6°C) and only two programs, CLIM2000v1.1 and TASEv3.0, produced values which deviated from the measurement by more than 3°C.

8.3. *South-facing vertical solar irradiances*

For the May period, all the programs predicted total south-facing vertical solar irradiances which were within the error bands (Fig. 13(a)).

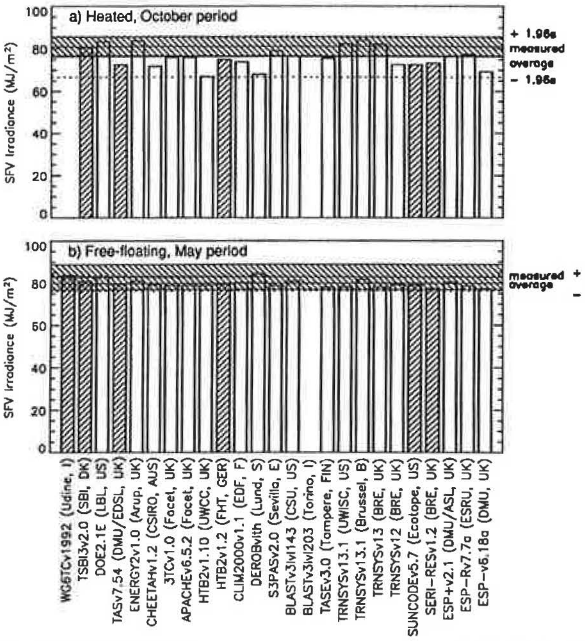

Fig. 13. Measured and predicted total south-facing vertical solar irradiances for the October and May periods.

For the October period, most programs predicted total irradiances that were below the error band (Fig. 13(b)), and the hourly values (Fig. 9(a)) showed a similar underprediction. This could explain why, during the middle of the day, they predicted heating energy demands that were above the measured demands (Fig. 9(c)). However, if radiation effects were dominant, one would expect, given this trend, that the total heating energy demands would be overpredicted. In fact, the reverse is true, so it is likely that other factors are more influential and causing the low energy demand predictions. (This is discussed further below.)

In all, 9 programs produced predictions of total irradiance in the October period which were within the estimated error band (Fig. 13(a)). The hourly plots (Fig. 9(a)) and the simple statistics (see Ref. [1]) confirmed the good predictions for TSBI3v2.0, DOE2.1E, ENERGY2v1.0, S3PASv2.0, TRNSYSv13, TRNSYSv13.1 (Brussels) and TRNSYSv13.1 (BRE). All these programs except S3PAS v2.0 used an anisotropic sky model.

8.4. *Overall program performance*

The foregoing observations and discussions focus on the performance of the programs when undertaking individual prediction tasks. Their overall performance is illustrated in Table 6, which shows that none of the programs produced predictions within the estimated error bands for all 12 of the parameters. It is equally clear however, that some programs performed much better than others.

9. Interpreting the results

In principle, the interpretation of the results is quite simple. Once predictions fall outside the error bands then the proba-

Table 6
Summary of results

	Heated, October Period						Free-floating, May Period						Sfvr		Number of Parameters within Bands[2]
	Double glazed			Opaque			Double gl.		Single gl.		Opaque		Oct.	May	
Program	E	$\hat{T}$	$\check{T}$	E	$\hat{T}$	$\check{T}$	$\hat{T}$	$\check{T}$	$\hat{T}$	$\check{T}$	$\hat{T}$	$\check{T}$			
WG6TCv1992 (Udine, I)	n/a	n/a	n/a	n/a	n/a	n/a	–	○	○	–	–	–	n/a	●	2 (out of 6)
TSBI3v2.0 (SBI, DK)	○	○	○	○	○	○	–	○	–	–	○	○	○	○	9
DOE2.1E (LBL, US)	–	–	●	–	●	–	●	–	●	–	●	–	●	●	5
TASv7.54 (DMU/EDSL, UK)	○	–	–	○	○	–	–	○	–[1]	–[1]	○	–	–	○	5
ENERGY2v1.0 (Arup, UK)	–	●	–	–	●	–	–	●	–	–	●	●	●	●	5
CHEETAHv1.2 (CSIRO, AUS)	●	–	–	–	●	–	–	●	–	–	●	●	–	●	5
3TCv1.0 (Facet, UK)	●	–	●	●	●	●	–	●	–	–	●	●	–	●	8
APACHEv6.5.3 (Facet, UK)	●	–	●	●	●	●	–	●	–	–	●	●	–	●	8
HTB2v1.10 (UWCC, UK)	–	–	–	●	●	–	–	–	–	–	●	–	–	●	3
HTB2v1.2 (FHT, GER)	○	–	–	–	○	–	–	○	–	–	–	○	–	○	4
CLIM2000v1.1 (EDF, F)	●	–	–	●	●	–	●	●	–	–	–	–	–	●	5
DEROBvlth (Lund, S)	–	–	–	–	●	–	–	●	–	–	–	●	–	●	3
S3PASv2.0 (Sevilla, E)	–	●	●	●	●	–	–	●	–	–	●	●	●	●	7
BLASTv3.0lvl143 (CSU, US)	●	–	–	–	●	–	–	–	–	–	●	●	–	●	4
BLASTv3.0lvl203 (Torino, I)	–	–	○	–	○	○	○	○	○	–	○	○	n/a	n/a	8
TASEv3.0 (Tampere, FIN)	●	●	–	–	●	–	●	●	–	–	●	●	–	●	7
TRNSYSv13.1 (UWISC, US)	–	–	●	–	●	–	–	●	–	–	●	●	●	●	5
TRNSYSv13.1 (Brussels, B)	–	●	●	–	●	–	–	●	–	–	●	●	●	●	6
TRNSYSv13 (BRE, UK)	–	–	●	–	●	–	–	●	–	–	●	●	●	●	5
TRNSYSv12 (BRE, UK)	–	–	–	–	●	–	–	●	–	–	●	●	–	●	4
SUNCODEv5.7 (Ecotope, US)	○	–	–	○	○	–	–	○	–	–	○	○	–	○	6
SERI-RESv1.2 (BRE, UK)	○	○	–	–	○	–	–	○	–	–	○	○	–	○	6
ESP+v2.1 (DMU/ASL, UK)	–	–	●	–	●	●	●	●	●	●	–	–	●	●	7
ESP-Rv7.7a (ESRU, UK)	–	●	●	–	–	–	–	●	–	–	–	–	●	●	3
ESPv6.18a (DMU, UK)	–	–	–	–	●	–	●	●	●	●	●	–	–	●	6

○ and ● indicate predictions within error bands, ○ for Phase 2 results and ● for Phase 1 results.

– indicate results outside the error band.

italics indicate programs for which Phase 2 results were obtained.

E = total heating energy consumption over 7 days; $\hat{T}$ = maximum temperature; $\check{T}$ = minimum temperature;

Sfvr = total South facing vertical solar irradiance

[1] a Phase 1 result; [2] Excludes South facing vertical irradiance (Svfr), so maximum possible score = 12

bility that the result could be due to uncertainties in the experimentation (an external error in the experiments) is so small (in this case about 1%) that it is reasonable to assume that the program contains internal errors. For this interpretation to stand, it is essential that all sources of experimental uncertainty have been identified and correctly accounted for and that no user errors remain.

9.1. Impact of program user

In general, mistakes by program users are very hard to eliminate, because the data input requirements of DSPs are onerous and diverse, and because, for many programs, the data input interface is cumbersome. In this validation exercise however, every effort was made to avoid these errors, by

(i) using experts, program vendors or program authors to produce the predictions whenever possible;

(ii) emphasizing to users the importance of accurate modelling and quality assurance checking;

(iii) adopting a two-stage process in Phase 1 to try and trap any user errors;

(iv) presenting the building descriptions and weather data in a form which was readily useable by DSP users;

(v) offering assistance through a hotline and newssheets; and

(vi) including Phase 2 in which program users had the opportunity to rectify any mistakes and/or to improve their modelling procedures.

These efforts, together with the broad agreement with the results from the other IEA intermodel comparison exercises, indicate that, as far as was reasonably practicable, the influence of program users has been eliminated. If user errors do remain, then they are also likely to exist when other, less skilled, operators employ the program, especially under the time constraints of a real design situation.

It is worth noting, however, that at least one program yielded different results when used by different people (compare for example, TRNSYS13.1 UWISC and TRNSYS13.1 Brussels). Such differences may be due to a lack of guidance on how to model a specific problem (by a program users' manual or through the user interface) or ambiguities and a lack of clarity in the user interface. In this context, the division between an external, user-introduced, error and an internal program error, becomes rather blurred. Either way it falls to program authors and vendors/distributors to rectify the deficiency. It is widely acknowledged that there is considerable scope for improving the interfaces of many thermal simulation programs.

9.2. *Accounting for experimental error*

It is beyond the scope of this paper to detail all the efforts made to identify, quantify and account for experimental error. However, the measures taken were much more rigorous than those usually adopted when collecting high quality data sets (see Ref. [1]). Some of the special measures undertaken to address specific concerns raised by the IEA participants are worth noting here. These included:

(i) partly dismantling one room to compare the actual construction with the description given to the modellers;

(ii) conducting a steady-state heating (co-heating) trial just two weeks after the October heating period;

(iii) conducting periodic 'matching trials' to ensure that rooms were thermally identical, except for the deliberately introduced differences; and

(iv) periodic air leakage tests.

Taken as a whole, the quality assurance procedures ensured that, to within experimental accuracies, the EMC rooms were thermally identical, that they had not deteriorated (e.g. due to rodent damage or water leakage) over the experimental period, and that their construction and air leakage characteristics conformed in all significant respects to the descriptions given to the program users.

These measures, together with the knowledge that the calculated uncertainty bands are likely to be over-, rather than underpredicted, indicate that all significant experimental errors have been properly accounted for.

9.3. *Impact of errors in programs*

The foregoing discussions (Sections 9.1 and 9.2) lead to the conclusion that, when programs lie outside the error bands, there are likely to be internal errors within the program (or deficiencies in the program documentation and interface). In this regard, no program produced predictions that were inside the error bands on all occasions (Table 6), however, some programs' predictions fell well outside the error bands on a number of occasions. These programs are a particular cause for concern, the likely sources of the internal errors are discussed in Section 10.

When programs perform well, it could be due to a number of factors:

(i) the program contains no internal errors;

(ii) any internal errors which do exist are benign, for the particular conditions being tested; and/or

(iii) internal errors exist but they are counteracted by other, compensating, internal errors.

In practical terms, it does not matter which of the above possibilities is true, provided the experiment reflects the real world conditions in which the program may actually be used. The EMC rooms stress fabric heat loss issues which are a feature of virtually all buildings. Thus, if programs perform well in this validation work, their credibility will be increased.

It is often argued that, in a real design situation, other factors are more important. These would include modelling capability, input and output style, extent of in-built databases, and speed of simulation. Certainly, modellers should consider these issues before selecting a DSP for a particular task. It is self-evident, however, that if the DSP gives unreliable predictions, these other features are of limited worth. Ultimately, it is the accuracy of the result which determines the final building design and its thermal performance. It is this aspect for which thermal modellers are responsible and for which they may be legally liable.

10. Possible sources of internal error

When using information from this empirical validation exercise in isolation, it is impossible to identify the precise sources of internal error in a particular program. However, it is worth exploring some possibilities and, to this end, the internal errors have been divided into three types:

(i) coding errors;

(ii) the omission of submodels or algorithms for thermal processes that are important; and

(iii) the use of inappropriate algorithms and assumptions.

It is generally accepted that modelling approximations or coding errors exist in all large computer programs. This is an inevitable consequence of the complexities of the programming task. It is when these errors have a noticeable impact on predictions that they are of concern.

Coding errors are rather difficult to isolate by empirical validation alone because they are compounded by internal errors of types (ii) and (iii). However, errors of this type were revealed in DOE2.1D, ESPv6.18a, an earlier version of TASE and TRNSYSv12.1, by the complementary IEA BESTEST work [2]. Some programs performed reasonably well in the empirical validation exercise whilst others, which used many of the same algorithmic assumptions, performed rather poorly. This suggests the existence of critical, but as yet unidentified, coding errors in some programs. In this context, the persistent underprediction of heating energy demands by the ESP programs is a source of concern.

Sensitivity analyses conducted by some of the participants during Phase 2, plus some analyses undertaken by DMU (as part of their quality assurance checks), revealed that the algorithms describing the convective heat transfer at internal surfaces, and the heater output and its distribution, can have a large impact on predictions.

The characteristics of the heaters in the EMC rooms were typical of those used in domestic UK buildings (they used a mix of radiant and convective output and inherent thermal inertia). Few of the programs could model the dynamics and some could not model heaters with radiant output. The inability to model radiant output is the most serious and is likely to lead to underprediction of heating energy demands. Programs which appeared to be unable to model radiant heater output are: DOE2v1E; CLIM2000v1.1; ENERGY2v1.0; and TRNSYS. The modelling of heater dynamics was probably less important, at least for the prediction of total heating energy demand.

There was evidence from the measurements, CFD predictions and explorations conducted by the user of TASv7.54, that more rapid air movement occurred around the walls of the room than in the middle. The importance of this issue is still a matter of debate. Stratification is a well-known phenomena in real buildings as is the occurrence of rising plumes over emitters and down draughts close to windows. The real question, however, is the extent to which these phenomena may be having a disproportionately large effect in the small test rooms compared to real buildings. At present, this question cannot be answered.

Although none of the programs considers internal air flows in detail, a number of them produce predictions of individual primary parameters that are close to the measurements and a small number perform well for the whole range of parameters. This may be fortuitous, however, a simpler explanation is that air circulation details do not need to be considered in order to produce reliable predictions for the test rooms. Hopefully this is indeed the case, otherwise the reliable thermal analysis of many buildings will become much more complex and possibly impractical.

Many assumptions are made within DSPs (internal errors of type (iii)) because they are difficult, if not impossible, to avoid. They are introduced because of circumstances beyond the control of model builders (e.g. lack of data, the need for reasonable solution times, etc.). The predicted total south-facing vertical solar irradiances for the October period displayed a wide variation yet the prediction depends only on the sky and ground modelling algorithms.

A common assumption is that conduction is one-dimensional, and that surface heat transfer coefficients are constant. Neither is true, but there is no clear evidence that, for most real buildings, these assumptions lead directly to poor predictions.

One benefit of empirical validation is that it tests the combined effect of all the internal errors in a program. Intermodel comparisons and analytic tests do not usually do this, and, in particular, they do not 'pick up' internal errors of type (ii). It is for this reason that programs which have previously done well in other tests may nevertheless perform less well in rigorous empirical validation exercise.

11. Benefits of the IEA empirical validation benchmarks

The results in this paper are for specific versions of parti ular programs. They cannot be used to make inferences abo the performance of different versions of these programs (o of course, other programs). However, the IEA benchmar have numerous long-lasting benefits.

The five empirical validation benchmarks contain description of the EMC test rooms, guidance on how to mod them and a diskette containing the weather data and the mea ured performance of the rooms. The uncertainties in all pa ameters are quoted. The IEA participants believe that it is good example of how to document a benchmark for validatin DSPs. It is available, from the BRE publishers, for use b others [1].

The benchmarks have many benefits for program user vendors and potential purchasers of DSPs:

(i) to evaluate the predictions of new programs, or nev versions of old programs by comparison with high qualit measurements and in the context of the performance of othe well-known state-of-the-art DSPs;

(ii) as a spur to improving the performance of existin programs and to track the changes in performance as modifications are made; for example, the IEA work promptec improvements to the internal solar distribution models in CLIM2000 and was used for charting the further developmen of CHEETAH [31];

(iii) they enable DSP buyers to test the accuracy of the product, which will mean that numerical accuracy becomes a more prominent issue than hitherto in discussions with vendors; conversely, they provide a vehicle by which vendors can demonstrate the capabilities of their products; and

(iv) they can provide a training vehicle for new users and a test of the in-house quality assurance procedures adopted by users.

In general, benchmarks such as this provide a common focus for discussions between the authors/vendors of programs, their purchasers and validators. As such they will help to establish a common understanding of the accuracy which can be expected from DSPs in general, and specific programs in particular.

12. Future empirical validation needs

It is difficult to extrapolate from the performance observed in the IEA study in order to quantify the likely accuracy of predictions for other circumstances, but it will never be possible to test programs under all possible circumstances. However, it would be useful to extend the work from test rooms into full-sized buildings, and, in particular, to offices and houses. Past experience indicates that this will be extremely difficult. However, given the expertise which is now available, it may be possible to develop empirical validation packages based on such buildings. Enhancements to the available validation techniques will assist the process—for example, advanced time-series-analysis methods, and refined Monte Carlo strategies. These issues have been explored in a joint BRE, DMU, EMC, Electricité de France and Creteil University (Paris) project.

Validation will always be needed—to test new programs, new versions of existing packages, or new algorithms, and new building forms which will continue to stretch the capabilities of programs. It would be useful, therefore, to develop empirically-based data which test the accuracy of individual algorithms e.g. for surface convection, heater modelling, glazing systems and air circulation. This is currently being explored at De Montfort University.

Because validation is an ongoing process, it is important to make the existing validation benchmarks available to vendors, authors, and users in a form that is easy to use, reliable and accurate. Steps have been taken towards developing a computer database of benchmarks [32].

The IEA exercise has shown that it is possible for programs to be evaluated in a reliable and unbiased manner whilst, at the same time, involving the program authors and vendors. It has also shown that there is a broad consensus about how empirical validation should be undertaken and how to collect and identify reliable data. An independent accreditation system could therefore be developed for DSPs. This would be invaluable for enhancing the credibility of thermal modelling in general and for giving users confidence in the abilities of specific programs.

13. Conclusions

(a) The empirical validation work in IEA Annex 21/Task 12 was the largest validation exercise of its type ever undertaken. It included virtually all the state-of-the-art detailed thermal simulation programs (DSPs) currently used within the member countries, and some from non-member countries. All were used by either the program developers, vendors or an expert user. In total, 25 combinations of DSP and user participated.

(b) Empirical validation is both costly and technically complex, much more so than other forms of verification (such as intermodel comparisons or analytic tests). It requires careful planning, considerable measurement expertise, the application of rigorous quality assurance procedures, and detailed uncertainty analysis.

(c) An empirical validation package, which can be used by others, has been produced. It contains a site handbook, a validation guidebook, and a data diskette; these describe five empirical validation benchmarks. This has been thoroughly tested and approved by the IEA.

(d) The six test rooms used in the study are a compromise between the needs for both realism and experimental accuracy. They stressed the algorithms in programs which model heat transfer through the building fabric. The benchmarks were refined during the validation exercise and are an exemplar of how such benchmarks should be documented.

(e) The IEA participants were unanimous in recommending that a well-designed empirical validation exercise should be undertaken in two phases. The first phase being conducted blind, i.e. without knowledge of the actual measurements, and the second phase being open, i.e. with all available measurements being given to the modellers.

(f) A blind phase gives 'added value': it can provide a clear, clean snap-shot of program capabilities when they are used in a similar way to that which is adopted for real design problems; it can identify programs which may be in error and the possible areas of weakness; and the ambiguities which may be introduced when modellers are able to fit predictions to measurements are avoided.

(g) If blind validation is attempted it is recommended that a feedback mechanism is set up to assist in resolving problems. The use of a hotline (which must provide a rapid response) and newssheets is one approach. All the IEA participants who chose to comment felt this was a good approach.

(h) The open phase of an empirical validation exercise enables program users to explore, and hopefully to understand, the reasons for any divergence between the predictions of their program and the measurements.

(i) The total measurement uncertainty in the IEA experiments is indicative of that which will be obtained in carefully designed experiments in passive rooms. The total uncertainty band width for the 10-day heating energy was about 16% and for the peak air temperatures was 6°C. This was much less than the range in the program predictions. Well-conducted empirical validation experiments can, therefore, provide a powerful test of thermal programs and can identify programs which may be in error.

(j) Six of the programs predicted 7-day heating energy demands within the measurement error band for both the opaque and double-glazed test rooms. Seven of the programs predicted peak temperatures in the heated rooms that were within the error band.

(k) Seventeen of the programs underpredicted peak temperatures in the free-floating rooms and only four programs produced predictions which lay within the error band.

(l) The results do not support the view that interprogram variability is reduced if DSPs are used to predict differences in the performance of two buildings, or the effects of changing the design of one building, rather than the absolute performance of a single building.

(m) The empirical validation exercise has identified programs which performed well overall and, as a result, the credibility of these programs has been enhanced. It has also identified some programs which are highly likely to contain significant internal errors, and caution should be exercised when using these programs to assess the performance of real buildings.

(n) The work gave an insight into the type and source of internal errors in some programs. As a direct result, some programs have been improved, and the inherent strengths and weaknesses of many more have been clarified.

(o) The results of well-founded empirical validation exercises are likely to attract considerable attention. Particularly when, as in this exercise, the data collection and analysis is conducted by a neutral third party with the program authors/vendors/experts producing blind predictions. The results are of great interest: (i) because well designed and managed validation exercises are rare; and (ii) because they address the 'bottom line' issue—how well can the programs predict what will actually happen?

(p) Developers and vendors should make full use of the validation benchmarks which already exist to ensure that their programs are as reliable as possible, and to make clear the limitations of their program to potential users. For example, if a program cannot model a radiant heating system, users should be advised not to apply it to situations in which this is required.

(q) This exercise has shown that it is possible for thermal programs to be evaluated in a reliable and unbiased manner, whilst involving the program authors and vendors. It indicates that an independent program accreditation system could be developed. Such a system would be valuable to enhance the credibility of thermal modelling in general and to give users confidence in the abilities of specific programs.

(r) There seems to be little research work directed at improving the absolute accuracy of detailed thermal simulation programs. Participants noted the following as areas where further research could be useful: surface coefficients; heater dynamics; room heater interaction; and glazing systems. There is also a need for more very high quality data sets from which validation benchmarks can be created.

Acknowledgements

The authors are grateful to the UK Department of Environment for funding this work. We wish to thank the joint IEA Annex 21 Subtask C, and Task 12 Subtask B members for their valuable contributions and support: Vittorio Bocchio/Augusto Mazza, Politecnico di Torino, Italy (BLASTv3.0); Pascal Dalicieux, EDF, France (CLIM2000-v1.1); Tapio Haapala/Timo Kalema, Tampere University of Technology, Finland (TASEv3.0); Shirley Hammond, BRE (SERI-RESv1.2); Michael Holtz, Architectural Energy Corporation, Colorado, USA; Ron Judkoff, National Renewable Energy Laboratory, Colorado, USA; Foroutan Parand, BRE (TRNSYSv12.1 and 13.1); Eduardo Rodriguez, Escuela Superiore Ingenieros Industriales, Seville, Spain (S3PAS-v2.0); and Peter Verstraete/Rik van de Perre, Vrije Universiteit Brussel, Belgium (TRNSYSv13.1).

We are also indebted to all those who participated, without any dedicated funding, by running their detailed thermal simulation programs and by providing critical comments on our documentation and reporting. Without these contributions the work would have been far less comprehensive: Lorenzo Agnoletto, Instituto di Fisica Technica, Udine, Italy (WG6TCv1992); Don Alexander, University of Wales College, Cardiff, UK (HTB2v1.10); Angelo Delsante, CSIRO, Australia (CHEETAHv1.2); Martin Gough/Alan Jones, EDSL (TASv7.54); Mike Holmes, Arup R and D (ENERGY2v1.0); Kjeld Johnsen, SBI, Denmark (TSBI3v2.0); Mike Kennedy, ECOTOPE, USA (SUNCODEv5.7); Brian Miller/Doug Hittle, Colorado State Universtiy, USA (BLASTv3.0); Peter Moors, DMU (TASv7.54); Peter Pfrommer, FHT Stuttgart, Germany (HTB2v1.2); Paul Strachan, ESRU (ESP-rv7.7a); Glenn Stuart, ASL Sterling (ESP + v2.1); Andrew Tindale, FACET (3TCv1.0 and APACHEv6.5.3); Jeff Thornton, University of Wisconsin, USA (TRNSYSv13.1); Maria Wall/Petter Wallenten, Lund University, Sweden (DEROBvlth); and Fred Winkelmann, LBL, USA (DOE2v1E).

The permission of ETSU (UK) for the use of data from the Energy Monitoring Company's test room is gratefully acknowledged.

Appendix A

A.1. Results and method of calculating the total uncertainty bands

In empirical validation, it is important to account for the uncertainty in the experiment in order to assess whether the difference between the predictions and the corresponding measurements are due to the inherent (experimental) uncertainty or to internal error(s) in the program(s) combined, perhaps, with mistakes made by the program user. Uncertainty analysis played a key role in this research, and this sets it apart from most previous studies.

The total uncertainty has two components: one due to measurement errors—which are relatively easy to quantify; and a second due to uncertainties in the program input data—which is more difficult to calculate. The difficulties arise because of the large number of uncertain inputs to DSPs and because these uncertainties must be propagated through each program to assess the impact on predictions.

Fortunately, a great deal of work was done in this area in earlier UK studies [7,33]. Based on this work, and preliminary studies using SERIRES [34], it was clear that only a limited sub-set of all the input parameters were significant for the EMC rooms. This was because some values were accurately known and for others the small uncertainties being introduced were swamped by the much larger uncertainties introduced by other parameters. The uncertainty in the significant parameters was defined such that there was only a

Table 7
Summary of predictions and their variability, and the measurements and their uncertainty

	Heated, October Period									**Free-floating, May Period**						**Sfvr**	
	Double glazed			**Single Glazed**			**Opaque**			**Double glazed**		**Single Glazed**		**Opaque**		**Oct.**	**May**
Program	E [MJ]	$\hat{T}$ [°C]	$\check{T}$ [°C]	E [MJ]	$\hat{T}$ [°C]	$\check{T}$ [°C]	E [MJ]	$\hat{T}$ [°C]	$\check{T}$ [°C]	$\hat{T}$ [°C]	$\check{T}$ [°C]	$\hat{T}$ [°C]	$\check{T}$ [°C]	$\hat{T}$ [°C]	$\check{T}$ [°C]	[MJ]	[MJ]
WG6TCv1992 (Udine, I)	-	-	-	-	-	-	-	-	-	*33.8*	*13.4*	*33.7*	*11.5*	*18.0*	*10.2*	-	*84.0*
TSBI3v2.0 (SBI, DK)	*80.8*	*37.3*	*12.0*	*95.8*	*36.5*	*9.5*	*111.1*	*30.0*	*14.1*	*29.1*	*12.7*	*29.0*	*11.2*	*16.6*	*9.4*	*80.9*	*81.1*
DOE2.1E (LBL, US)	65.6	40.8	11.7	69.3	41.2	10.3	100.7	30.2	11.5	30.3	11.5	31.7	10.6	16.1	7.6	83.5	83.3
TASv7.54 (DMU/EDSL, UK)	*83.3*	*36.5*	*10.8*	*97.7*	*35.6*	*7.1*	*109.4*	*30.0*	*12.9*	*28.5*	*11.6*	*30.4*	*11.0*	*16.1*	*8.3*	*72.5*	*79.6*
ENERGY2v1.0 (Arup, UK)	70.9	37.7	11.3	78.6	38.6	9.4	99.8	30.0	12.6	28.4	12.0	30.8	11.0	16.5	9.1	83.8	81.3
CHEETAHv1.2 (CSIRO, AUS)	85.2	35.1	10.9	103.1	34.0	8.1	103.0	30.0	12.7	29.2	12.1	29.1	10.4	16.7	9.3	71.9	79.4
3TCv1.0 (Facet, UK)	85.1	36.3	12.3	102.2	34.5	10.5	117.1	30.0	14.1	27.0	12.2	27.3	11.3	15.9	9.1	76.0	79.1
APACHEv6.5.2 (Facet, UK)	86.1	36.3	12.6	102.1	35.4	10.3	118.6	30.1	14.8	26.9	12.1	27.5	11.0	15.8	8.9	76.0	79.1
HTB2v1.10 (UWCC, UK)	94.4	33.3	9.8	110.3	32.7	8.1	108.2	30.1	13.0	26.4	10.8	27.1	9.9	16.6	8.3	67.0	79.0
HTB2v1.2 (FHT, GER)	*82.1*	*36.0*	*10.3*	*94.8*	*35.2*	*8.5*	*103.9*	*29.7*	*12.8*	*28.4*	*12.1*	*28.7*	*11.1*	*18.2*	*9.3*	*75.1*	*79.5*
CLIM2000v1.1 (EDF, F)	83.3	41.4	10.5	98.8	38.4	8.6	108.9	29.9	12.4	30.8	11.9	29.2	10.0	15.0	7.7	73.9	80.4
DEROBvlth (Lund, S)	57.3	35.4	9.3	73.1	34.7	7.0	82.6	30.0	10.6	35.0	12.6	35.0	10.8	17.8	9.6	68.1	84.7
S3PASv2.0 (Sevilla, E)	78.0	38.8	12.1	89.5	38.2	10.0	105.7	30.0	13.7	29.1	11.9	30.4	11.0	17.4	9.4	79.1	79.0
BLASTv3lvl143 (CSU, US)	83.8	36.4	10.2	99.1	35.3	8.6	123.4	30.0	11.2	28.1	11.1	28.4	9.9	16.8	8.9	76.5	81.3
BLASTv3lvl203 (Torino, I)	*68.5*	*41.1*	*13.3*	*80.2*	*40.3*	*11.8*	*97.3*	*30.1*	*14.5*	*32.5*	*13.1*	*32.5*	*11.6*	*16.9*	*10.0*	-	-
TASEv3.0 (Tampere, FIN)	79.2	39.4	10.6	102.9	36.5	7.7	101.1	30.1	13.0	32.7	12.6	30.8	10.4	17.1	9.6	75.6	78.1
TRNSYSv13.1 (UWISC, US)	57.1	41.5	12.9	65.7	42.4	11.4	87.3	30.0	13.8	29.1	12.3	30.2	10.8	16.8	9.7	82.5	78.5
TRNSYSv13.1 (Brussels, B)	62.8	36.9	12.3	70.0	38.4	10.6	88.3	30.0	13.2	27.3	12.0	29.2	10.7	16.8	9.6	84.1	81.9
TRNSYSv13 (BRE, UK)	66.6	36.1	11.6	78.2	36.6	8.7	93.4	30.0	12.9	27.7	12.4	29.6	11.0	17.4	9.9	82.4	78.4
TRNSYSv12 (BRE, UK)	71.2	34.7	11.3	83.8	35.0	8.5	93.8	30.0	12.9	27.8	12.5	29.7	11.1	17.4	9.9	72.6	79.4
SUNCODEv5.7 (Ecotope, US)	*80.1*	*36.4*	*10.7*	*94.8*	*35.3*	*8.6*	*111.9*	*30.0*	*12.5*	*28.8*	*11.9*	*29.2*	*10.6*	*16.7*	*9.0*	*72.8*	*79.3*
SERI-RESv1.2 (BRE, UK)	*82.2*	*36.8*	*11.1*	*95.7*	*36.1*	*8.9*	*103.8*	*30.0*	*13.2*	*28.9*	*11.8*	*29.5*	*10.6*	*16.7*	*9.2*	*73.5*	*77.6*
ESP+v2.1 (DMU/ASL, UK)	55.5	43.8	13.8	66.3	43.6	12.2	92.7	29.9	14.3	31.7	13.6	32.0	11.9	15.5	8.4	76.8	80.7
ESP-Rv7.7a (ESRU, UK)	69.5	40.3	12.4	78.5	39.9	10.7	100.3	30.3	12.8	28.9	11.7	29.5	10.8	15.4	7.9	77.7	78.7
ESPv6.18a (DMU, UK)	61.1	42.7	14.0	72.3	42.7	12.0	94.7	30.0	14.0	32.6	13.5	33.3	12.0	15.7	8.4	69.4	77.5
Average prediction	74.6	38.0	11.6	87.6	37.4	9.5	102.4	30.0	13.1	29.6	12.2	30.1	10.8	16.6	9.1	76.2	80.0
Maximum prediction	94.4	43.8	14.0	110.3	43.6	12.2	123.4	30.3	14.8	35.0	13.6	35.0	12.0	18.2	10.2	84.1	84.7
Minimum prediction	55.5	33.3	9.3	65.7	32.7	7.0	82.6	29.7	10.6	26.4	10.8	27.1	9.4	15.0	7.6	67.0	77.5
Standard deviation of predictions	10.7	2.7	1.2	13.5	2.9	1.5	10.0	0.1	1.0	2.2	0.7	2.0	0.6	0.8	0.7	4.9	1.9
Range in predictions	38.9	10.5	4.7	44.6	10.9	5.2	40.8	0.6	4.2	8.6	2.8	7.9	2.6	3.2	2.6	17.1	7.2
Range as % of average	51	-	-	51	-	-	40	-	-	-	-	-	-	-	-	22	9
Measured value	89.3	37.8	11.9	-	-	-	117.1	29.8	14.6	31.0	12.2	32.6	12.1	16.8	9.2	81.1	82.8
Upper uncertainty bound	92.7	40.5	13.9	-	-	-	122.3	30.2	16.4	33.4	13.6	35.0	13.6	17.5	10.0	85.5	88.8
Lower uncertainty bound	78.1	36.5	11.5	-	-	-	105.3	29.4	14.0	29.6	11.6	31.2	11.6	15.7	8.6	76.7	76.8
Error band width	14.6	4.0	2.4	-	-	-	17.0	0.8	2.4	3.8	2.0	3.8	2.0	1.8	1.4	8.8	12.0
Width as % of measurement	16	-	-	-	-	-	15	-	-	-	-	-	-	-	-	11	14

italics indicate new Phase 2 results, remainder from Phase 1.

E = total heating energy consumption over 7 days; $\hat{T}$ = maximum temperature; $\check{T}$ = minimum temperature; Sfvr = total South facing vertical solar irradiance