A FULL-SCALE EMPIRICAL VALIDATION STUDY APPLIED TO THERMAL SIMULATION PROGRAMS

Paul Strachan[1], Jon Hand[1], Katalin Svehla[1], Ingo Heusler[2], Matthias Kersken[2]

[1]Energy Systems Research Unit, University of Strathclyde, Glasgow, Scotland, UK
[2]Fraunhofer Institute for Building Physics IBP, Holzkirchen, Germany

ABSTRACT

A large-scale study for validating building energy simulation programs against measured data was undertaken within IEA ECB Annex 58 "Reliable building energy performance characterization based on full scale dynamic measurements". The study included 12 commercial and research programs from 15 organisations. The experimental data was gathered in two experiments conducted at the Fraunhofer IBP test site at Holzkirchen in Germany. The paper briefly summarises the methodology of the study, key aspects of the experiments and the main research outcomes. An important objective of the paper is to present lessons learned from the study, from the perspectives of experimentalists, modellers and data analysts, which may be of benefit in future empirical validation exercises.

INTRODUCTION

Energy regulations for new and retrofitted buildings require increasing attention to energy efficiency as part of the worldwide drive to reduce carbon emissions and reliance on fossil fuels. Many authors (e.g. BRE 2015) have observed a so-called "performance gap" between the as-designed energy consumption and that experienced in practice, for which there are many reasons, including poor controls, poor workmanship, poor commissioning, different weather conditions, different operating conditions and user behaviour. Nevertheless, there is still a need for design tools to predict energy consumption for compliance purposes, for identifying energy and carbon efficient designs, and to ensure good indoor environmental quality. Except for relatively simple buildings, this involves the use of simulation programs to represent dynamic response and to predict overheating risk, indoor air quality, lighting comfort etc. There needs to be evidence that the programs used to make these predictions are both adequate and appropriate.

In spite of significant international effort, it is perhaps surprising that there are still questions regarding the reliability of commercial and research programs to predict energy and environmental performance accurately. Certainly, many papers claim, following a limited monitoring exercise, that the program being used has been "validated", but in reality this may only be true for a particular building type, climate, operational regime, construction type etc. There are also questions regarding how extensive and accurate these monitored datasets are, whether all the relevant influencing parameters have been measured, and whether in fact the level of agreement can be classified as "good". Other programs have been claimed to be "validated" by comparing with standard benchmark tests, particularly BESTEST which was developed within IEA Annex 21 (1995) but which did not include measured data. The danger is that empirical validation exercises yield inconclusive results because of too much uncertainty in inputs and/or experimental measurements to be useful for diagnosing sources of disagreement between simulation results and empirical data.

Because of the complexity of detailed thermal simulation programs, a detailed validation methodology is necessary. The overall methodology is well-established and comprises elements of analytical, inter-program comparison and empirical test (Judkoff and Neymark 2006, Jensen 1995). Although analytical and comparative tests have an important role, there is a need for empirical validation, to ensure that the physical-based models at the heart of simulation programs represent real-world performance.

At the start of IEA Annex 21, a comprehensive worldwide review of existing datasets suitable for empirical validation was reported (Lomas et al 1997). The majority of the datasets investigated were found to be of limited use for program validation, primarily because of missing monitored data of key parameters. There have subsequently been some successful large-scale international projects for empirical validation purposes, but these have been at component level, e.g. for testing micro-cogeneration models in IEA Annex 42 (2007), or on outdoor test cells, e.g. IEA Annex 43 (2007). The reason why few large-scale whole building empirical validation tests have been undertaken is a combination of the lack of suitable test facilities, cost and time.

The criteria for suitable validation-quality experimental datasets are exacting. Building on previous work by Lomas et al (1997) and Judkoff et al (2008), the following are important requirements:

- The test building should be unoccupied because of the difficulty of separating uncertainty in calculating building performance from uncertainty in occupant behaviour.

- There should be options for heating/ cooling and an accurate/flexible control system.
- All program data inputs should be measured (building parameters, schedules etc).
- Detailed weather data must be measured on-site to record all the weather inputs required by the simulation programs.
- The instrumentation system should be comprehensive and reliable with traceability for all sensor calibrations. Data recording should be sub-hourly. The experimental team should be able to respond to modeller requests for additional sensors to be added.
- Data should be recorded for the overall building to enable the calculation of the overall building heat loss and effective capacity.
- As far as possible, data should be recorded for individual heat transfer paths. It can help to identify causes of differences between measurements and predictions, but it becomes more difficult when moving from test cells to larger scale buildings. Alternatively, if two essentially identical buildings are available, side-by-side experiments can be devised to focus on individual heat transfer processes.
- Infiltration and ventilation should be controlled or measured.
- Measured data should include uncertainty estimates.
- The experimental team should be experienced in detailed high quality dataset collection and the test facility must be well documented.
- The test facility should be available for extended test periods to cover a range of weather conditions.

TWIN HOUSE EXPERIMENTS

High quality outdoor test facilities are increasingly being constructed, as documented in Janssens (2014). At the start of IEA Annex 58 (2015), a review of these test facilities was undertaken against the criteria identified in the previous section in terms of suitability for empirical validation and availability. The Twin Houses (one of which is shown in Figure 1) operated by the Fraunhofer Institute IBP at Holzkirchen in Germany were selected. These identical detached houses were equipped with extensive measurement and control equipment, and there is comprehensive weather monitoring on site. Previous testing showed infiltration and heat loss coefficients of the houses were within measurement error. External walls are externally insulated with U-values in the range 0.20 to 0.28 W/m²K. Windows are double glazed with a glazing U-value of 1.2 W/m²K and with electric external roller blinds. The focus for the validation experiment was the ground floor (7 rooms), with air temperatures in the cellar and the attic measured and supplied to modelling teams as boundary conditions. A constant flow rate balanced mechanical ventilation system was operated.

Two experiments were undertaken:
Experiment 1: A side-by-side experiment in August and September 2013 in which the house were operated identically, except that blinds were down on the south façade on one house and up on the other for most of the experiment.
Experiment 2: An experiment on one of the houses in April and May 2014 with fixed temperatures in boundary spaces, reduced mechanical ventilation rate and some additional temperature and heat flux sensors.

In both cases the testing was split into different phases:
Period 1: Initialization phase (7 days) with steady conditions
Period 2: Room air temperatures constant (7 days)
Period 3: A pseudo-random heating sequence in the living room (14 days) to ensure the solar and heat inputs are uncorrelated
Period 4: Room air temperatures constant (7 days)
Period 5: Free-floating period (7 days)

Full details of the specification and experiments are given elsewhere (within IEA Annex 58 (2015) and in Strachan et al (2015). Only key aspects are given here.

Figure 1 View from South of one of the Twin Houses

METHODOLOGY

The overall empirical validation methodology applied in this study was similar to that employed in previous IEA validation studies (e.g. Loutzenhiser et al 2007, Kalyanova et al 2009, Lomas et al 1997). The steps were as follows:

Experimental design: specify test sequences, experimental configuration and monitoring scheme.
Experimental set-up: calibrate and install sensors; program the heating and ventilation.
Experimental specification: develop the building and test specification required for modelling.
Experiment: undertake the experiment and process the experimental data.
Blind validation: modellers predict internal conditions using the experimental specification, measured climate data and operational schedules but without knowledge of internal conditions. They submit

predictions and modeller reports with details of the programs used and assumptions made.

First stage analysis: compare predictions against experimental data for internal temperatures and heat fluxes.

Re-modelling: disseminate all measured data. Modelling teams are encouraged to investigate differences between measurements and predictions and resubmit predictions and updated reports. Only changes that correct user modelling errors or alter a modelling assumption (with documented rationale) are allowed.

Final analysis and reporting: provide definitive documentation of the analysis and outcomes.

Archiving of high quality data sets: for use by program developers and testers.

It was not possible to undertake a blind validation in Experiment 2, because at that stage measured data from Experiment 1 had been made available to modelling teams, so they had an opportunity to compare with measured data. Nevertheless, the measured temperature data of the free-float period of Experiment 2 was initially withheld so an element of blind validation was maintained.

MODELLING

An important part of a validation exercise is to ensure there are a large group of modellers willing to participate – this helps to ensure a variety of simulation programs are used and that the specification is thoroughly scrutinised. It is also important that a mechanism is established whereby questions can be raised, with answers circulated to all teams. In this study, it proved useful for clarifying some aspects of the specification; it also led to some additional measurements of dimensions and surface absorptivity, and more information on thermal bridges. The analysis of the first experiment also resulted in additional sensors to measure temperature stratification in more of the rooms and external wall heat fluxes in the second experiment, and additional measurements and calculations of ventilation duct losses.

In the blind validation part of Experiment 1, 21 datasets were submitted from 13 organisations in 10 countries with 11 distinct simulation programs. For the re-modelling, after data was released, there were 18 datasets (including 4 new ones that were not submitted in time for the blind validation) from 15 organisations with 12 programs. For Experiment 2, 13 datasets were submitted from 12 organisations with 10 programs. Tables 1 and 2 list the participating programs and organisations. As can be seen, this was a well-supported exercise, with a good mix of organisations using both commercial and research programs.

Table 1: Participating organisations

CIEMAT	University of Liege
Czech Technical Univ.	Politecnico di Milano
Danish Technical Univ.	IES
University of Gent	University Innsbruck
Hong Kong City Univ.	University of Leuven
University of Strathclyde	HS Stuttgart
Equa Solutions	RWTH Aachen
Fraunhofer Institute for Building Physics	

Table 2: Programs

TRNSYS	4
Modelica	4
EnergyPlus	2
ESP-r	2
EES	2
INSEL	2
Matlab	2
eQuest	1
IDA-ICE	1
Wufi	1
IESVE	1
Dynbil	1

RESULTS

It is not possible to present all the results of the comparison between experimental data and modelling predictions in this paper. They will be presented in detail in the final IEA Annex 58 report. Key results from Experiment 1 have been presented in Strachan et al (2015).

Important considerations are:

Estimation of experimental uncertainty of variables used in the comparison. In these experiments, the comparison was made with internal air temperature predictions for the pseudo-random heat input and free-float periods, and with heat inputs for the constant temperature setpoint periods. The heating power accuracy is ±1.5 %; individual calibrated shielded temperature sensors have an accuracy of better than 0.15 °C. However, stratification was observed in rooms as discussed later in this paper, so modelling choices made in selection of appropriate room average temperatures may account for offsets of perhaps 1 °C to 2 °C, especially during higher heating power inputs.

Selection of comparative metrics. There are two categories of comparative metrics. The first is a timestep comparison; this is usually a time series display of all or part of the test sequence. It is largely based on a visual comparison and is useful for observing general trends and time shifts between measurements and predictions. The other category involves comparative statistics of the output variables (mean values, peak values, integrated heating energy input). These are useful for summarising the overall goodness of fit, but they do not provide information in cases where there is good fit over part of the period and poor fit elsewhere.

In the analysis carried out, the two main metrics used to summarise the level of agreement were as follows.
1. The magnitude fit: this was defined as the absolute average difference between measurement and prediction for each experimental period in each room.
2. The shape fit: the level of correspondence in the shape of the profile was given by Spearman's rank correlation coefficient (Kendall and Gibbons 1990) between predictions and measurements.

In the experiments here, given the high levels of instrumentation, it is also possible to compare measurements with predictions of more focused areas of the building (e.g. surface temperatures and heat fluxes) and heat transfer processes (e.g. longwave and shortwave fluxes on different orientations). In addition, differences in predictions for the two houses can be compared with differences in measurements.

Assessment of goodness-of-fit. There are no definitions of "acceptable" bands within empirical validation, so these need to be guided by experimental uncertainty and subjective judgement. In the analysis undertaken to date, agreement of average absolute temperatures within 1°C is classed as "good" (see Table 3), as is the heating power within 100W for the constant temperature periods. However, it is recognised that it would be useful to establish a more rigorous basis for categorising the level of agreement.

A few example results are shown in Figures 2 and 3, and Table 3.

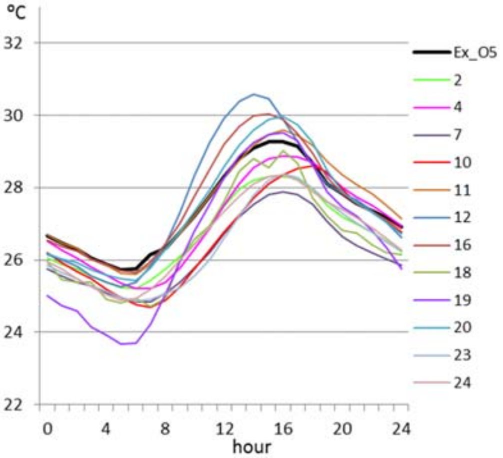


Figure 2: One day in free float period for Experiment 2: predictions versus experiment (Ex_O5)

DISCUSSION OF RESULTS

Two of the important findings from the study were as follows.

User modelling errors. These were identified by all teams who submitted at both the blind validation and re-modelling stages: in some cases, these were minor; in other cases, they were substantial. Examples of user errors reported include:

- The thermal bridges between the internal walls and ceiling/floor were not modelled.
- An error was found in modelling of the control of the heaters. It was thought that the heater power was limited to 500W throughout, not just the pseudo-random heat input period.
- There was faulty control of the blinds due to mistake in the modelling.
- For the output, the actual value for the end of each hour was sent, rather than mean values as requested.
- By mistake, both supply and exhaust airflow were modelled in the corridor.
- The north wall in bedroom was modelled with the wrong construction type.

Predictive accuracy. For the blind validation, and considering the absolute average difference in temperature measurements and predictions in the periods of pseudo-random heat injections and free-float, no program predicted temperature in every room and every period within 1°C although two simulations came close. Ten out of the 21 datasets submitted predicted the absolute average difference in temperature within better than 2°C of the measured values for all the main rooms for both houses (with the exception of the kitchen, as discussed later), as well as for the difference in temperatures between the two houses. For experiment 2, 8 out of 13 programs predicted within this limit. The results for heat inputs in the constant temperature heating periods are more variable – for example, only 5 out of 13 programs predicted the total heating energy in the southern rooms to be within 12% of the measured data.

LESSONS FOR FUTURE EMPIRICAL VALIDATION STUDIES

Perspective of Modellers

The key requirement for modellers is that the specification is clear and comprehensive.

From the point of view of the simulation tool user the Twin House experiments presented a number of challenges and highlighted issues of model design, quality checking, sensitivity studies and the discovery of patterns in performance data.

Firstly, model design needed to reflect the available information on the composition of the building, monitoring kit, and controls applied during the experiments. Many aspects of the buildings were specified at a greater level of detail that is typically found in commercial practice, as would be expected for a validation project. For example, all the constructional material thermophysical properties, thicknesses and surface properties were specified according to manufacturers' data or were measured. Therefore, this data had to be entered, rather than using standard constructional databases that form the basis of many commercial modelling exercises.

Secondly, there was much more measured data than is commonly available in modelling studies. It was necessary to create an appropriate climate file (preferably with the 10-minutely data to improve time resolution) in the particular format used by the program. It was also necessary to use time-varying heat inputs rather than regular scheduled inputs for part of the test sequence, and to use measured temperature setpoints for other parts.

Producing QA reports to confirm details of form, composition and building use, together with getting colleagues to check the model, proved invaluable for confirming the inputs. The modeller reports indicated that different modelling teams undertook varying degrees of QA. In some cases a thorough report was produced with all the model details and assumptions; in other cases it seemed that no checking was done except by the individual modeller.

There were also instances where user judgement was required to decide on the best modelling approach. Four examples are illustrative:

a) the external roller blind system used in the experiments is a common system, but it is difficult to model conditions between the blind and the glass and possible air movement regimes, so modellers were left to decide when the blinds were down whether to consider the air gap as sealed or whether to model some ventilation. More research is needed to improve understanding of and assessment of roller blind systems.

b) modelling of thermal bridges to the outside is represented in several tools (typically as a linear thermal transmittance, or by adding additional constructions to account for the extra heat loss). In the Twin Houses, there were also significant internal thermal bridges between the ground floor experimental rooms and the basement and roof space. Again, modelling teams adopted different approaches varying from ignoring these thermal bridges to adding additional constructions to represent these heat loss paths.

c) many programs assume that buildings are largely static with simple scheduling of heating, cooling, ventilation, lighting etc. In these validation experiments, the test regime imposed a number of step changes to controls, blinds and internal gains at precise times on different days. Constraints on scheduling facilities in some programs made it difficult to coordinate these changes.

d) most programs assume well-mixed air in each thermal zone. In these experiments, air temperature was measured at different height in some rooms. Different modellers made different assumptions – either averaging, or taking the mid-point temperature as the setpoint. Stratification is commonly overlooked in simulation programs.

When the measured data was made available, it was possible to compare these with predictions. Visual inspection helped to identify any obvious misalignment of control periods, heat injections etc. However, identifying causes of discrepancies is difficult, and in most cases modellers who investigated this undertook ad hoc sensitivity studies to gauge the effect of modelling input assumptions (e.g. appropriate internal convection coefficients). This indicates the need for simulation programs to support users in undertaking structured sensitivity analyses, not only for validation studies, but also for calibration exercises that are becoming more common now that the availability of measured data of building performance is increasing.

The project also highlighted working practices in terms of model documentation. As the model evolves, it is necessary to document changes comprehensively to make subsequent QA easier.

Perspective of Experimental Team

The experiments carried out on the Twin Houses were designed based on the previous experience of the experimental team. Even with the significant effort that was invested in the design and implementation of this experiment, some new lessons can be learned for future empirical validation studies.

Sensor calibration is critical to ensure reliable results. It is not sufficient to trust the specifications of a device since the quality procedures of the manufacturer are often not known, or damage may have occurred during transport or prior use. Some devices can be calibrated in-house; others need to be tested by a professional calibration laboratory to ensure traceability. Certain sensor positions may require special attention related to accuracy, for example when measurement of thermal power requires the detection of small temperature differences in the case of some ventilation systems, duct heat losses or hydronic heating systems.

Attention should also be focused on the physical installation. Air temperature sensors need to be equipped with radiation shields reducing the influence of solar and thermal radiation on the sensor. In the case of electrical heating, the power meters should be calibrated in the installed condition if possible. The power calibrator should be inserted into the power line where it enters the room to include cable resistance losses in the supply room but not those in the test room since the latter ones add to the room’s heat load. For hydronic systems, care must be taken that the flow meters do not start to drift during the experiment because of dirt and/or metallic residues in the heat carrier fluid. To prevent this, the heat-carrying medium can be replaced by completely demineralized water or the flows should be measured with two different independent measurement principles. Mechanical ventilation systems are sensitive to external wind so the mass flows to and from the rooms need to be measured.

The information available for buildings usually concentrates on the external components, because these are the most relevant for the energy consumption of the entire building that traditionally is of primary interest. When the attempt is made to model a building with validation quality, also internal processes such as airflow and heat fluxes become important for inter-room heat transfer. For example, a small temperature difference can cause the heating in one room to start.

When this heat is transported to the next room by air movement, the heater in the second room will stay inactive. This example relies on knowledge of air movements between the rooms, especially through open internal doors, but this air movement was not monitored in these experiments. It is recommended that multi-zone tracer gas measurements, using different tracer gases in each room, are included in future experiments to check on this heat transfer path.

Even if all desired details on the measurement facility are available and included in the specification, there may still be deviations e.g. from poor craftsmanship or the presence of humidity inside the construction. A co-heating test as part of the experimental design can help with an overall heat loss coefficient (Bauwens and Roels 2014) and serve as a reference for the building's performance.

When air ducts or hydronic pipes pass through rooms, there are two possibilities for considering the associated heat losses. If they are well insulated, an estimation of losses is required during the experimental design phase to ensure an unacceptable uncertainty is not introduced. If the losses are expected to have a significant influence on the measurement result, instrumentation needs to be installed to measure these losses and/or detailed modelling undertaken of the duct losses, as was done for the supply air duct through the kitchen in the experiment described in this paper.

In these experiments, significant stratification (Figure 4) was observed in the rooms dependent on the heating power, the test period, interaction with internal air movement and the positions of the mechanical ventilation inlet and extracts. There are two possible approaches to deal with stratification. It is a common occurrence, so it could be expected that simulation programs should be able to model it (although this is not usually the case in practice in whole building simulation programs), and additional sensors measuring the air temperature distribution could be added to check on the modelling. Alternatively, an additional experimental phase with homogenous room air temperatures provided by mixing fans could be introduced for part or all of the experiment, as is done in co-heating tests. When fans are used during the experiment their electric consumptions needs to be taken into account. The direction of airflow should be chosen to have minimal impact on the air speed at the external walls, to minimize the influence on the internal convective heat transfer coefficients. Textile hoses can additionally be used to diffuse the airflow from the fans homogeneously. One suggestion to study the effect in detail is to undertake a side-by-side experiment: one test house can be equipped with fans while in the other one, natural stratification is allowed to occur.

It is important to maintain the integrity of the detailed monitoring system over extended periods. In future experiments, it is recommended that an uninterruptible power supply be used. To guard against other failures such as malfunctioning control programs or hardware failures, frequent checking of the recorded measurement data, at least on daily basis, is required. In case gaps in the measurement data need to be filled, these time periods and the methods chosen to do this must be included in the experimental specification.

It is recommended that the test specification is distributed to the participating modelling teams before the measurements are started, together with some representative climate data for the site. Even if the experiment is designed by a very experienced team, multiple experienced modelling teams will raise additional questions, not thought of during experimental design. The designers cannot be specialists in all participating software programs, which can often have different input requirements. If modelling teams can prepare a model and run preliminary simulations, any questions can be raised in time to add additional sensors, check dimensions, etc.

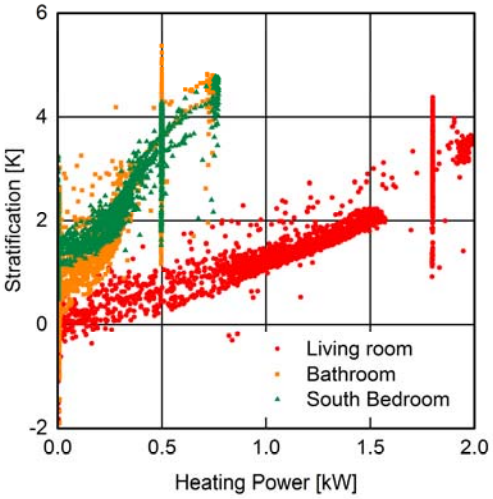


Figure 4: Stratification as a function of applied heating power

Perspective of Analyst/Co-ordinator

In a large empirical validation study as described in this paper, planning and co-ordination are essential. Key considerations are:

Modelling teams. Ensuring there are a sufficient number of modelling teams – having many modellers scrutinising the specification and running simulations is essential for developing a useful validation dataset.

Experimental design. It is essential to ensure the test house is subject to a comprehensive test schedule, and that experimental uncertainties are reduced to a minimum. This requires sensitivity studies with the proposed test configuration and schedules in advance of the experiment to determine key parameters that need to be measured and the level of accuracy required. For example, in this experiment the initial simulations were used to show that ground reflectivity should be measured; they were also used to set the magnitude of the heat injections and ventilation flow rates to produce useful variation in internal temperatures without excessive overheating.

Question and answer "hot-line". Inevitably, modellers have questions on the specification, so the co-ordinator and experimental team need to be responsive by providing additional detail for dissemination to all modelling teams. This proved useful in the Twin House experiment for improving the overall specification.

Standardised format. Given the large number of submitted datasets, it is necessary that modelling teams are required to follow the specified format, to make the data processing efficient and to reduce the possibility of errors which can be difficult to detect with such large quantities of data.

Good quality modelling reports. For analysing the results, modelling reports should cover details of all assumptions made, a summary of the main methods in the program used for modelling the various heat transfer paths, the experience of the modeller/modelling team and any modelling difficulties encountered. In this study, the modelling reports were of varying quality, from rather terse descriptions to high levels of detail. This made it difficult to discern patterns between prediction accuracy and modelling detail, for example. Even more important, in the re-modelling stage after all measured data has been disseminated, the reports should detail changes made (e.g. correction of user input errors) to make it clear that no calibration or tuning has been undertaken with arbitrary input adjustments to improve the fit. It was difficult to monitor this in practice, and it is one aspect of the overall methodology that could be improved. Possibilities are to introduce some sort of peer review, where modelling teams using the same software would review each other's models. Alternatively, modelling teams could also be asked to submit all their models, so that others could potentially check what changes had been made.

CONCLUSIONS

There are many benefits of empirical validation studies. The resulting specification and datasets can be used by developers of new programs or by those extending existing programs to test their predictions; they can help to identify program deficiencies; and they can be used in training for program users. The experiments described in this paper have already been used for these purposes. In addition, software vendors can use them to provide evidence of prediction reliability.

Nevertheless, the time and effort required is substantial. It requires a high-quality test facility, an experienced experimental team, a good number of modelling teams and a range of simulation programs with users who are highly knowledgeable about the simulation programs they are using and have QA procedures in place.

It is believed the experiments summarised in this paper are a useful resource, and form a step up from experiments in small outdoor test cells. The experiments are already being used for model development (Masy et al 2015). There is also additional measured information (e.g. measured long-wave and short-wave radiation on different orientations) that has not yet been fully analysed. The experiments do have limitations – the buildings have a comparatively simple construction, they are unoccupied, and no systems were included in the experiments except for the mechanical ventilation system and electric heaters.

Some programs performed very well, even in the blind validation phase of Experiment 1 and free-float period of Experiment 2. It has been shown that it is possible for programs to predict to within 1°C to 2°C for the average temperature in buildings such as the Twin Houses, which have large thermal mass and high solar gains.

However, there were significant numbers of user errors even by experienced modellers. It proved difficult to disaggregate factors such as experience of the modeller, amount of QA undertaken by colleagues of the modeller, and the attention and time devoted to the study by the modeller. In addition, not all submissions can be classified as program validation: the full capability of programs was not always used (e.g. aggregating rooms, using combined surface coefficients …)

Certainly, more such experiments are needed on other building types, but it should be noted that large resources are required, both in time and money. Future studies are also needed that focus on the types and impacts of user errors on larger-scale building designs, with a view to informing program developers.

ACKNOWLEDGEMENTS

The UK authors acknowledge support from the Engineering and Physical Sciences Research Council (Ref: EP/K01532X/1). The German authors acknowledge support from the Federal Ministry for Economic Affairs and Energy (Ref: 03ET1144A).

REFERENCES

Bauwens G and Roels S, 2014. Co-heating Test: a State of the Art, Energy and Buildings, Vol 82, pp163-172, doi:10.1016/j.enbuild.2014.04.039

BRE, 2015. Bridging the Performance Gap. Understanding Predicted and Actual Energy Use of Buildings, Information Paper 1/15.

IEA Annex 21, 1995. Environmental Performance http://www.ecbcs.org/annexes/annex21.htm

IEA Annex 42, 2007. The Simulation of Building-Integrated Fuel Cell and Other Cogeneration Systems (COGEN-SIM), http://www.ecbcs.org/annexes/annex42.htm

IEA Annex 43, 2007. Testing and Validation of Building Energy Simulation Tools, http://www.ecbcs.org/annexes/annex43.htm

IEA Annex 58, 2015. Reliable Building Energy Performance Characterisation Based on Full Scale Dynamic Measurements, http://www.kuleuven.be/bwf/projects/annex58

Janssens A, 2014. State of the Art of Full Scale Test Facilities for Evaluation of Building Energy Performances. IEA Annex 58 Subtask 1 report, http://www.kuleuven.be/bwf/projects/annex58/index.htm

Judkoff R and Neymark J, 2006. Model Validation and Testing: The Methodological Foundation of ASHRAE Standard 140, Ashrae Conference, Quebec City, Canada

Judkoff R, Wortman D, O'Doherty B and Burch J, 2008. A Methodology for Validating Building Energy Analysis Simulations, Technical Report NREL/TP-550-42059, National Renewable Energy Laboratory.

Kalyanova O, Heiselberg P, Felsmann C, Poirazis H, Strachan P and Wijsman A, 2009. An Empirical Validation of Building Simulation Software for Modelling of Double Skin Façades, Proc 11th IBPSA Conference, Glasgow, pp27-30.

Loutzenhiser PG, Manz H, Felsmann C, Strachan P and Maxwell GM, 2007. An Empirical Validation of Modeling Solar Gain through a Glazing Unit with External and Internal Shading Screens, Applied Thermal Engineering, Vol 27, Issues 2-3, pp528-538.

Lomas KJ, Eppel H, Martin CJ and Bloomfield D, 1997. Empirical Validation of Building Energy Simulation Programs, Energy and Buildings, 26, pp252-275.

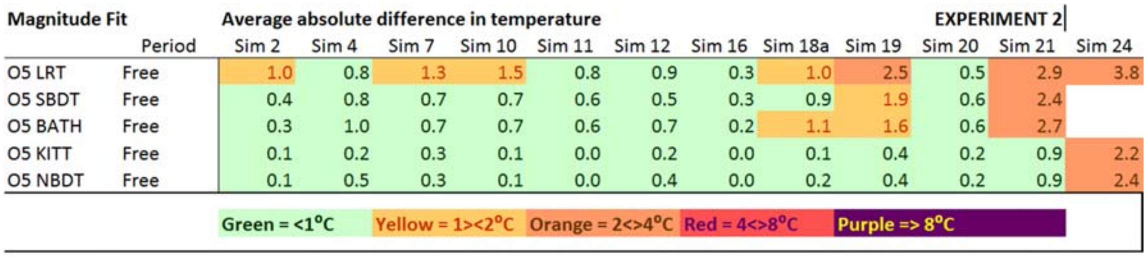
Masy G, Rehab I, André P, Georges E, Randaxhe F, Lemort V and Lebrun J, 2015. Lessons Learned from Heat Balance Analysis for Holzkirchen Twin Houses Experiment, 6th Intl Building Physics Conference, IBPC 2015, Torino.

Jensen SO, 1995. Validation of Building Energy Simulation Programs: A Methodology, Energy and Buildings 22(2), pp 133–144

Kendall M and Gibbons J D, 1990. Rank Correlation Methods, 5th edition, Edward Arnold, London, pp 69-77

Strachan P, Svehla K, Heusler I and Kersken M, 2015. Whole Model Empirical Validation on a Full-Scale Building, Journal of Building Performance Simulation,
doi: 10.1080/19401493.2015.1064480.

Table 3: Example of average absolute temperature differences between measured and predicted in the free-float period of Experiment 2, for living room, south bedroom, bathroom, kitchen and north bedroom



Magnitude Fit		Average absolute difference in temperature										EXPERIMENT 2	
	Period	Sim 2	Sim 4	Sim 7	Sim 10	Sim 11	Sim 12	Sim 16	Sim 18a	Sim 19	Sim 20	Sim 21	Sim 24
O5 LRT	Free	1.0	0.8	1.3	1.5	0.8	0.9	0.3	1.0	2.5	0.5	2.9	3.8
O5 SBDT	Free	0.4	0.8	0.7	0.7	0.6	0.5	0.3	0.9	1.9	0.6	2.4	
O5 BATH	Free	0.3	1.0	0.7	0.7	0.6	0.7	0.2	1.1	1.6	0.6	2.7	
O5 KITT	Free	0.1	0.2	0.3	0.1	0.0	0.2	0.0	0.1	0.4	0.2	0.9	2.2
O5 NBDT	Free	0.1	0.5	0.3	0.1	0.0	0.4	0.0	0.2	0.4	0.2	0.9	2.4

Green = <1°C Yellow = 1><2°C Orange = 2<>4°C Red = 4<>8°C Purple => 8°C

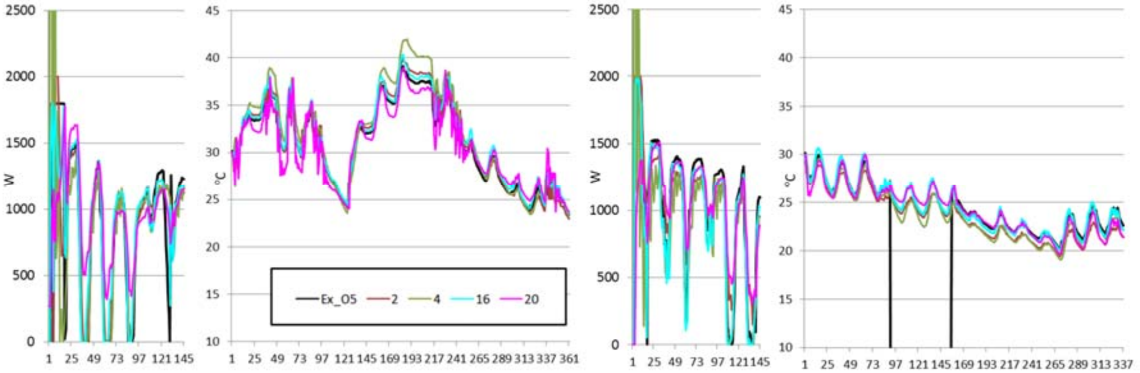


Figure 3: An example of four test periods of Experiment 2: from left to right, constant temperature period with heat input predicted; pseudo-random heat inputs with temperature predicted; constant temperature period with heat input predicted; and free float with temperatures predicted. The black line is the measured data; there is a period of missing data in the 4th period. Four example prediction datasets are shown.