CALIBRO: an R Package for the Automatic Calibration of Building Energy Simulation Models

Filippo Monari, Paul Strachan
University of Strathclyde, Glasgow, UK

Abstract

Bayesian probability theory offers a powerful framework for the calibration of building energy models (Bayesian calibration). The major issues impeding its routine adoption are its steep learning curve, and the complicated setting up of the required calculation. This paper introduces CALIBRO, an R package which has the objective of facilitating the undertaking of Bayesian calibration of building energy models. An overview of the techniques and procedures involved in CALIBRO is given, as well as demonstrations of its capability and reliability through two examples.

Introduction

Previous research has highlighted a significant gap between building energy performance predicted in the design stage and that achieved in the operational stage. This issue may be addressed by ensuring that building energy models are equally applicable to the operation stage by reconciling their predictions with field observations. Recently, significant research effort has been focused on the development of methods for tuning model parameters in order to improve predictions, that is the calibration of Building Energy Simulation (BES) models against measured data. A novel approach is to employ probabilistic emulators based on Gaussian Process Regression (GPR) (Rasmussen and Williams (2006)) in a quasi-Bayesian framework (Bayesian calibration). Such an approach has several advantages with respect to previously employed methodologies like manual iterative procedures and the employment of optimisation techniques.

Manual iterative procedures have been largely applied in the past (Clarke et al. (1993), Pedrini et al. (2002), Yoon et al. (2003), Westphal and Lamberts (2005), Raftery et al. (011a)) and Raftery et al. (011b). Their high dependence on the skills and expertise of the analyst makes them very heterogeneous. This factor together with the absence of a strong mathematical framework has led to criticisms about the consistency of this approach. The main issue is the inadequate consideration of modelling uncertainties.

In few recent studies (Reddy et al. (2007a), Reddy et al. (2007b), Sun and Reddy (2006), Johnson and Hu (2012) and Coakley et al. (2011) optimisation techniques, especially Monte Carlo techniques, have been used to aid the calibration of BES models. This approach follows a more rigorous mathematical basis. However the consideration of modelling uncertainties is not satisfactory. Only a few solution vectors, not sufficient to characterise the uncertainties related to parameter estimates, are provided.

Bayesian calibration improves on the two approaches mentioned, by offering the opportunity to treat the calibration problem probabilistically (Kennedy and O'Hagan (2001)). It allows for a comprehensive treatment of the uncertainties (observation errors, parameter uncertainties and model deficiencies), which are consistently considered throughout the calibration process. This influences the solution, which is returned in the form of full posterior probability density distributions, reflecting the uncertainties about parameter estimates. Furthermore, the Bayesian paradigm allows the rigorous inclusion of modeller knowledge in the calculations, through the model structure and parameters' prior probability density distributions. For these reasons, Bayesian calibration can be applied effectively for the calibration of BES models under uncertainties as shown in Heo et al. (2012), Heo et al. (2013), Kim et al. (2014) and Heo et al. (2015).

The central problems impeding its widespread adoption, are its steep learning curve and the absence of means for its routine application. CALIBRO is an R package which aims to overcome these problems by facilitating the setting up and the undertaking of the calculations underpinning the Bayesian analysis of calibration problems, thus allowing the automatic Bayesian calibration of BES models.

CALIBRO

As far as the authors are aware, CALIBRO is the first software package dedicated to the Bayesian calibration of BES models. CALIBRO has been based upon the mathematical framework established in Monari (2016), which builds on the methods depicted in Kennedy and O'Hagan (2001), Bayarri et al. (2005), and especially Higdon et al. (2008). The main usage of CALIBRO is envisaged to be the correction of models built at the design phase, by tuning their parameters according to information gained through operational monitoring.

CALIBRO uses a black-box approach and can calibrate any building energy model. The information it requires is:

- measurements of the observed performance indicators;

- measurements of the boundary conditions influencing the observed performance indicators;
- a sample of model outputs, predicting the observed performance indicator, and the corresponding model inputs.

This set of information is called a calibration dataset. Different calibration datasets can be used in the same calibration thus considering different performance indicators and boundary conditions simultaneously. It is advised to carry out the model simulations according to Latin Hyper Cube design of the model inputs. The parameters boundaries should be established according to a previous uncertainty analysis. A sample size of about 10 times the number of model inputs is suggested.

Figure 1 gives an overview of the steps and techniques combined in CALIBRO.

At first, in order to simplify and speed up the calculations, the dimensionality of the calibration dataset is reduced by applying Principal Component Analysis (Jolliffe (2002)) as in Higdon et al. (2008). Structural identifiability is then solved by means of a variance based sensitivity analysis. The method in Ratto and Pagano (2010) is applied as indicated in Lamboni et al. (2011) to automatically select the most influential parameters (calibration parameters). This technique allows the calculation of the estimates of the parameters' Sobol' first order indices (S_i) and of the relative standard errors (se). A parameter is deemed to be influential if its S_i is significantly higher than 0 (i.e. > 0 with 95% probability assuming S_i normally distributed with mean the calculated estimate and standard deviation the corresponding standard error).

In the subsequent phase Gaussian Process Regression in a quasi-Bayesian framework is used to build a black box emulator of the BES model, considering only the calibration parameters. This emulator is optimised with the Nelder-Mead method (Nelder and Mead (1965)). The training is considered successful if a Determination Coefficient at least equal to the sum of the calibration parameters' S_i is achieved.

Subsequently, the trained emulator is used to infer the calibration parameters' joint posterior probability density distribution (Monari (2016)), which is integrated with the Adaptive Metropolis-within-Gibbs algorithm (Rosenthal (2007)) to find the posterior probability density distribution for each calibration parameter. The convergence of the MCMC simulations is monitored with the R package CODA (Plummer et al. (2006)). When only one chain is simulated, its convergence is assessed depending on the ratio between the chain mean and the relative corrected standard error being smaller than a given threshold. When multiple chains are run in parallel, the Gelman and Rubin's convergence diagnostic (Gelman and Rubin (1992)) is employed. Finally the calibration parameters' estimates, consisting of the Maximum A Posteriori (MAP) values and the relative confidence intervals are returned.

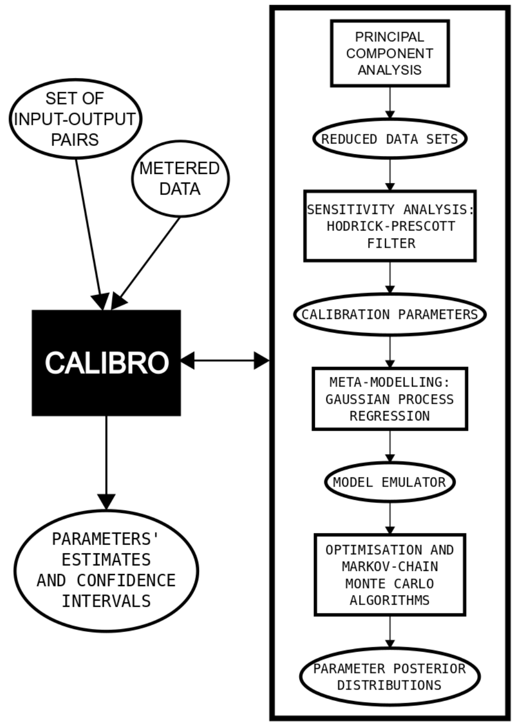

Figure 1: steps and techniques involved in CALIBRO.

Experiments

In the following two examples of CALIBRO applications are described. Two very different models are calibrated in order to demonstrate the flexibility of the method implemented in the calibration software package. The first model is of a single building component, namely a multilayer wall. The second model represents a small domestic building, and takes into account complex phenomena like wind driven infiltration. For both of these case studies, two kinds of experiments are described:

- virtual experiments: where the models are calibrated against synthetic observations for which the true parameter values are known, in order to test the correct implementation of the method;
- real experiments: where the model is calibrated against the measured observations, in order to test practical applications of CALIBRO.

The wall

The first is the calibration of a model aiming to reproduce the measured heat flux through a multilayer wall of a laboratory in an insulation factory in the south of Sweden. The experiment was conducted by the EC Joint Research Centre, Institute for Energy and Transport in ISPRA, Italy, and consisted of monitoring the heat flux through the wall, and the temperatures at the internal and external surfaces, for a period of one month. Data with a 30 minutes time step were collected. The construction element had three layers: a central core of gas concrete blocks (GC) of thickness 150 mm and insulation glass fibre boards (FB) of thickness 27 mm at both sides. At the end of the experiment, samples were taken from the test component in order to determine the properties of each material. For FB, values for density and conductivity of 116.6 kg/m^3 and 31.27 $mW/(mK)$ respectively were specified. For GC, only the density (552 kg/m^3) was provided.

The model to be calibrated consisted of two thermal zones, one representing the inside of the laboratory, and the other depicting the exterior environment, divided by the test wall. The boundary conditions, namely the internal and external temperatures, were imposed on the two faces of the test component though suitable control laws. A three layer construction, reproducing the given specifications, was used to represent the test wall layer structure. The unknown model parameters consisted of gas concrete block conductivity (GC_k), gas concrete block specific heat (GC_c) and glass fibre board specific heat (FB_c). The initial values and prior probability density distributions assumed for these parameters are indicated in Table 1. A set of 30 model input-output pairs was used as input for CALIBRO.

Table 1: The Wall: calibration parameters' prior probability density distributions and initial values.

PARA-METER	INITIAL VALUES	PRIOR DIST
GC_k ($\frac{W}{mK}$)	0.12	$Uniform(0.05, 0.15)$
GC_c ($\frac{J}{kgK}$)	800.00	$Uniform(600, 1000)$
FB_c ($\frac{J}{kgK}$)	800.00	$Uniform(600, 1000)$

In Table 2 are reported the results of the sensitivity analysis phase under the form of estimated fractions of model output variance attributable to each model parameter by itself (S_i). GC_k clearly dominated the model, GC_c had only a marginal role, while FB_c, having a negligible effect, was deemed not identifiable and it was not considered as a calibration parameter by CALIBRO.

Table 2: The wall: parameters' main effects (S_i) and relative standard errors (se).

PARAMETER	S_i	se
GC_k	0.954	0.007
GC_c	0.043	0.070
FB_c	0.002	0.018

Since this example required a relatively short computational time, at first, a series of virtual tests was undertaken to demonstrate the performance of CALIBRO depending on different levels of noise in the data. Synthetic target observations were generated for the initial values in Table 1 and then Gaussian white noise was added according to 34 increasing noise to variance ratios (NVRs). Ten calibrations were carried out, by simulating single Markov chains, for each NVR for a total of 340. In order to perform the analysis in a reasonable time a relatively high convergence tolerance was used (0.002). Figures 2 and 3 summarise the results of this set of trials. Here the blue box plots represent the samples of 0.025 and 0.975 quantiles calculated for each NVR. Similarly, the black box plots indicate the distribution of calibration parameter estimates. Blue and black dots are the average values resulting from considering only two significant figures.

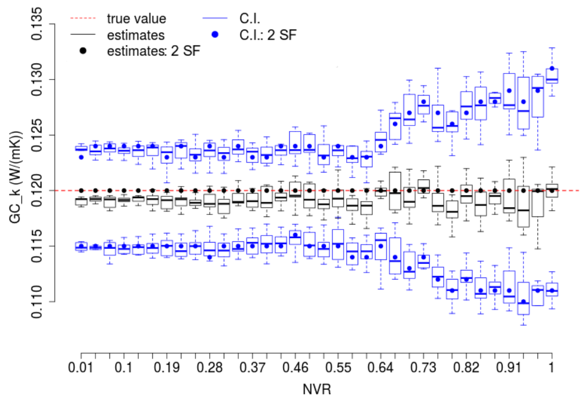

Figure 2: The wall–virtual experiments: GC_k calibration results for different NVRs. SF indicates Significant Figures.

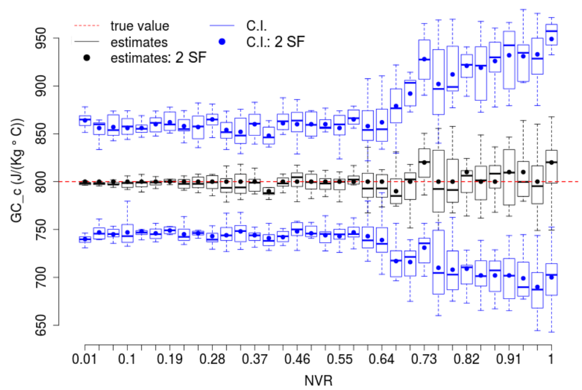

Figure 3: The wall–virtual experiments: GC_c calibration results for different NVRs. SF indicates Significant Figures.

Subsequently, the real observations were considered. The calibration parameters were inferred by running three different Markov chains in parallel. Figures 4 and 5 show the evolution of mean, 0.025 and 0.975 quantiles for such chains during the simulation, respectively for GC_k and GC_c. The calibration results are summarised in Table 3, and a comparison between calibrated model prediction and observations is depicted in Figure 6.

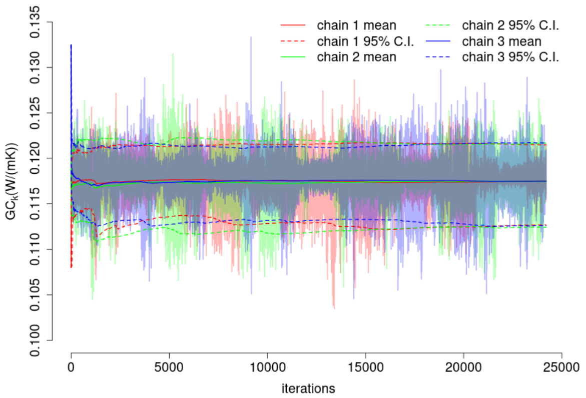

Figure 4: The wall–real experiment: evolution of three Markov chains for GC_k: traces and summary statistics.

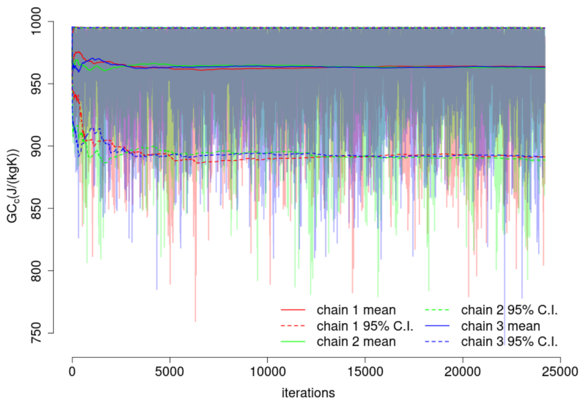

Figure 5: The wall–real experiment: evolution of three Markov chains for GC_c: traces and summary statistics.

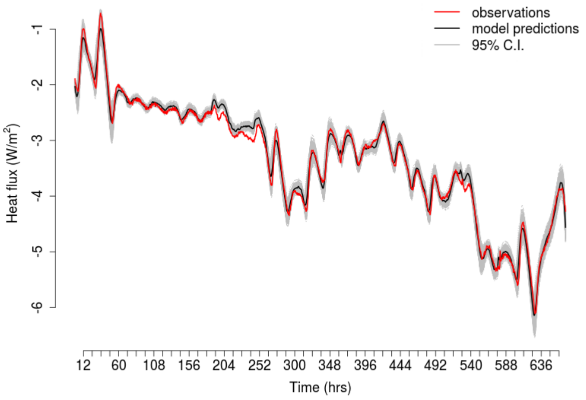

Figure 6: The wall–real experiment: comparison between calibrated model predictions (black line) and observations (red line). The 95% confidence intervals are indicated in grey.

Table 3: The Wall–real experiment: calibration parameters estimates and 95% confidence intervals.

PARAMETER	MAP	q0.025	q0.5	q0.975
GC_k $(\frac{W}{mK})$	0.118	0.113	0.118	0.122
GC_c $(\frac{J}{kgK})$	993	890	971	995

The house

In the second example the object of the analysis was to calibrate a model aiming to predict the ground and upper floor average temperatures in a small domestic building (Figure 7), which was used as test facility in developing model predictive control systems. The building is located at the BRE Innovation Park in Motherwell (Scotland), and it is a prototype for a modular mass market low energy house design. Its main heat source is 5 kW exhaust air heat pump feeding underfloor heating. The unit also includes a small hot water store. A 1 kW array of photovoltaic panels is located on the roof. Being a test facility, the house was unoccupied, and its properties were extensively investigated through dedicated tests performed by the research team. A blower door test was undertaken to characterise air tightness, and its thermal behaviour was observed subject to dedicated thermal pulse testing, and to a normal heating regime. The monitoring period lasted approximately one month, collecting data with a 10 minute time step. The data comprised external temperature, wind velocity, wind direction, relative humidity, diffuse solar radiation, direct solar radiation, heat injected in the ground floor and heat injected in the first floor. This dataset was used by the research team to manually calibrate a BES model. The results that are presented in the following have been achieved by applying CALIBRO independently afterwards, with the purpose of evaluating the eventual benefit of its employment in future studies.

Figure 7: The house.

One virtual experiment and two real experiments (real 1 and real 2) were undertaken. In the virtual and real 1 experiments, 55 parameters were considered, consisting of conductivities, densities, window thermal resistances, crack lengths, and constant volume flow rates. Construction properties (conductivities, densities, window thermal resistances) were fairly well known, while much more uncertain were deemed to be the airflow network parameters (crack lengths, and constant volume flow rates). Therefore the former were varied within ±15% and the latter were changed within ±80% of their design values according to uniform probability density distributions. Subsequently (real 2 experiment), windows optical transmissions were included in the calibration parameters, since correlation was noticed between the solar radiation and the residuals between model predictions and observations. These were varied within ±90% of their design values. A sample of 550 model input-output pairs was used as input for both the analyses.

Virtual and real 1 experiments

The outcomes from the sensitivity analysis showed that the model was highly conditioned by the value of the parameter HP_{ext} which represents the volume flow rate of the heat pump exhaust air extraction. However, by breaking down the sensitivity results according to the principal components, CALIBRO identified other minor effects from some parameters representing lengths of cracks around the frames of south and west windows located at the ground floor of the building (*crck2l* and *crck3l*). This group of model inputs was considered as calibration parameters. Their sensitivity indexes are displayed in Table 4, and the relative initial values and prior probability density distributions are listed in Table 5.

Table 4: The house - virtual and real 1 experiments: estimates of parameters' main effects (S_i) and relative standard errors (se).

PARAMETER	S_i	se
HP_{ext}	0.965	0.001
$crck2_l$	0.006	0.003
$crck3_l$	0.006	0.005

Table 5: The house - virtual and real 1 experiments: calibration parameters' prior probability density distributions and initial values.

PARAMETER	INITIAL VALUES	PRIOR DIST
HP_{ext} $(\frac{m^3}{10^3 s})$	4.3	$Uniform(0.86, 7.74)$
$crck2_l$ (m)	4.6	$Uniform(0.92, 8.28)$
$crck3_l$ (m)	6.2	$Uniform(1.24, 11.16)$

Firstly the model was calibrated against synthetic observations, which were obtained by running the model for the initial values in Table 5. In this case no noise was added, in order to assess the capabilities of the software of identifying very weak parameters like $crck2_l$ and $crck3_l$. The evolutions of the three Markov chains used in this calibration are shown in Figures 8, 9 and 10.

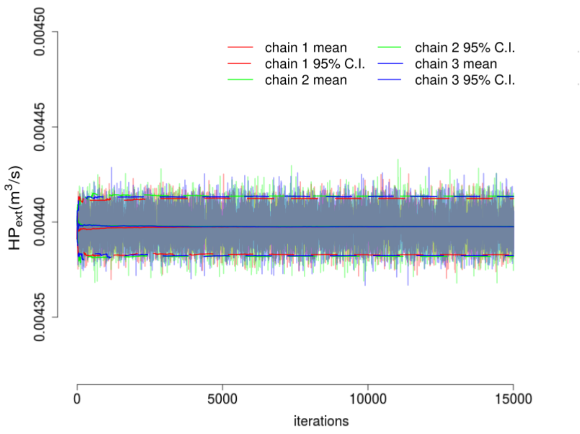

Figure 8: The house - virtual experiment: evolution of three Markov chains for HP_{ext}: traces and summary statistics.

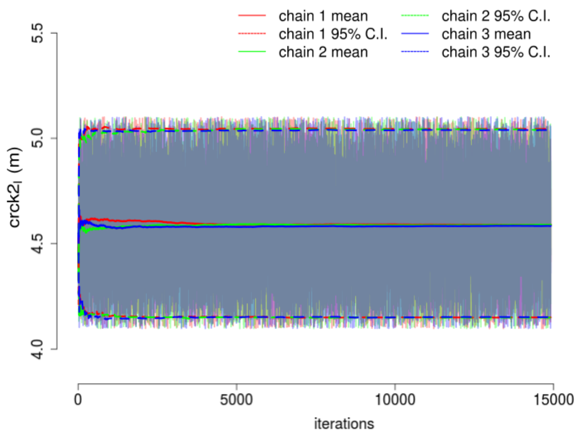

Figure 9: The house - virtual experiment: evolution of three Markov chains for $crck2_l$: traces ans summary statistics.

After having substituted the synthetic observations with the measured ground floor (GF) and upper floor (UF) temperatures, a new calibration was undertaken. The results are summarised in Table 6 containing calibration parameters' estimates and confidence intervals.

Table 6: The house - real 1 experiment: calibration parameters estimates and 95% C.I..

PARAMETER	MAP	q2.5	q50	q97.5
HP_{ext} $(\frac{m^3}{10^3 s})$	4.53	4.5	4.53	4.55
$crck2_l$ (m)	4.11	4.10	4.16	4.37
$crck3_l$ (m)	5.63	5.62	5.78	6.29

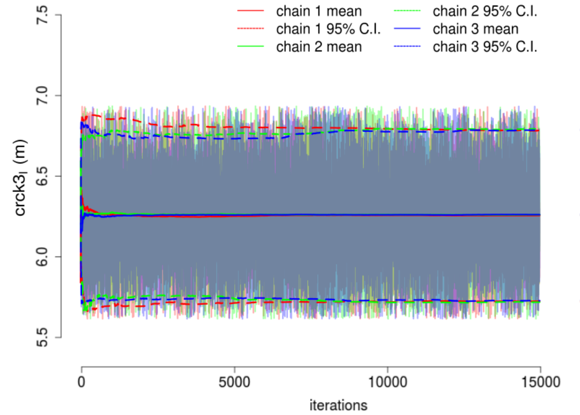

Figure 10: The house - virtual experiment: evolution of three Markov chains for $crck3_l$:traces and summary statistics.

Real 2 experiment

In this experiment the the optical transmission of the ground floor windows (Tr_{GF}) and the optical transmission of the upper floor windows (Tr_{UF}) were added to the free parameters of the model, after having noticed correlation between the solar radiation and the residuals between observed temperatures and predictions of the model calibrated in the real 1 experiment. The outcomes from the sensitivity analysis step (Table 7) highlighted that Tr_{GF} and Tr_{UF} had effects comparable to the main parameter's one (HP_{ext}).

Table 7: The house - real 2 experiment: estimates of calibration parameters' main effects (S_i) and relative standard errors (se).

PARAMETER	S_i	se
Tr_{GF}	0.14	0.041
Tr_{UF}	0.39	0.045
HP_{ext}	0.57	0.037
$crck2_l$	0.02	0.016
$crck3_l$	0.005	0.009

Table 8: The house - real 2 experiment: optical transmission parameters' prior probability density distributions and initial values.

PARA-METER	**INITIAL VALUES**	**PRIOR DIST**
Tr_{GF} (-)	0.346	$Uniform(0.01, 0.66)$
Tr_{UF} (-)	0.346	$Uniform(0.01, 0.66)$

The calibration was undertaken by running three Markov chains in parallel. Their evolutions for the three most significant calibration parameters (Tr_{GF}, Tr_{UF} and HP_{ext}) are summarised in Figures 11, 12 and 13. The inferred estimates and confidence intervals are contained in Table 9. The model calibrated in this test had the best predictive performances (Table 10). The achieved matches between its predictions and measured temperatures are graphically described in Figure 14 and 15.

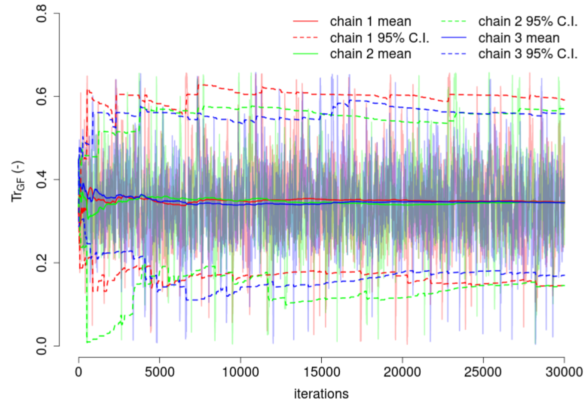

Figure 11: The house - real 2 experiment: evolution of three Markov chains for TR_{GF}: traces and summary statistics.

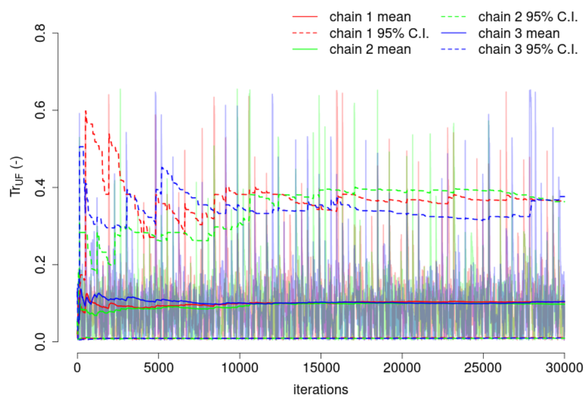

Figure 12: The house - real 2 experiment: evolution of three Markov chains for TR_{UF}: traces and summary statistics.

Table 9: The house - real 2 experiment: calibration parameters estimates and 95% C.I..

PARA-METER	**MAP**	**q2.5**	**q50**	**q97.5**
Tr_{GF} (-)	0.345	0.154	0.34	0.57
Tr_{UF} (-)	0.101	0.001	0.080	0.367
HP_{ext} ($\frac{10^3 m^3}{s}$)	3.10	2.17	2.92	5.52
$crck2_l$ (m)	4.43	1.22	4.29	8.08
$crck3_l$ (m)	3.02	1.32	2.66	6.83

Discussion and result analysis

The undertaken virtual experiments demonstrated the capability of CALIBRO to identify effectively calibration parameter values, even when significant levels of noise contaminate the measurements.

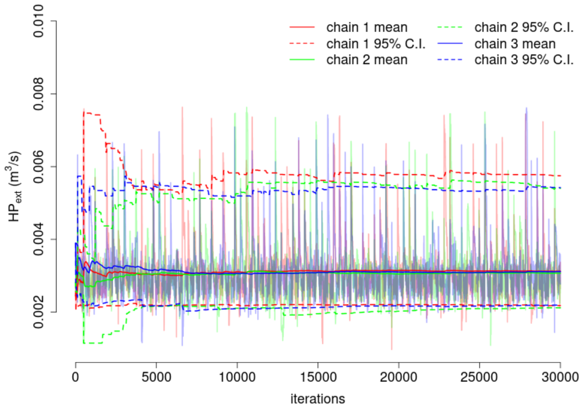

Figure 13: The house - real 2 experiment: evolution of three Markov chains for HP_{ext}: traces and summary statistics.

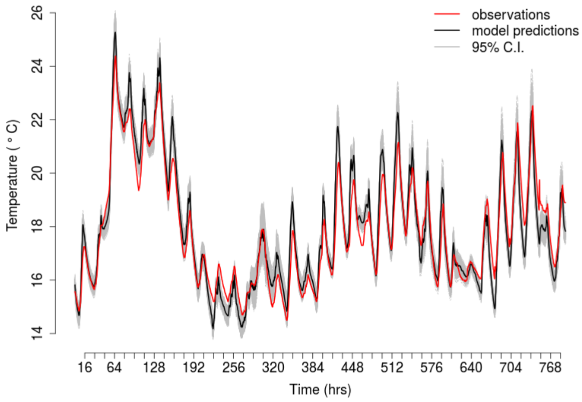

Figure 14: The house - real 2 experiment: comparison between ground floor temperature predictions and observations.

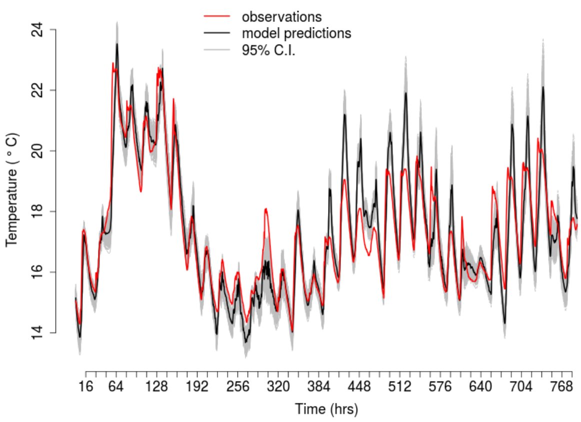

Figure 15: The house - real 2 experiment: comparison between upper floor temperature predictions and observations.

In the first series of virtual trials (Figures 2 and 3), CALIBRO was always able to correctly estimate the considered model inputs even for high noise to variance ratios. Indeed, despite little inaccuracy, that may be due to the premature stop of the MCMC simulations because of the high convergence threshold adopted, the true parameters values are always well centred within the calculated confidence intervals and close to the inferred estimates. If the effects of such small inaccuracies are diminished by rounding the results to suitable precision, that is two significant figures all the calibrated values are equal to the actual parameter values.
Similar conclusions can be drawn from the outcomes of the calibration against synthetic observations performed on the house model. All the MCMC simulation converged to the same stationary distribution (Figures 8, 9 and 10) and all the calibration parameters were estimated correctly. HP_{ext} has been negligibly overestimated ($0.0044 \approx 0.0043\frac{m^3}{s}$) and even $crck2_l$ and $crck3_l$, although having almost indiscernible effects, were well identified. In particular, their empirical probability density distributions, despite having large variances, are clearly peaked around the respective true values (Figures 16 and 17). These last results are consistent with those depicted in Figure 2 and 3. As indicated by the box plots, the calculated values become increasingly uncertain with the noise. Similarly the average confidence intervals broaden as the NVR grows, especially for GC_c which is the weakest of the two calibration parameters. Thus, measurement uncertainties are correctly influencing the calculations, and the results correctly follow the sensitivity of the model, returning higher uncertainties for the less important parameters.

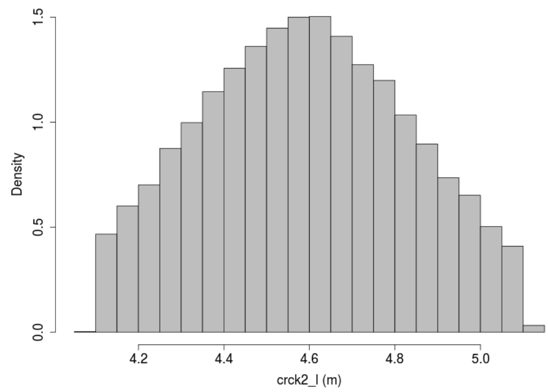

Figure 16: The house–virtual experiment :$crck2_l$ empirical probability density distribution.

The performed real experiments proved the ability of CALIBRO to tune a model's inputs in order to improve the efficacy of such a model in predicting observed performance indicators. It is difficult to judge the actual correspondence between the estimated parameters and the real properties of the wall that they represent, since the specifications of the wall have not been disclosed, while airflow network parameters, for the house experiment, are naturally unknown due to

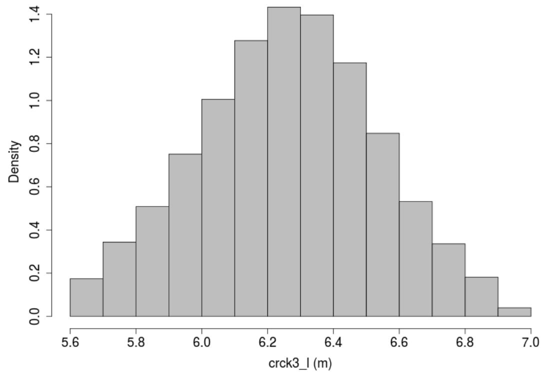

Figure 17: The house–virtual experiment: $crck3_l$ empirical density distribution.

the difficulty of measuring them. However, confidence in the goodness of the calibration results is given by Figures 4, 5, 11, 12 and 13, which show that during the calibrations of the model against real measurements all the MCMC simulations converged to the same probabilistic solutions. Especially the strongest parameters (GC_k, HP_{ext}, TR_{GF}, and TR_{UF}) showed reasonable estimates well centred in the respective confidence intervals, and characterised by bell shaped empirical probability density distributions, indicating that these variables were effectively estimated. Conversely, the weakest parameters (GC_c, $crck2_l$ and $crck3_l$) showed estimates close to their confidence boundaries and large variances. Probably, their estimations were influenced by observation errors and model deficiencies. In this case CALIBRO tends to excessively modify the least important model parameters resulting in a little over-fit of the measurements, as discussed more in detail in Monari (2016). Nevertheless, due to their minor effects, errors in the estimates of these parameters do not affect sensibly the model predictions.

As shown by the graphical comparison in Figure 6, it was possible to achieve a very good match between the heat flux through the wall predicted by the calibrated model and that measured. In particular, it is not possible to notice any significant discrepancy between predictions and measurements, which are consistently with the 95% confidence intervals. This is further highlighted by the low Root Mean Square Error (RMSE), equal to 0.1 W/m^2.

The final calibrated model for the house was able to fit well the mean temperature of the ground floor (Figure 14), while it showed some inaccuracies in predicting the average temperature of the upper floor (Figure 15). In particular the predicted upper floor temperature seems too sensitive to heat gains from solar radiation and thermal pulses. It is interesting to notice the two different inferred values for the window optical transmission. Tr_{UF} had an MAP value less than one third of Tr_{GF}. Different window specifications for the two floors are deemed to be unlikely, and such difference may be due to accumulated dirt or some phenomena not recorded during the monitoring. A more likely explanation might be a lack of thermal mass in the upper floor. Material densities were not highlighted by the sensitivity analyses performed as significant parameters. The cause may be their individual consideration instead of taking into account more global parameters governing the building thermal capacity. This may suggest that the inclusion of details in the model should be done gradually, depending on the outcome of the calibration. In particular it is believed that, at first, it would be convenient to consider few important macro-parameters acting on different locations of the model, and then break down such macro-parameters according on the results of analysis. The correction of this model deficiency requires more investigation which will be object of future studies. For the purpose of this research it is more useful to compare the manual iterative calibration process and its outcomes with the application of CALIBRO and its results.

The time spent to manually calibrate the model was two days. The process involved the empirical identification of the calibration parameters according to the expertise and knowledge of the modeller, and their iterative variation until a satisfying match with the observed temperatures was achieved. In the same amount of time it was possible to apply CALIBRO twice. The most time consuming phases were generating the required samples of simulations, that were carried out overnight. The actual calibrations required a few hours, during which the identification of calibration parameters was performed automatically through a detailed variance based sensitivity analysis, and the identification of a solution was performed according to robust probabilistic techniques allowing the assessment of the reliability of the inferred model parameter estimates. It should also be noted that a relatively short lasting manual calibration process, as in this case, is more an exception, due to the large amount of information available, than a rule. Furthermore, if the analyst was willing to perform the selection of the calibration parameters based upon sensitivity analysis, the required time would have increased greatly. Finally, both the models calibrated with CALIBRO achieved better matches with the measurements (Table 10), having overall 34% and 55% lower RMSEs.

Table 10: The house–real experiment: RMSE.

CALIBRATION TYPE	GF	UF	GLOBAL
manual iterative	0.69	1.72	1.31
CALIBRO - real 1	0.60	1.07	0.86
CALIBRO - real 2	0.62	0.81	0.72

Conclusions

This paper has introduced CALIBRO, an R package for the Bayesian calibration of building energy models. Such a software package allows the easy undertaking of the calibration of BES models according to a Bayesian framework, providing the analysts with the possibility of rigorously introducing their knowledge in the calculations, and to prove or disprove their hypothesis against the information provided by field observations. The main features of CALIBRO are the following.

- Model independence: adopting a black-box approach, the calibration process can be applied to any building energy model, requiring only the information defined in the calibration dataset as inputs.
- Multi-objective calibration: multiple performance indicators can be contained in one calibration dataset, and in turn multiple calibration datasets can be considered simultaneously during the calibration.
- Sensitivity analysis: the selection of the calibration parameters is performed automatically according to a detailed variance based sensitivity analysis.
- Robust probabilistic and statistical techniques: the inference of the calibration parameter values most likely to reduce the gap between model predictions and field measurements is undertaken though optimisation and MCMC algorithms, which provide assessment of the reliability of the calculated estimates.

The capabilities of CALIBRO have been demonstrated in the described virtual and real experiments. The results relative to the former class of tests, showed good performance in estimating calibration parameters even if the effects of such parameters on the model output are small, and also when significant noise contaminates the measurements. The inferred calibration parameter estimates were reliable, reflecting in their uncertainties the magnitude of the observation errors and the sensitivity of the model. The outcomes from the real examples, demonstrated that CALIBRO is effective at improving the predictive capabilities of BES models. With respect to the house example, the model calibrated with CALIBRO produced better fits of the measured internal temperatures than that manually calibrated. However, this does not mean that the application of the software package will yield systematically a better fit with respect to other approaches. The assessment of its performance in comparison to other kinds of calibration procedure would require a statistical investigation which could be object of future research. Nevertheless, the calibration of the house model with CALIBRO required less time than the manual calibration of the same model. Furthermore, the outputs from CALIBRO could be used to asses the reliability of calculated calibration parameter values, building up confidence in the analyst that a good solution was found. The information returned by CALIBRO is also effective in spotting model deficiencies and consequent model improvements, thus making the software package a useful tool for the analysis and diagnosis of BES models against monitored data. Therefore the authors are confident in saying that, compared to the usual practice of manually calibrating building energy models, the employment of CALIBRO offers a theoretical edge, achieving more reliable results in a shorter period of time.

However, the software has limitations that require improvements. Calibration parameters are identified only according to the structural identifiability, while it would be important to take into account the experimental identifiability as well. It is currently impossible to consider vectorial model parameters, such as pressure coefficients, time varying convection coefficients, unknown heat gains or occupancy schedules, which are an important class of inputs for BES models. Also a development priority is to provide a means to compare different calibrated models through reliable model selection criteria.

CALIBRO has completed a first development phase, and it is now undergoing a testing stage in the context of Hit2Gap European project (Costola et al. (2017)). The authors are confident that its robustness and reliability will be further strengthened by this testing process. A first version of the software can be found at `http://www.esru.strath.ac.uk/software.htm`. The final purpose is to provide, with CALIBRO, a sophisticated tool supporting the development of novel practices and systems, advancing Building Energy Simulation and allowing its routine application during the operational phase of a building life cycle.

References

Bayarri, M. J., J. O. Berger, D. Higdon, M. C. Kennedy, A. Kottas, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C. H. Lin, and J. Tu (2005). A framework for validation of computer models. Technical report, National Institute of Statistical Sciences, 19 T. W. Alexander Drive, PO Box 14006, Research Triangle Park, NC 27709-4006, www.niss.org.

Clarke, J. A., P. A. Strachan, and C. Pernot (1993). An approach to the calibration of building energy simulation models. *ASHRAE Trans*, 917–927.

Coakley, D., P. Raftery, P. Molloy, and G. White (2011, 14-16 November). Calibration of a detailed bes model to measured data using an evidence-based analytical optimisation approach. In *Proceedings of Building Simulation 2011: 12th Conference of International Building Performance Simulation Association, Sydney.*

Costola, D., J. Clarke, N. Kelly, and F. Monari (2017). A big data approach to the application of building performance simulation to improve the operational performance of large estates. Accepted for publication in *Proceedings of Building Simulation 2017: 15th Conference of the International Building Performance Simulation Association, San Francisco.*

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Heo, Y., G. Augenbroe, and R. Choudhary (2013). Quantitative risk management for energy retrofit projects. *Journal of Building Performance Simulation 6(4)*, 257–268.

Heo, Y., R. Choudhary, and G. Augenbroe (2012). Calibration of building energy models for retrofit analysis under uncertainty. *Energy and Buildings 47*, 550–560.

Heo, Y., D. J. Graziano, L. Guzowski, and R. T. Muehleisen (2015). Evaluation of calibration efficacy under different levels of uncertainty. *Journal of Building Performance Simulation 8*(3), 135–144.

Higdon, D., J. Gattiker, B. Williams, and M. Rightley (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association 103(482)*, 570–583.

Johnson, N. and H. Hu (2012). Developing an effective and reliable approach for building energy audit in support of retrofit decision making. In *International High Performance Building Conference, Purdue.*

Jolliffe, I. (2002). *Principal Component Analysis (Springer Series in Statistics).* Springer.

Kennedy, M. and A. O'Hagan (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B 63*, 425–464.

Kim, Y., K. Kim, C. Park, and I. Kim (2014, November). Deterministic vs. stochastic calibration of energy simulation model for an existing building. In *ASim 2014, IBPSA Asia Conference, Nasoga Japan, 2014.*

Lamboni, M., H. Monod, and D. Makowski (2011, Apr). Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety 96*(4), 450459.

Monari, F. (2016). *Sensitivity Analysis and Bayesian Calibration of Building Energy Models.* Ph. D. thesis, University of Strathclyde, Glasgow, UK.

Nelder, J. A. and R. Mead (1965, January). A simplex method for function minimization. *The Computer Journal 7*(4), 308–313.

Pedrini, A., F. Westphal, and R. Lambertsb (2002). A methodology for building energy modelling and calibration in warm climates. *Building and Environment 37*, 903–912.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News 6*(1), 7–11.

Raftery, P., M. Keane, and A. Costa (2011a). Calibrating whole building energy models: Detailed case study using hourly measured data. *Energy and Buildings 43*(12), 3666 – 3679.

Raftery, P., M. Keane, and J. O'Donnell (2011b). Calibrating whole building energy models: An evidence-based methodology. *Energy and Buildings 43*(9), 2356 – 2364.

Rasmussen, C. and C. Williams (2006). *Gaussian Process for Machine Learning.* the MIT Press.

Ratto, M. and A. Pagano (2010, Dec). Using recursive algorithms for the efficient identification of smoothing spline ANOVA models. *AStA Advances in Statistical Analysis 94*(4), 367388.

Reddy, T., I. Maor, and C. Panjapornpon (2007a). Calibrating detailed building energy simulation programs with measured data–part ii: application to three case study office buildings (rp-1051). *HVAC & R Research 13*(2), 221–241.

Reddy, T., I. Maor, I., and C. Panjapornpon (2007b). Calibrating detailed building energy simulation programs with measured data–part i: general methodology (rp-1051). *HVAC & R Research 13*(2), 243–265.

Rosenthal, J. S. (2007, Aug). AMcmc: An r interface for adaptive MCMC. *Computational Statistics & Data Analysis 51*(12), 5467–5470.

Sun, J. and T. Reddy (2006). Calibration of building energy simulation programs using the analytic optimization approach (rp-1051). *HVAC&R RESEARCH 12(1)*, 177–196.

Westphal, F. S. and R. Lamberts (2005). Building simulation calibration using sensitivity analysis. In *Ninth International IBPSA Conference Montral, Canada August 15-18.*

Yoon, J., E. Lee, and D. Claridge (2003). Calibration procedure of energy performance simulation model for a commercial building. *Journal of Solar Energy Engineering 125(August)*, 251–257.