Bayesian Calibration of Building Energy Models for Uncertainty Analysis Through Test Cells Monitoring

Carmen María Calama-González[1*], Phil Symonds[2], Giorgos Petrou[3], Rafael Suárez[1,] Ángel Luis León-Rodríguez[1].

[1] Instituto Universitario de Arquitectura y Ciencias de la Construcción, Escuela Técnica Superior de Arquitectura, Universidad de Sevilla, Av. Reina Mercedes 2, Seville 41012, Spain.
[2] UCL Institute for Environmental Design and Engineering, Central House, 14 Upper Woburn Plc, London WC1H 0NN, U.K.
[3] UCL Energy Institute, Central House, 14 Upper Woburn Plc, London WC1H 0NN, U.K.
* Corresponding author: Tel.: +34 954559517
E-mail address: ccalama@us.es (C.M. Calama-González)

Highlights

- Calibrating energy simulation models is crucial when assessing existing buildings.
- Sensitivity analysis is key to reduce computational time in the calibration process.
- Uncertainty techniques may be applied to assess energy models' accuracy.
- Test Cells allow the performance of building simulation tools to be estimated.

Abstract

Improving the energy efficiency of existing buildings is a priority for meeting energy consumption and CO_2 emission targets in buildings. Building simulation tools play a crucial role in evaluating the performance of energy retrofit options. In this paper, a Bayesian calibration approach is applied to reduce the discrepancies between measured and simulated temperature data. Through its application to a test cell case study, the incorporation of sensitivity analysis and Bayesian calibration techniques are proven to improve the level of agreement between on-site measurements and simulated outputs, whilst accounting for both experimental and simulation uncertainties. The accuracy of a building simulation model developed using EnergyPlus was evaluated before and after calibration. Uncalibrated models were within the uncertainty ranges specified by the ASHARE Guidelines, with hourly simulation data over-predicting measurements by 3.2 ºC on average. After Bayesian calibration, the average maximum temperature difference was reduced to around 0.68 ºC, an improvement of almost 80%.

Keywords: Bayesian calibration; sensitivity analysis; uncertainty analysis; building energy modelling; Mediterranean climate; housing stock.

1. Introduction

In the European Union, the number of existing dwellings is about 196 million [1]. The average annual rate of construction of new buildings is around 1.1%, with an estimated annual ratio of building replacement of only 0.07% [2]. The existing European building stock accounts for over 40% of total energy consumption, of which residential represents 63% [3].The building sector is responsible for approximately 30% of carbon dioxide emissions [4], which means that retrofit and refurbishment must be one of the key priorities in meeting the objectives proposed for 2030 [5].

Building Energy Modelling (BEM) allows for the evaluation of alternative options for achieving building energy efficiency. Although originally intended for the building design and operation stages [6], nowadays BEM is increasingly being used in other stages of a project [7], especially in the refurbishment phase [8]. Nevertheless, BEM has its limitations, and may only capture limited parts of a large number of dynamic, stochastic and probabilistic elements (e.g. building geometry, thermal zones, material properties, Heating, Ventilation and Air Conditioning (HVAC) systems, occupant behaviour, appliance, use scheduling, etc.) [9], inevitably leading to a simplified prediction of real building performance. In addition, software limitations, construction, users, inputs, weather data inaccuracy and errors in measurements may lead to a significant performance gap [10] between real and simulated data 11]. The dynamic complexity of calculation methods and the choice of simulation tool can also result in the inaccuracy [12] and uncertainty [13] associated with BEM.

For the above reasons, to ensure the accuracy and reliability of the simulation results and reduce the performance gap when assessing existing buildings, model calibration is a key step in the BEM to minimize the discrepancies between predicted and monitored data [14]. During this process, information about the building is collected and used to tune the BEM, in order to achieve a greater level of accuracy [15]. Although there is no single generally accepted methodology for BEM calibration [16], Clarke, Strachan and Pernot [17] devised a proposal that was later revised by Reddy, Maor and Panjapornpon [18], by which model calibration may be classified into four different categories: (1) heuristic or pragmatic intervention, (2) graphical-based calibration methods, (3) calibration based on special tests and (4) automated techniques. During the calibration process, different methods may be combined, as established by Clarke,

Strachan and Pernot [17]: the heuristic technique involves selecting parameters and manually calibrating them based on monitored data, through trial-and-error, usually changing one variable at a time to be compared to the original model; graphical-based calibration is normally used in combination with manual methods and consists of time-series and scatter-plot representations; analytical calibration includes special tests which do not involve statistical procedures and are normally quite invasive, for instance blower door or thermal transmittance tests and audits; lastly, automated techniques apply mathematical and statistical tests, involving optimization functions, parameter estimation or uncertainty incorporation.

A large number of previous studies have used manual calibration [19]: Royapoor and Roskilly [9] apply a heuristic iterative approach to calibrate an office building running 19 models, each with incremental manual input adjustments, related in particular to the electricity and HVAC systems. Raftery, Keane and Costa [13] also calibrate an office building by manually varying internal loads and the HVAC systems' characteristics, representing the building to a high level of detail and taking considerable time and resources. Parker, Cropper and Shao [20] conducted up to 118 individual parameter modifications for an airport terminal building, iteratively updating construction properties, systems details, equipment energy and airflow, among other variables, needing extensive information.

One of the major problems of manual calibration is the high dependency on parameter and value selection [21], which may significantly reduce the calibration quality and effectiveness. This is because analysts rely on their subjective judgment to select the input variables they believe will most likely influence the outputs iteratively running simulations to determine the scenario where differences between simulated and measured data are reasonably small. A drawback of this approach is reported by Heo, Choudhary and Augenbroe [22]: identifying a single combination of parameter values that leads to a good fit does not guarantee that those values represent reality with confidence. Besides, calibrating every parameter that influences the simulation may also be a poor use of time and resources [23]. Given that a typical building energy model could have an average of 3,000 parameters, manual calibration could require months through a trial-and-error process [24], depending on the model complexity. Thus, manual calibration involves limited simulation runs, may become time-consuming, its credibility may be questioned given its subjectivity and it cannot be easily scaled up to other models [24].

Automated calibration is now being increasingly used by the scientific community [25], although it only represents around 26% of existing studies [26]. Several approaches may be used, such as sensitivity analysis, meta-models, optimization-based problems or Bayesian techniques [27]. It is generally accepted that the fewer parameters to be optimized, the more efficient the optimization [19]. Thus, sensitivity analysis is normally incorporated into any calibration method [28] to reduce the number of parameters to be calibrated [29].

In the meta-model approach, a surrogate model is created to reduce the complexity of the original model through a mathematical function determined by a limited number of input-output combinations [27]. One advantage of this is the reduced computation time, allowing a large number of scenarios to be analysed, this method has been extensively used in BEM: O'Neil and Eisenhower [30] apply a meta-model to calibrate an office building by sampling nominal values for all parameters within the model, identifying which parameter combinations provide the best fit to monitored data. Manfren, Aste and Moshksar [31] also applied meta-model calibration to assess an office building, standardizing and categorizing the input-output parameters to construct a dataset and run many simulations to account for input variation. Thus, meta-models have the advantage of considering all the parameters in the model,allowing numeric algorithms to decide which ones are the most critical in terms of calibration. Gaussian processes (GP) is the most widely used meta-model due to its robustness in interpolation. Although meta-modelling considerably reduces computational burden during calibration, GP can be computationally intensive compared to other meta-modelling approaches [32].

Optimization methods define an objective function through an optimization algorithm to identify the best parameter combination to minimise the difference between monitored and simulated data. Optimization algorithms often require a large amount of computer resources. For example, Hong, Kim, Jeong, Lee and Ji [33] calibrated a school building using a genetic algorithm to find global optimal solutions; and Sanyal, New and Edwards [34] automatically calibrated a BEM using trained machine agents from a large set of parametric simulations through machine learning, exploiting supercomputing resources. A general disadvantage of optimization methods is the dependency of the calibration procedure on the optimization settings [35].Difficulties may arise in selecting optimization hyperparameters settings, which can lead to local minima problems [32].

Quantifying uncertainties in BEM is recognized as fundamental for evaluating the cumulative impact on simulated outputs' reliability [36]. Most building calibration methods tackle the incorporation of uncertainties regarding measured data, model predictions [37] and input parameters, [38], through classical statistical methods, such as uncertainty [39] or sensitivity analysis that determines the impact of uncertain variables on simulation outcomes [40]. Among the various calibration methods used during recent decades, optimization algorithms and Bayesian techniques have become the most favourable [32]. The main advantage of Bayesian methods is allowing the incorporation of uncertainties into the calibration process using statistical inference through probabilistic predictions. This can considerably contribute to the improvement of parameter estimation and model resolution. Given its expandability and accuracy, Bayesian calibration has received increasing attention, as reported by Lim and Zhai [41], which provides an extensive list of studies where this technique is used in BEM at both individual and stock levels. As concluded by Riddle and Muehleisen [42]: this approach balances the ability to reach a good fit between monitored and simulated data, with the knowledge about the uncertainty involved with parameter estimation.. Furthermore, Bayesian calibration can be automated with minimal user input, allowing the algorithm to identify which parameter values are most likely, whilst accounting for various probability distributions. However, audits and measurements of a real building's performance are vital for achieving a good calibration, thus affecting the effectiveness of this method [43].

2. Objectives and research scope

Given the benefits previously established, this research addresses model calibration through the use of Bayesian techniques, explained in detail in subsection 3.5. The computational burden of Bayesian calibration may noticeably increase when adding more output or calibration parameters [42], thus, most BEM studies are limited to using monthly energy data. However, indoor comfort assessment is also crucial when retrofitting existing buildings and is normally reported hourly.

As such, the scope of this study is to determine if Bayesian calibration techniques can adequately provide an accurate calibration of BEM based on an hourly basis analysis.

Simulation predictions have been compared with measured indoor air temperatures under various data taking protocols, quantifying the level of agreement according to current guidelines. In order to determine whether the proposed calibration methodology is suitable and applicable, a case study with controlled boundary conditions and high resolution measured data has been conducted. Test Cells have been commonly used to evaluate experimental and dynamic performance of building components under controlled conditions [44]. Data monitored in a pair of Test Cells, located in Seville, southern Spain (Mediterranean climate) have been used. The main aim of this paper is to provide a global and representative first approach to an hourly-based statistical calibration-simulation Bayesian technique which combines sensitivity analysis, evaluating its appropriateness for obtaining reasonable results within computing-time and resource constraints. An extensive comparison of the benefits of this technique in comparison with other research in the field is described in section 5.

3. Methods

The methodology used in this research combines empirical monitoring, thermodynamic building simulations and statistical techniques. It follows a Bayesian method to calibrate the temperature predictions of a building simulation model using high quality monitoring data from a pair of Test Cells. On site indoor air temperature measurements taken during different experimental protocols have been compared to the outputs reported by the simulation tool. The following analysis steps (Figure 1) have been followed, which are explained in detail in subsequent subsections:

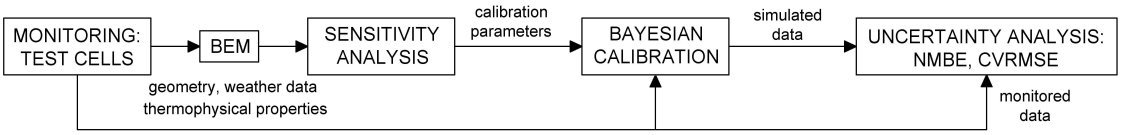


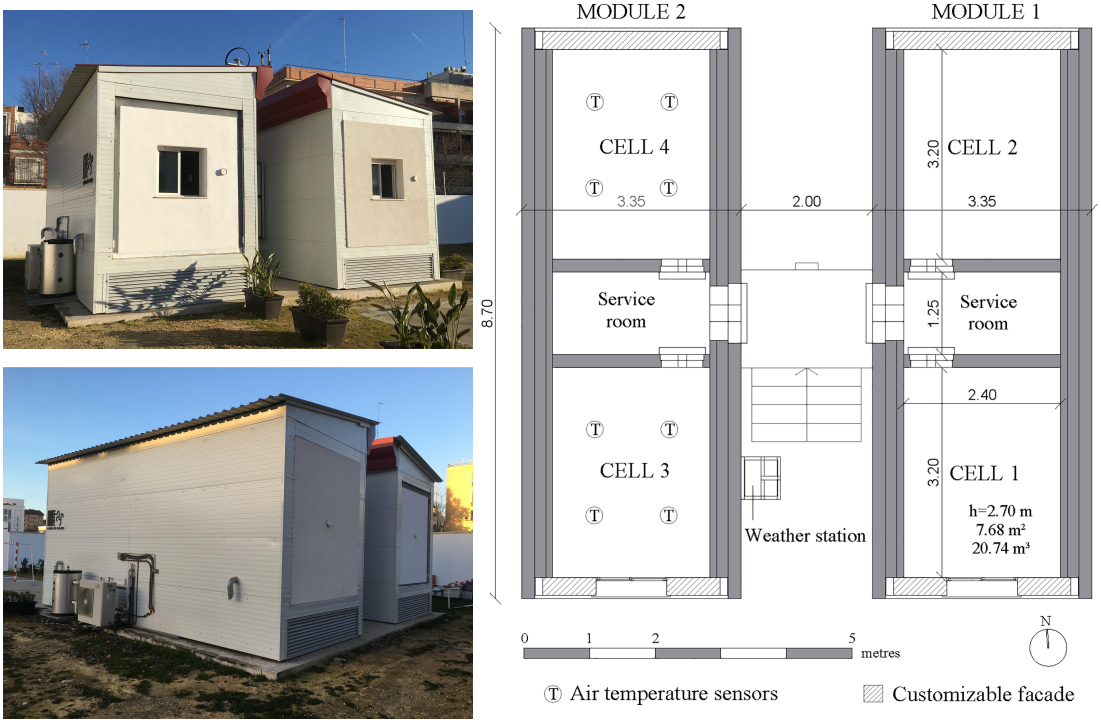
Figure 1. Methodology followed.

In the first phase of the methodology; “Experimental setup: Test Cells and Protocols description” (described in subsections 3.1 and 3.2), an experimental case study is selected and monitored during different protocols to obtain high quality measurements of indoor and outdoor ambient variables. In the following phase; “Construction of the energy building model”, a building

simulation model of the case study is constructed in an energy simulation tool, making the necessary geometrical, physical and constructive assumptions and simplifications, described in subsection 3.3. Prior to model calibration, the most influential input parameters on output results are determined, with these parameters subsequently used in calibration. This is achieved through phase 3 of the analysis, “Sensitivity analysis of input setting variation”, detailed in subsection 3.4. Once the parameters with the highest impact on outputs have been identified, they are analysed in detail and calibrated through Bayesian techniques during phase four, “Calibration of the most influential parameters” (subsection 3.5). Finally, the viability of the calibrated model is tested during phase 5, “Uncertainty analysis: accuracy measurement of the energy building model”, as explained in subsection 3.6. Here, calibrated model predictions are compared with monitored data recorded in phase 1.

3.1 Experimental setup: Test Cell Description

The selected case study consists of two Test Cells located in a Mediterranean area of southern Spain (37º 23’ N, 5º 58’ W), which are modelled on a typical Andalusian bedroom (Figure 2). Each cell is autonomous and records high quality data on the performance of different façade components, under real outdoor conditions.



(a) (b)

Figure 2. (a) Experimental Test Cells; (**b**) Floor plan of the Test Cells.

For this research, only the south-facing Cell 3 (with window) and north-facing Cell 4 (without window) were considered. Further information regarding the Test Cells' geometry, thermophysical properties (e.g. U-values, blower door test…) and technical aspects can be found in [45]. For monitoring purposes, four sensors were installed inside the Cells, to measure dry-bulb air temperatures at 5-min intervals. Outdoor ambient variables were recorded by a local weather station, located on the roof of the Cells, which were used to develop a weather file for the building simulations. Outdoor variables monitored were: dry-bulb air temperature, relative humidity, wind speed and direction and solar radiation (global horizontal, diffuse horizontal and direct normal irradiance). Detailed technical characteristics of the probes installed can be found in [46].

3.2 Experimental protocols

To assess the impact of different model input parameters, four experimental protocols were defined (Table 1), with each protocol having a slightly different emphasis.

Table 1. Characteristics of the experimental protocols.

Protocol	Cell	Monitoring period*	Training period*	Testing period*	Window blinds (% aperture)	Mechanical Ventilation
C4MVOFF	C4	10-16/07/2017	11/07/2017	12-16/07/2017	No window	OFF
C4MVON	C4	12-18/09/2017	13/09/2017	14-18/09/2017	No window	ON (22-8h), 1.75 ACH
C3MVOFF50	C3	10-16/07/2017	11/07/2017	12-16/07/2017	50% open	OFF
C3MVON50	C3	12-18/09/2017	13/09/2017	14-18/09/2017	50% open	ON (22-8h), 1.75 ACH

* On site measurements were recorded in the Test Cells during a 168-hour period. In the calibration phase, a 24-hour training period was considered. Calibration was then tested using an independent 120-hour period.

Sensitivity analysis and Bayesian calibration were conducted for each of the four protocols, via the following steps. Firstly, a protocol with no influence of solar radiation nor Mechanical Ventilation (MV) was analysed (C4MVOFF), in order to focus on the uncertainty related to the envelope's thermophysical properties (cell 4). Subsequently, MV was incorporated into the study (C4MVON) to assess the impact of the MV system within simulations. In parallel, cell 3 was configured with the window's blinds half open to quantify the parametric uncertainty

associated with solar radiation (C3MVOFF50). Finally, MV was once again incorporated into the previous protocol (C3MVON50).

3.3 Construction of the energy building model

The Cells described in the previous section were modelled in DesignBuilder v.4.7.0.027, recognized by the US-DOE [47], and coupled with the EnergyPlus v.9.0.1 simulation engine [48]. In this process, geometric, technical and construction information have been input into the simulation tool. Weather data measured by a local weather station were incorporated as known variables into the weather file used in simulations, assuming to have a negligible measurement error [49]. Since weather data provides hourly information for ambient variables, the temporal resolution of the model was also hourly, allowing thermal comfort assessment to be done.

In developing the building simulation model for the test cell, some simplifications and assumptions had to inevitably be made, mainly due to physical and simulation tool limitations. First, even though only Cells 3 and 4 were analysed, the model included all four cells and services rooms to take into account their shading and thermal influence. Each Cell and service room was considered to be a thermal zone. The thermal envelope was modelled by introducing the material layers of each construction system. The physical properties included in the available technical sheets were introduced into the simulation software, although there was some missing information, particularly in relation to density, conductivity and solar absorptance of materials.

Regarding MV, the schedule and air change rates set in the simulation model were as established in Table 1. Both air chambers in the roof and floor of the cells were modelled, as well as the interior and exterior MV air grids.

For considering a window blind aperture of 50% open, the window had to be modelled in two parts (upper and lower), so that the model resembles reality as close as is feasibly possible.

Since the cells are unoccupied, neither occupation loads nor occupation schedules were required, so it was not possible to assess the users' influence in this research. The Cells were analysed in free-running conditions, meaning no HVAC systems were considered, no lighting systems were activated and no extensive flow path equations were implemented for CFD analysis during the selected protocols. Lastly, given that the case study is located in an open

space, no shading effect by surrounding buildings was modelled. Shadow calculations were done following the Sky Diffuse Modelling Algorithm. The inside and outside surface convention algorithms applied were TARP [50] and DOE-2 [51], respectively. Surface heat conduction was modelled using the Conduction Transfer Function.

3.4 Sensitivity analysis of input setting variation

Sensitivity analysis has been conducted to determine the impact on output variation due to modifications in the input settings. The ultimate objective is to reduce the number of parameters to be considered in the calibration process, since the Bayesian method is computationally prohibitive in a high-dimensional parameter space [52] and an increase in the number of calibration parameters may lead to inaccuracy and ineffectiveness [53].

The parameter screening technique considered is the Morris method [49], later extended by Campolongo, Cariboni and Saltelli [54]. This method is widely used in building performance analysis because of its balance between low computational cost and accuracy [55], when compared to other approaches, such as Sobol' [56] or Standardized Rank Regression Coefficient [57]. Petersen, Kristensen and Knudsen [58] reported an extensive literature review where Morris method is used for BEM analysis.

The Morris method discretizes the parameter space, creating a grid of values from a pre-selected number of levels, dividing each parameter interval. Starting from an initial fixed point, the movement in the space is carried out along the axes, changing one parameter value at a time, while maintaining the remaining values. This one-step-at-a-time procedure allows the determination of an elementary effect (EE) for each trajectory (r) and for each parameter (k) [59], evaluating the influence of uncertain parameters over their whole range. The Morris method requires k+1 model simulations to calculate one EE for each of the k input variables [60]. Variables are ranked taking into account their relative effect on the reported output. It calculates the standard deviation (σ) (Equation 1) for each parameter's elementary effects (which provides a measure of the parameter's interaction with other parameters), as well as the modified mean (μ*) (Equations 2 and 3) (which quantifies the parameter's impact on the model output):

$$\sigma = \sqrt{\frac{1}{r} \cdot \sum_{n=1}^{r} (EE_n - \mu)^2} \quad \text{(Equation 1)}$$

$$\mu = \frac{1}{r} \cdot \sum_{n=1}^{r} EE_n \quad \text{(Equation 2)}$$

$$\mu^* = \frac{1}{r} \cdot \sum_{n=1}^{r} |EE_n| \quad \text{(Equation 3)}$$

Where:

σ: standard deviation

r: set of trajectories in which the space grid is sampled (independent EE)

EE_n: elementary effect (measures interactions with other parameters)

μ: mean of the value of the elementary effects

μ*: modified mean of the finite distribution of absolute values of the EE

According to Campolongo and Braddock [61], the standard deviation σ determines the spread (variance) of the finite distribution of EE values, indicating possible interactions with other variables. The same authors define the μ index as the sensitivity strength between the input variable and the reported output, caused by all first- and higher- order effects. The larger the μ index, the higher sensitivity an output has to an input variable. To provide a true importance measure and avoid cancellation effects due to negative elements of non-monotonic models, μ^* as proposed by Campolongo, Cariboni and Saltelli [54] is used.

For the sensitivity analysis, the necessary adjustments to the simulation file were made through a parametric definition [62], improving the management of large and complex BEM. This allows simulations to be automatically run to explore different values of the parameterized variables, reducing manual variation tasks. For this, the open source tool jEPlus v.1.7.2 [63] was used in combination with the R v.3.5.3 [64]statistical programming environment, through its sensitivity package v.1.16.2 [65].

In this case study, given that free-running conditions are considered (without heating and cooling systems), the impact of the input variables on the hourly indoor air temperatures of the Test Cells has been assessed to determine the dominant input parameters. Variable selection is explained in detail in subsection 4.1. The recommendations reported by Petersen, Kristensen

and Knudsen [58] regarding consistency in parameter ranking, were followed for setting the necessary variables in the Morris method: the number of repetitions of the design (r = 500), the number of levels (levels = 12) and the number of levels that are increased/decreased for computing effects (grid jump = 6), which Morris himself recommends to be at least half the number of levels [4949].

Table 2 shows the variables considered in this analysis for the four protocols. Design (central), minimum and maximum values are described for each parameter distribution, representing the uncertainties of the initial parameter values in the model. Minimum and maximum values correspond to the boundary conditions (upper and lower limits) of the building energy model variables and the ranges are determined considering the values established in the Spanish Technical Building Code [66]. Design values were the expected values of each parameter and were informed by the technical sheets of the building equipment and materials used in the Test Cells construction.

Table 2. Prior distribution functions considered in the experimental protocols for the sensitivity analysis.

ID	Parameter description	Protocols	Design value	Min/Max. value	Unit
BRICKc	Conductivity of brick	C4MVOFF/ON C3MVOFF50	0.8	0.7/0.9	W/m·K
BRICKd	Density of brick	C4MVOFF	1700	1600/1800	kg/m^3
BRICKsh	Specific heat of brick	C4MVOFF	1000	950/1050	J/kg·K
FAN	Fan efficiency (MV)	C4MVON C3MVON50	0.6	0.6/0.9	-
FLOW	Ventilation rate (MV)	C4MVON C3MVON50	1.75	1.50/2.00	ACH
FRAMEc	Conductance of frame	C3MVOFF/ON50	8.0	6.0/10.0	$W/m^2 \cdot K$
FRAMEsa	Solar absorptance of frame	C3MVOFF/ON50	0.6	0.5/0.8	-
INFIL	Infiltration rate	C4MVOFF/ON C3MVOFF/ON50	0.2	0.1/0.4	ACH
MASS	Thermal mass: thickness of concrete layer	C4MVOFF	0.2	0.1/0.3	m
MWc	Conductivity of mineral wool	C4MVOFF/ON C3MVOFF/ON50	0.035	0.03/0.04	W/m·K
MWd	Density of mineral wool	C4MVOFF	100	90/110	kg/m^3
MWsh	Specific heat of mineral wool	C4MVOFF/ON C3MVOFF50	840	830/850	J/kg·K

PURc	Conductivity of polyurethane	C4MVOFF/ON C3MVOFF50	0.02	0.02/0.03	W/m·K
PURd	Density of polyurethane	C4MVOFF/ON C3MVOFF50	40	30/50	kg/m^3
PURsh	Specific heat of polyurethane	C4MVOFF	1000	950/1050	J/kg·K
RENDta	Thermal absorptance of exterior mortar rendering	C4MVOFF/ON C3MVOFF/ON50	0.5	0.4/0.6	-
RUBta	Thermal absorptance of rubber	C4MVOFF	0.9	0.8/0.95	-
STEELta	Thermal absorptance of steel	C4MVOFF/ON C3MVOFF/ON50	0.9	0.8/0.95	-
WDWc	Conductivity of glazing	C3MVOFF/ON50	0.9	0.8/0.95	W/m·K
WDWst	Solar transmittance of glazing	C3MVOFF/ON50	0.9	0.8/0.95	-

Note: Min/Max. value refer to the boundary conditions considered in the building simulation model.

3.5 Calibration of the most influential parameters.

In the calibration process, Bayesian techniques were implemented due to the following advantages: (1) easy incorporation of prior information and expert knowledge; (2) capacity of computing probabilistic outcomes as reasonable expectations; (3) analysis of uncertainties associated to the predictions of model parameters; (4) possibility to consider multiple sources of uncertainty, regarding model inputs, model discrepancies due to the physical limitations of BEM, errors in field observations or noisy measurements; (5) updating posterior distributions based on prior knowledge and measured data [67]. The statistical formulation of this method was established by Kennedy and O'Hagan (Equation 4) [68]:

$$y(x) = \eta(x,t) + \delta(x) + \epsilon(x) \quad \text{(Equation 4)}$$

Where:

$y(x)$: field observations at known conditions x. Known conditions x considered in this paper were: dry-bulb outdoor air temperature, outdoor relative humidity and global horizontal irradiance. $\eta(x,t)$: energy model outputs, computed at known conditions x and with calibration parameters t.

$\delta(x)$: model discrepancy or bias (discrepancies between model and true physical behaviour).

$\epsilon(x)$: errors in measurements and observations.

This technique allows for consideration of parameter uncertainty, by specifying a prior distribution for parameters. This distribution includes the most likely range of possible values [42] taking into account building specifications, surveys or expert judgment. Prior distributions are then updated given measured data through Bayes' rule [67]: model simulation runs are used to identify which parameters are most likely to lead to the observed measurements and, then, Bayes' theorem is implemented to calculate a posterior parameter distribution. Thus, posterior distributions result from the combination of prior distributions (prior knowledge) and a likelihood that a set of parameters would yield the observed measurements. The likelihood function is based on how observed data $y(x)$, relates to calibration parameters t and observable inputs x, considering the model prediction $\eta(x,t)$, the model inadequacy $\delta(x)$ and the measurement error $\epsilon(x)$ [42]. The model discrepancy term captures the model bias and prevents the model overfitting during calibration [69].

In the Bayesian approach adopted, the building energy model outputs $\eta(x,t^f)$ and the discrepancy term $\delta(x)$ are modelled using GP, capturing the effects and interactions of individual parameters on the outputs through a covariance matrix (nonlinear multivariable region). GP does not impose a fixed functional form, defining its properties by its mean and covariance functions (Equations 5 to 9). Thus, GP are used as a surrogate model for the EnergyPlus tool, since the calibration approach would otherwise be toocomputationally expensive. A more detailed explanation of this is provided in [52].

$$\sum_{n,ij} = \frac{1}{\lambda_n} \cdot \exp\cdot\{-\sum_{k=1}^{p} \beta_k^n \left|x_{ik} - x_{jk}\right|^2 - \sum_{k\prime=1}^{q} \beta_{p+k\prime}^n \left|t_{ik\prime}^f - t_{jk\prime}^f\right|^2\} \quad \text{(Equation 5)}$$

$$\sum_{p,ij} = \frac{1}{\lambda_\delta} \cdot \exp\cdot\{-\sum_{k=1}^{p} \beta_k^\delta \left|x_{ik} - x_{jk}\right|^2\} \quad \text{(Equation 6)}$$

$$\mathrm{L}\left(z\middle|t^f,\beta^n,\lambda_n,\beta^\delta,\lambda_\delta,\lambda_\epsilon\right) \propto [\sum_z]^{-1/2}\cdot\exp\cdot\{-\frac{1}{2}(z-\mu)^T \sum_z^{-1}(z-\mu)\} \quad \text{(Equation 7)}$$

$$\sum_z = \sum_n + \begin{vmatrix} \sum_p + \sum_y, & 0 \\ 0 & 0 \end{vmatrix} \quad \text{(Equation 8)}$$

$$\sum_y = I_{\mathrm{n}} / \lambda_\epsilon \quad \text{(Equation 9)}$$

Where:

$\sum_n$: covariance function (depends on known input conditions x and calibration parameters t^f).

$\sum_\delta$: covariance function that depends on known input conditions x.

$\sum_y$: covariance matrix that accounts for observation errors.

β^n: correlation hyperparameter of the GP model for the simulator η.

β^{δ}: correlation hyperparameter of the GP model for the discrepancy term δ.

λ_{n}: precision hyperparameter of the GP model for the simulator η.

$\lambda_{\delta,}$: precision hyperparameter of the GP model for the discrepancy term δ.

λ_{ϵ}: precision hyperparameter of the GP model for the observation errors ϵ.

p: number of known input conditions x.

q: number of calibration parameters t.

z: n+m vector (establishes the relationship between the field observations y and predictions η).

L: refers to the normal likelihood function [52].

Although, GP is proven to provide the best accuracy of various meta-modelling techniques [70], its computational costs are high [71]. As building simulation is typically nonlinear, the posterior distribution usually results in an intractable expression [72], making it difficult to analytically sample from high-dimensional posterior distributions. To solve this, several studies recommend using Markov Chain Monte Carlo (MCMC) sampling [69], whose stationary distribution is the posterior distribution [73]. Following the step-by-step guidance provided by Chong and Menberg [52], the No-U-Turn Sampler, an extension of the Hamiltonian Monte Carlo algorithm was used, avoiding the random walk behaviour[67], allowing faster convergence for high-dimensional posterior distributions [74].

Latin hypercube sampling (LHS) was used to construct the simulated training set for the Bayesian calibration. LHS aims to effectively sample in the multi-dimensional space of the calibration parameters [75] and a generally accepted rule is to consider 10 LHS samples per parameter [76]. Thus, the number of simulations m considered was m=40, for 4 calibration parameters. Since the assessment of indoor thermal conditions is usually based on temporal data, it requires a higher precision than that of energy consumption (generally evaluated at a 12-month resolution: 12 measurements) [41], a 24-hour training period was used, with a total of 960 number of simulated points (Table 1).

The MCMC creates chains in the high probability regions (those with the highest impact on the output) for approximating the posterior distributions. The number of chains was fixed at 3, minimum value recommended by Annis, Miller and Palmeri [77]. The bigger the number of chains, the more computationally time-consuming the task [7171]. The number of iterations the algorithm runs was set at an initial value of 500 and the warm-up argument (number of steps

used to automatically tune the sampler) was set at 250, following the same authors suggestions [77]. Since the model normally begins in regions far away from the more plausible values (low probability regions), the warm-up argument reduces the influence of the starting values [52].

3.6 Uncertainty analysis: accuracy measurement of the building simulation model

As established by Ruiz and Bandera [78], the accuracy of a calibrated model may be measured using uncertainty analysis. Even though there is a lack of consensus on BEM calibration standards [26], the most common uncertainty indices used in calibration are defined in the ASHRAE Guideline 14:2002 [79]: the Normalized Mean Bias Error (NMBE), the Coefficient of Variation of the Root Mean Square Error (CVRMSE) and the Coefficient of Determination (R^2) (Equations 10 to 12), where: m_i are the measured values, s_i refers to the simulated data, n is the number of measured data points, ρ is the number of adjustable model parameters and $\bar{m}$ represents the mean of measured values.

$$NMBE=\frac{1}{\bar{m}} \cdot \frac{\sum_{i=1}^{n}(m_i-s_i)}{n-\rho} \cdot 100(\%) \quad \text{(Equation 10)}$$

$$CV(RMSE)=\frac{1}{\bar{m}} \cdot \sqrt{\frac{\sum_{i=1}^{n}(m_i-s_i)^2}{n-\rho}} \cdot 100(\%) \quad \text{(Equation 11)}$$

$$R^2=\left(\frac{n \cdot \sum_{i=1}^{n} m_i \cdot s_i - \sum_{i=1}^{n} m_i \cdot \sum_{i=1}^{n} s_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} m_i^2 - \left(\sum_{i=1}^{n} m_i\right)^2\right) \cdot \left(n \cdot \sum_{i=1}^{n} s_i^2 - \left(\sum_{i=1}^{n} s_i\right)^2\right)}} \right)^2 \quad \text{(Equation 12)}$$

According to the ASHRAE Guideline, the difference between the monitored and simulated data is sufficiently small if the mentioned metrics are within the thresholds shown in Table 3. Although these indices are generally used to calibrate energy consumption [14] and demand in simulation models [23], there are also several studies that use them to calibrate ambient data: outdoor temperatures, wind speed [80], solar radiation [81],indoor temperatures [82] or humidity [83].

Table 3. Uncertainty ranges according to ASHRAE Guideline 14:2002.

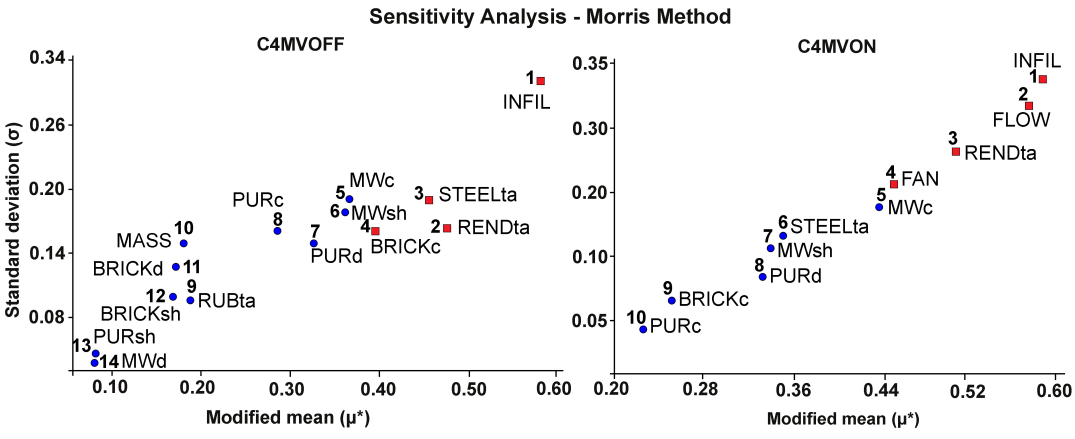
Calibration Frequency	Index	ASHRAE
Monthly	NMBE	±5%
	CV(RMSE)	15%
Hourly	NMBE	±10%
	CV(RMSE)	30%
Suggested	R^2	>0.75

4. Analysis and Results

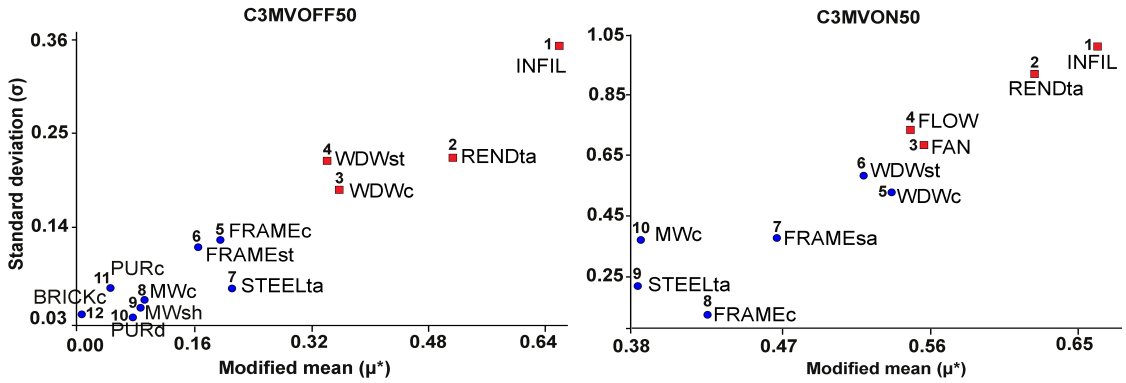
4.1 Sensitivity Analysis Results

As outlined in section 3.4, a sensitivity analysis has been conducted for each protocol, with the aim of determining the most influential parameters to be used in calibration.

Using the Morris method, the variables analysed in each protocol have been ranked by order of importance, based on the modified mean (μ*) and standard deviation (σ) values obtained (Figure 3). In each graph, the top right corner represents the variables with the highest μ* and σ values, corresponding to the most influential variables on the modelled indoor air temperature. On the contrary, the bottom left corner groups the variables with less importance on the output results.



(**a**) (**b**)



(**c**) (**d**)

Figure 3. Sensitivity analysis results (Morris Method): parameter ranking in **(a)** C4MVOFF; (**b**) C4MVON; (**c**) C3MVOFF50; and (**d**) C3MVON50. The calibration parameters used in each protocol are in red.

Results obtained for the first protocol (C4MVOFF) were used to screen-out non-sensitive parameters when assessing the second protocol (C4MVON). Specifically, the followingvariables were ruled out for the C4MVON protocol: density of mineral wool (MWd), specific heat of polyurethane (PURsh), density and specific heat of brick (BRICKd and BRICKsh), thermal mass (MASS) and thermal absorptance of rubber (RUBta). Similarly, from the third (C3MVOFF50) to the fourth protocol (C3MVON50), the least influential parameters were also screened-out: specific heat of mineral wool (MWsh), density of polyurethane (PURd) and conductivity of both polyurethane (PURc) and brick (BRICKc).

Only the top 4 most influential variables in each protocol (indicated in red in Figure 3) were used in calibration, since a higher number of variables may lead to a loss of posterior precision [52].

4.2 Bayesian Calibration Results

The Bayesian technique has been applied to calibrate the top 4 most influential parameters for each protocol, as defined in subsection 4.1. For the sake of brevity, only results of the C3MVON50 protocol (south-facing cell with MV and blinds 50% open), are provided in this section to provide an example of the calibration process. Results obtained in the remaining protocols are included in the Appendix A (Figures A.1 to A.3), along with the prior distributions used for all calibration parameters (Table A.1).

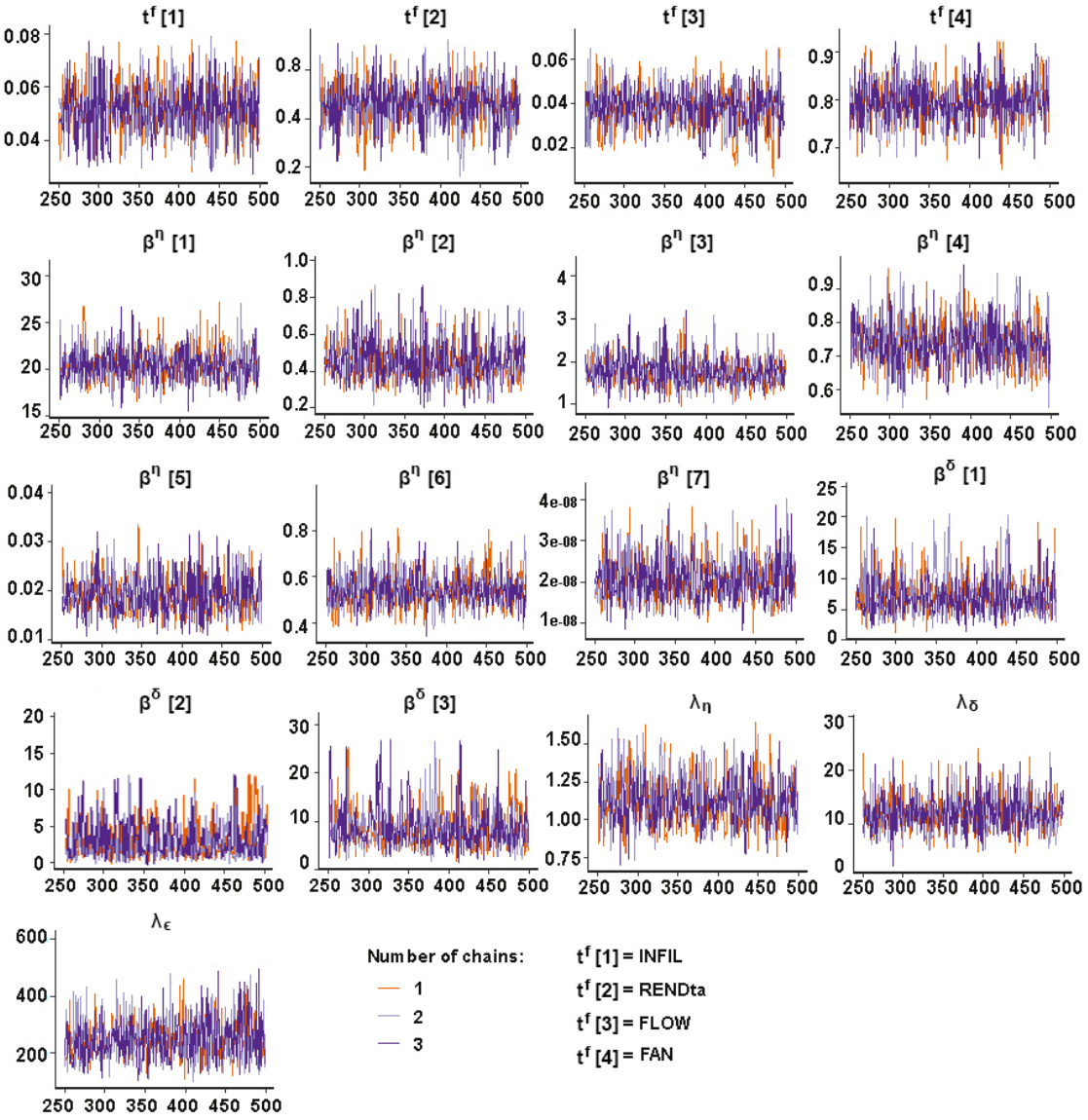


Figure 4. Trace plots of the four calibration parameters (t^f), GP correlation hyperparameters (β^η and β^δ) and GP precision hyperparameters (λ_η, λ_δ and λ_ϵ) set in the C3MVON50 protocol. X-axis represents the number of iterations (note that warm-up was set at 250) and Y-axis corresponds to the parameter values.

Figure 4 shows the trace plots of the four calibration parameters (t^f), GP correlation (β^η and β^δ) and GP precision hyperparameters (λ_η, λ_δ and λ_ϵ), used in the Bayesian calibration process for the C3MVON50 protocol. Their prior distributions were defined according to Chong and Menberg [52], and are as follows: β^η (a=1,b=0.3), β^δ (a=1, b=0.3), λ_η (a=10, b=10), λ_δ (a=10, b=0.3) and λ_ϵ (a=10, b=0.03), where a refers to the mean and b to the standard deviation.Visual inspection of the plots suggests that the sampling algorithm is exploring the posterior distribution efficiently, since the plots obtained look similar to a “fuzzy caterpillar” [84], being difficult to distinguish between individual chains and remaining around a constant value.

The low variability between and within the three MCMC chains, indicates that the number of chains being considered is enough to ensure model convergence to a common stationary distribution, with no need to increase the number of iterations or chains [6767]. Moreover, when comparing the variations between chains and the variations within chains for each parameter in the model, the potential scale reduction statistic (R_{hat}) is within 1.0±0.1 in all four protocols, proving that convergence was successfully achieved [73].

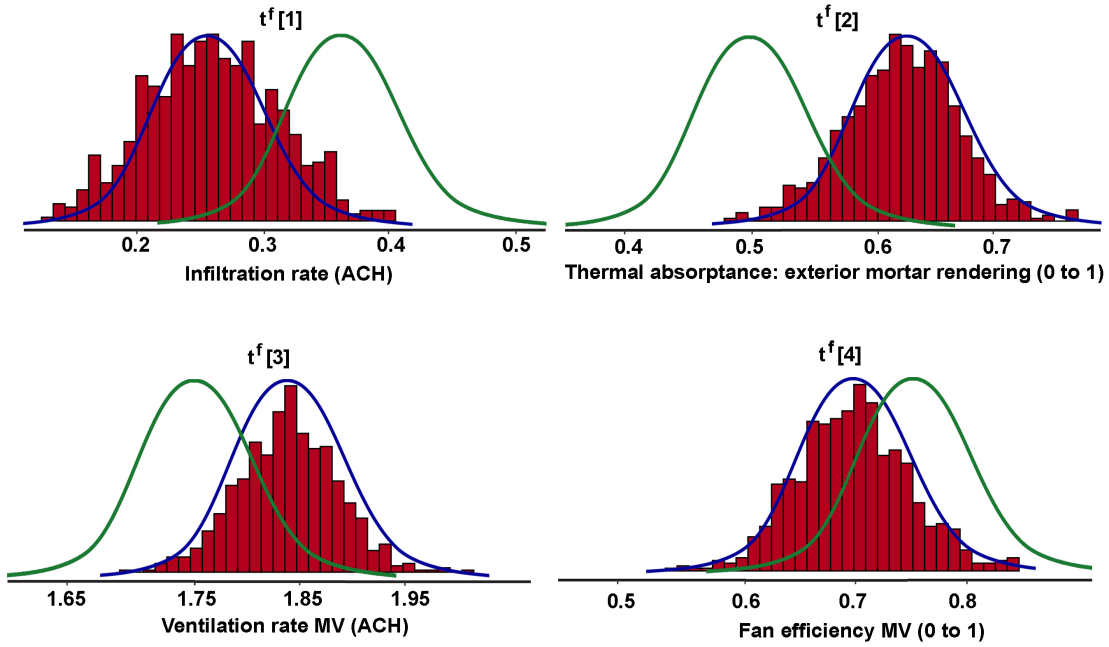


Figure 5. Posterior distributions obtained for C3MVON50 (red histogram), determined from a normal distribution of calibration parameters (blue line). The green line refers to prior uncertainty distributions.

The plausible normal posterior distributions of the four calibration parameters for C3MVON50, reported in the Bayesian process, are shown in Figure 5. Green lines represent prior uncertainty distributions used as input to the calibration process. When compared to the posterior distributions (blue lines indicate the most likely values to result based on the observed measurements) obtained in the Bayesian approach, it can be observed that previous estimations of the parameters needed to be refined in the model, particularly in the case of the thermal absorptance of the exterior mortar rendering (RENDta = tf[2]) and the infiltration rate (INFIL = tf[1]).

4.3 Uncertainty Analysis Results

This section assesses the accuracy of the calibrated building energy model in comparison to on-site measurements during the training and testing periods. Results for protocol C3MVON50 are shown here as an example. Results obtained for the remaining protocols can be found in Appendix B (Figures B.1 to B.6).

Figure 6 represents the comparison between monitored indoor air temperature (blue line) and simulated data, for both the uncalibrated model (green line) and after the Bayesian calibration (red line), during the 24-h training period.

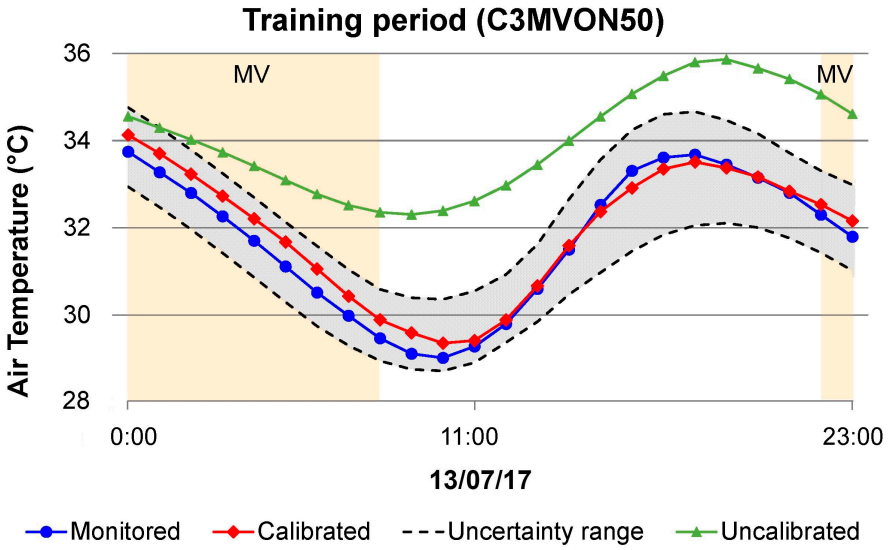


Figure 6. Comparison between monitored and simulated data during the 24-hour training period for the C3MVON50 protocol. The red line shows the prediction of the calibrated model. The uncertainty range (95% confidence intervals) relating to the posterior distributions of the calibrated model is shown in grey. MV indicates that mechanical ventilation is on.

It is observed that the calibration process has significantly improved the simulation model when compared to the uncalibrated model, particularly during minimum and maximum temperatures. Moreover, monitored data is within the uncertainty range (95% confidence intervals) of the calibrated model (indicated in grey), which is obtained by considering variations within the posterior distribution ranges.

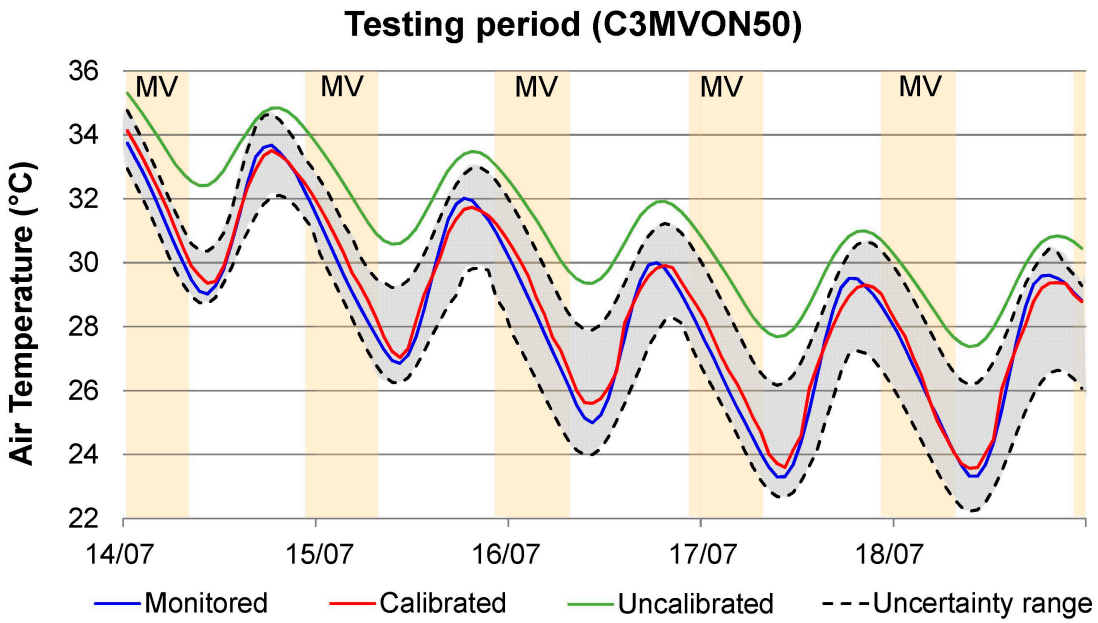


Figure 7. Comparison between monitored and simulated data during the 120-hour testing period for the C3MVON50 protocol. The red line shows the prediction of the calibrated model. The uncertainty range (95% confidence intervals) relating to the posterior distributions of the calibrated model is shown in grey. MV indicates that mechanical ventilation is on.

After Bayesian calibration, results obtained for the model for each protocol were tested for a longer period of time (120 hours) in order to check for bias in the evaluation process. The graphical results for the C3MVON50 protocol during the testing period are shown in Figure 7, where on site air temperature measurements (blue line) are compared to simulation outputs. Once again, both uncalibrated results (green line) and calibrated data (red line) are included. The graphic indicates model improvement using Bayesian calibration, which is crucial for hourly thermal comfort analysis. The uncertainty range (95% confidence intervals) is shown in Figure 7.

Statistical performance indices calculated following AHSRAE Guidelines (subsection 3.6) for each protocol are summarised in Table 4. A comparison is made between the performance of calibrated and uncalibrated models. To check for bias in the evaluation process, these indices are determined using an independent 120-hour dataset (testing period), different from the one used in the calibration process (training period). The values presented for the calibrated models refers to the best fit of the calibration (i.e. the central estimates for calibration parameters were used).

Table 4. Statistical indices obtained in the assessment of the calibrated models' accuracy with respect to test cell data for all four experimental protocols.

Protocol	Calibrated	NMBE (±10%)	CVRMSE (<30%)	R^2 (>0.75)	Max. $T_{difference}$ (°C)	P-value (T-Test)
C4MVOFF	No	8.92%	10.04%	0.61	2.63	0.49
	Yes	0.44%	0.90%	0.98	0.51	0.82
C4MVON	No	-7.46%	11.57%	0.72	2.41	0.25
	Yes	-0.22%	1.42%	0.97	0.74	0.70
C3MVOFF50	No	-8.35%	8.57%	0.84	3.31	0.38
	Yes	-0.22%	0.81%	0.98	0.75	0.78
C3MVON50	No	-9.31%	9.81%	0.93	4.42	0.43
	Yes	-0.78%	1.45%	0.98	0.74	0.72

Note: Results of the calibrated models are represented in grey.

Results show that, although the discrepancy between monitored and simulated air temperatures in all the four uncalibrated models were within the uncertainty ranges established by ASHRAE Guidelines, the maximum temperature differences (Max. $T_{difference}$) between measurements and predictions were significantly large. The worst agreement was obtained for C3MVON50 (south-facing cell with MV and blinds 50% open), which exceeded the monitored temperature by 4.0 ºC, which is unreasonable accuracy for hourly thermal comfort assessment. It should be highlighted that the statistical indices were noticeably improved after calibration in all four cases, significantly reducing the maximum temperature differences to below 0.75 ºC. Bearing in mind that the accuracy of the monitoring probes is ±0.5 ºC for temperature measurements in the range 10-30 ºC and ±1.0 ºC for 30-55 ºC, the model is considered to be well calibrated. In addition, p-values obtained are significantly higher than 0.05, meaning there is not enough evidence to conclude that the difference between the means of monitored and simulated data is greater than zero (null hypothesis is that the means of both samples are the same).

In the case of C3MVON50, despite meeting the requirements of ASHRAE, the uncalibrated model clearly differs from the monitored data. Although the energy simulation model reproduces the overall performance of the cell with no time shift, it generally overpredicts air temperatures, with significant discrepancies at the maximum and minimum peaks. Although the uncalibrated model does not reproduce the thermal conditions of the cell, the calibrated model tackles this issue. Calibration results obtained for C4MVOFF, C4MVON and C3MVOFF50 protocols,

regarding the thermophysical properties of the cells, were taken into account for the calibration of C3MVON50 protocol, which was possible given the different datasets analysed.

5. Discussion and comparison to other studies

It should be borne in mind that Bayesian calibration for highly accurate energy building models can be a time-consuming procedure, with significant computational costs, especially for large and high-definition spaces. As a result, monthly calibration is usually performed and studies tend to focus on energy consumption and demand assessment: Kang and Krarti [85] apply Bayesian methods to calibrate electricity and gas consumption, considering a model resolution of 12 months (12 data points with monthly average values); Sokol, Cerezo and Reinhart [86] use the same resolution taking into account the residential building stock as the calibration target for monthly and annual energy consumption; Nagpal, Mueller, Aijazi and Reinhart [87] calibrate electricity and chilled water usage for a 12-point resolution building model; or Yuan, Nian and Su [88] evaluates Bayesian posterior distributions of an office building model using average monthly electricity consumption. Few studies conducted to date, use hourly training periods. Again, this tend to be for energy consumption [73] and utility data analysis [89] rather than thermal conditions. Moreover, most of the mentioned studies assess Bayesian calibration in commercial buildings such as offices [88] and university buildings [90], while a small number of them present residential buildings as calibration targets [89], again putting great efforts into analysing energy use data.

The research approach adopted in this paper therefore presents several scientific innovations. Firstly, Bayesian calibration was implemented using hourly training periods (temperature data), instead of the commonly monthly approach, to provide a basis for thermal comfort assessment. This is done given that, instead of energy consumption and demand analysis, indoor thermal assessment is targeted, so monthly calibration with average values may be insufficiently detailed to tackle this analysis. Secondly, Test Cells that reproduce a bedroom of a typical residential Mediterranean building were used as case study adding to the building typologies studied in previous research. This allowed the viability of the calibration technique to be tested under a controlled setting. Thirdly, in this research a 24-hour training period and a 120-hour testing period were considered, which meant a larger amount of computational running time

was required compared to monthly calibration (Bayesian hourly calibration for this case study took on average 12-hours processing time for each protocol). To tackle this issue and in comparison to other studies where only Bayesian approaches were implemented [90], this technique was complemented with a sensitivity analysis for identifying the most influential variables on the output results. The combination of both methods has affordably reduced computation time and has led to a considerable improvement in prediction, in contrast to uncalibrated results. Besides, incorporating sensitivity analysis may tackle the inconvenience of needing large amount of input data to properly calibrate BEM to an accurate level.

Besides, using the Bayesian calibration, the accuracy performances indices established by the ASHRAE Guidelines were noticeably improved: NMBE and CVRMSE indices were below 1.0% and 1.5%, respectively, in contrast to the initial values (which were up to 9.3% and 11.57%, respectively), improving the model's accuracy by around 80%, while other Bayesian calibration studies reported improvement results of around a 40% [86]. Performances indices of the ASHARE for BEM accuracy verification have proven to be substantially permissive when calibrating thermal conditions (since temperature differences were significantly high), given that guideline ranges tend to be used for energy demand and consumption [26]. Nevertheless, this paper has presented a viable complementary calibration methodology that achieves better and higher-resolution performance for hourly thermal comfort assessment.

6. Conclusions: Limitations and Future Research

This paper demonstrates the viability of a calibration methodology, based on both sensitivity analysis and Bayesian techniques, for building energy models, when predicting hourly temperature data. Monitored Test Cell dry-bulb indoor air temperature data was compared to simulated predictions under different experimental protocols, with the accuracy of the energy models assessed through uncertainty indices. Bayesian calibration was able to achieve a significantly better prediction of the Test Cell temperature data compared to the uncalibrated protocols, with an average improvement of around 80%.

Nonetheless, the specific results reported in this paper may only be applied to housing buildings or small single zone units (such as flats or small offices) with limited ventilation and few wall partitioning, where a global performance of the space or unit is assessed. Furthermore, a key

limitation of this research was that it was carried out in an unoccupied, highly controlled Test Cell environment with free-running conditions (no HVAC systems).

Whilst the use of Tests Cells may be ideal for assessing the suitability of the proposed methodology under various experimental protocols, the findings may not be applicable to more complex case studies or real buildings. Taking into account the conclusions reported in this paper, future research should test this methodology in real building models, with the evaluation of its viability and accuracy in models with different grades of complexity and definition. In addition, given the limitations of the current research, the impact of adding more study variables into the building energy models (e.g. HVAC systems, occupant density and use schedules, flow paths, etc.) should also be extensively analysed. Likewise, comparing the impact of using both simplified and complex forms of algorithms and energy equations in building energy models may also provide useful information in addressing the simplification process whilst reducing simulation time.

In terms of indoor thermal assessment calibration, an extensive evaluation of the combination of both detailed hourly and monthly on site data of representative seasonal periods may also be a strategy worth analysing on the results' accuracy, since such results were not found in the literature. Besides, they may also report useful and clarifying results for calibration comparison between the building and stock levels for retrofitting purposes.

Summarizing, despite of a possible further calibration optimization, which highly depends on computational burden, model complexity, input measurements, building level information or resources availability, these results prove that implementing a first level statistical calibration-simulation methodology, combining sensitivity analyses and Bayesian techniques, significantly improves the accuracy of hourly indoor air temperature simulation outputs when compared to uncalibrated models. Therefore, the methodology presented is considered to be a useful and novel automated building calibration approach that can achieve good improvement in simulation results, within the limitations of computational resources and building data availability.

Funding and Acknowledges: The authors wish to acknowledge the financial support provided by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund through the research project "Parametric Optimisation of Double Skin

Facades in the Mediterranean Climate for the Improvement of Energy Efficiency in Climate Change Scenarios" (BIA2017-86383-R). Calama-González also acknowledges the support of the FPU Program of the Spanish Ministry of Education, Culture and Sport (FPU17/01375) and economical support to conduct a temporary stay in the IEDE-UCL (EST18/00296).

Author Contributions: Á.L.L.-R. and R.S. conceived and designed the experiments; all authors performed the experiments; all authors analysed the data; all authors have written, reviewed, and approved the final manuscript.

Declarations of interest: none. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References:

[1] Norris M, Shiels P. Regular national report on housing developments in European countries. Synthesis report. Dublin, Ireland: The Housing Unit; 2004
[2] R. Hartless R. Application of energy performance regulations to existing buildings. Final report of the Task B4, ENPER TEBUC Project, SAVE 4.1031/C/00-018. Watford, UK: Building Research Establishment; 2003.
[3] Balaras CA, Gaglia AG, Georgopoulou E, Mirasgedis S, Sarafidis Y, Lalas DP. European residential buildings and empirical assessment of the Hellenic building stock, energy consumption, emissions and potential energy savings. Build Environ 2007;42:1298–314. https://doi.org/10.1016/j.buildenv.2005.11.001.
[4] Li Y, Rezgui Y, Zhu, H. District heating and cooling optimization and enhancement–Towards integration of renewables, storage and smart grid. Renewable and Sustainable Energy Reviews 2017;72:281-294. https://doi.org/10.1016/j.rser.2017.01.061
[5] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A Policy Framework for Climate and Energy in the Period from 2020 to 2030, 2014. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2014%3A15%3AFIN (accessed on 06 April 2019).
[6] Gao H, Koch C, Wu Y. Building information modelling based building energy modelling: A review. Appl Energy 2019:320–343. https://doi.org/10.1016/j.apenergy.2019.01.032.
[7] Augenbroe G. Trends in building simulation. Build Environ 2002;37:891–902. https://doi.org/10.1016/S0360-1323(02)00041-0.
[8] Taylor J, Symonds P, Wilkinson P, Heaviside C, Macintyre H, Davies M, et al. Estimating the Influence of Housing Energy Efficiency and Overheating Adaptations on Heat-Related Mortality in the West Midlands, UK. Atmosphere (Basel) 2018;9:190-207. https://doi.org/10.3390/atmos9050190.
[9] Royapoor M, Roskilly T. Building model calibration using energy and environmental data. Energy Build 2015. https://doi.org/10.1016/j.enbuild.2015.02.050.
[10] de Wilde P. The gap between predicted and measured energy performance of buildings: A framework for investigation. Autom Constr 2014. https://doi.org/10.1016/j.autcon.2014.02.009.
[11] van Dronkelaar C, Dowson M, Spataru C, Mumovic D. A Review of the Regulatory Energy Performance Gap and Its Underlying Causes in Non-domestic Buildings. Front Mech Eng 2016. https://doi.org/10.3389/fmech.2015.00017.
[12] Karlsson F, Rohdin P, Persson ML. Measured and predicted energy demand of a low energy building: Important aspects when using building energy simulation. Build Serv Eng Res Technol 2007. https://doi.org/10.1177/0143624407077393.
[13] Raftery P, Keane M, Costa A. Calibrating whole building energy models: Detailed case study using hourly measured data. Energy Build 2011. https://doi.org/10.1016/j.enbuild.2011.09.039.

[14] Monetti V, Davin E, Fabrizio E, André P, Filippi M. Calibration of building energy simulation models based on optimization: A case study. Energy Procedia Elsevier Ltd 2015;78:2971–2976. https://doi.org/10.1016/j.egypro.2015.11.693.
[15] Snyder SC, Maor I. Calibrated Building Energy Simulation in Practice: Issues, Approaches, and Case Study Example. ASHRAE Trans 2015.
[16] Fumo N. A review on the basics of building energy estimation. Renew Sustain Energy Rev 2014. https://doi.org/10.1016/j.rser.2013.11.040.
[17] Clarke JA, Strachan PA, Pernot C. Approach to the calibration of building energy simulation models. ASHRAE Trans., 1993.
[18] Reddy TA, Maor I, Panjapornpon C. Calibrating detailed building energy simulation programs with measured data—part I: General methodology (RP-1051). HVAC R Res 2007. https://doi.org/10.1080/10789669.2007.10390952.
[19] Reddy T.A. Literature review on calibration of building energy simulation programs: Uses, problems, procedure, uncertainty, and tools. In ASHRAE Transactions 2006;12:226–240.
[20] Parker J, Cropper P, Shao L. A Calibrated Whole Building Simulation Approach to Assessing Retrofit Options for Birmingham Airport. First Build Simul Optim Conf 2012.
[21] Reddy T.A., Maor I. Procedures for Reconciling Computer-Calculated Results with Measured Energy Data, Technical Report Prepared for ASHARE Research Project 1051-RP, Atlanta, GA, 2006.
[22] Heo Y, Choudhary R, Augenbroe GA. Calibration of building energy models for retrofit analysis under uncertainty. Energy Build 2012;47:550–560. https://doi.org/10.1016/j.enbuild.2011.12.029.
[23] Martínez S, Erkoreka A, Eguía P, Granada E, Febrero L. Energy characterization of a PASLINK test cell with a gravel covered roof using a novel methodology: Sensitivity analysis and Bayesian calibration. J Build Eng 2019;22:1–11. https://doi.org/10.1016/j.jobe.2018.11.010.
[24] Chaudhary G, New J, Sanyal J, Im P, O'Neil Z, Garg V. Evaluation of autotune calibration against manual calibration of building energy models. Applied Energy 2016;182:115-134. http://doi.org/10.1016/j.apenergy.2016.08.073.
[25] Martínez S, Eguía P, Granada E, Moazami A, Hamdy M. A performance comparison of multi-objective optimization-based approaches for calibrating white-box building energy models. Energy Build 2020. https://doi.org/10.1016/j.enbuild.2020.109942.
[26] Coakley D, Raftery P, Keane M. A review of methods to match building energy simulation models to measured data. Renew Sustain Energy Rev 2014. https://doi.org/10.1016/j.rser.2014.05.007.
[27] Fabrizio E, Monetti V. Methodologies and advancements in the calibration of building energy models. Energies 2015. https://doi.org/10.3390/en8042548.
[28] Tahmasebi F, Mahdavi A. Optimization-based simulation model calibration using sensitivity analysis. Simulace budov a Tech. prostředí 2012 7° Konf. IBPSA-CZ, 2012.
[29] Roberti F, Oberegger UF, Gasparella A. Calibrating historic building energy models to hourly indoor air and surface temperatures: Methodology and case study. Energy Build 2015. https://doi.org/10.1016/j.enbuild.2015.09.010.
[30] Zheng O, Eisenhower B. Leveraging the analysis of parametric uncertainty for building energy model calibration. Build Simul 2013. https://doi.org/10.1007/s12273-013-0125-8.
[31] Manfren M, Aste N, Moshksar R. Calibration and uncertainty analysis for computer models - A meta-model based approach for integrated building energy simulation. Appl Energy 2013. https://doi.org/10.1016/j.apenergy.2012.10.031.
[32] Chen J, Gao X, Hu Y, Zeng Z, Liu Y. A meta-model-based optimization approach for fast and reliable calibration of building energy models. Energy 2019. https://doi.org/10.1016/j.energy.2019.116046.
[33] Hong T, Kim J, Jeong J, Lee M, Ji C. Automatic calibration model of a building energy simulation using optimization algorithm. Energy Procedia, 2017. https://doi.org/10.1016/j.egypro.2017.03.855.
[34] Sanyal J, New J, Edwards R. Supercomputer assisted generation of machine learning agents for the calibration of building energy models. ACM Int. Conf. Proceeding Ser., 2013. https://doi.org/10.1145/2484762.2484818.
[35] Cornaro C, Bosco F, Lauria M, Puggioni VA, De Santoli L. Effectiveness of automatic and manual calibration of an office building energy model. Appl Sci 2019. https://doi.org/10.3390/app9101985.
[36] Menberg K, Heo Y, Choudhary R. Influence of error terms in Bayesian calibration of energy system models. J Build Perform Simul 2019;12:82–96. https://doi.org/10.1080/19401493.2018.1475506.
[37] Sun K, Hong T, Taylor-Lange SC, Piette MA. A pattern-based automated approach to building energy model calibration. Appl Energy 2016;165:214–224. https://doi.org/10.1016/j.apenergy.2015.12.026.
[38] Mihai A, Zmeureanu R. Journal of Building Performance Simulation Bottom-up evidence-based calibration of the HVAC air-side loop of a building energy model Bottom-up evidence-based calibration of the HVAC air-side loop of a building energy model. J Build Perform Simul 2017;10:105–123. https://doi.org/10.1080/19401493.2016.1152302.
[39] Faggianelli GA, Mora L, Merheb R. Uncertainty quantification for Energy Savings Performance Contracting: Application to an office building. Energy Build 2017;152:61–72. https://doi.org/10.1016/j.enbuild.2017.07.022.
[40] Menberg K, Heo Y, Choudhary R. Sensitivity analysis methods for building energy models: Comparing computational costs and extractable information. Energy Build 2016;133:433–445. https://doi.org/10.1016/j.enbuild.2016.10.005.

[41] Lim H, Zhai ZJ. Influences of energy data on Bayesian calibration of building energy model. Appl Energy 2018:686–698. https://doi.org/10.1016/j.apenergy.2018.09.156.
[42] Riddle M, Muehleisen RT. A guide to Bayesian calibration of building energy models. 2014 ASHRAE/IBPSA-USA Build. Simul. Conf., 2014.
[43] Heo Y, Graziano DJ, Guzowski L, Muehleisen RT. Evaluation of calibration efficacy under different levels of uncertainty. J Build Perform Simul 2015;8(3):135-144. http://dio.org/10.1080/19401493.2014.896947.
[44] Cattarin G, Causone F, Kindinis A, Pagliano L. Outdoor test cells for building envelope experimental characterisation – A literature review. Ren Sust Energy Rev 2016;54:606-625. http://dio.org/10.1016/j.rser.2015.10.012.
[45] León-Rodríguez ÁL, Suárez R, Bustamante P, Campano MÁ, Moreno-Rangel D. Design and Performance of Test Cells as an Energy Evaluation Model of Facades in a Mediterranean Building Area. Energies 2017;10:1816-1832. https://doi.org/10.3390/en10111816.
[46] Calama-González CM, Suárez R, León-Rodríguez ÁL, Domínguez-Amarillo S. Evaluation of thermal comfort conditions in retrofitted facades using test cells and considering overheating scenarios in a mediterranean climate. Energies 2018;11,788-807. https://doi.org/10.3390/en11040788.
[47] *DOE U. S.* Department of Energy, Washington DC, 2017. Available online: http://www.energy.gov (accessed on 06 April 2019).
[48] U.S. Department of Energy. *EnergyPlus Energy Simulation Software*, 2017. Available online: http://apps1.eere.energy.gov/buildings/energyplus (accessed on 06 April 2019).
[49] Morris MD. Factorial sampling plans for preliminary computational experiments. Technometrics 1991;33:161–174. https://doi.org/10.1080/00401706.1991.10484804.
[50] Walton GN. Thermal Analyisis Research. Program Reference Manual. National Bureau of Standards. U.S. Department of Commerce. In NBS Publications, Washintong, DC;1-296 1983, p.
[51] Lawrence Berkeley Laboratory (LBL), 1994. DOE2.1E-053 source code.
[52] Chong A, Menberg K. Guidelines for the Bayesian calibration of building energy models. Energy Build 2018;174:527–547. https://doi.org/10.1016/j.enbuild.2018.06.028.
[53] Lim H, Zhai ZJ. Review on stochastic modeling methods for building stock energy prediction. Build Simul 2017;10:607–624. https://doi.org/10.1007/s12273-017-0383-y.
[54] Campolongo F, Cariboni J, Saltelli A. An effective screening design for sensitivity analysis of large models. Envir Mod Software 2017;22:1509-1518. https://doi.org/10.1016/j.envsoft.2006.10.004.
[55] Wei T. A review of sensitivity analysis methods in building energy analysis. Renew Sustain Energy Rev 2013;20:411–419. https://doi.org/10.1016/j.rser.2012.12.014.
[56] Tian W, Song J, Li Z, de Wilde P. Bootstrap techniques for sensitivity analysis and model selection in building thermal performance analysis. Appl Energy 2014;135:320–328. https://doi.org/10.1016/j.apenergy.2014.08.110.
[57] Escandón R, Ascione F, Bianco N, Mauro GM, Suárez R, Sendra JJ. Thermal comfort prediction in a building category: Artificial neural network generation from calibrated models for a social housing stock in southern Europe. Appl Therm Eng 2019:492–505. https://doi.org/10.1016/j.applthermaleng.2019.01.013.
[58] Petersen S, Kristensen MH, Knudsen MD. Prerequisites for reliable sensitivity analysis of a high fidelity building energy model. Energy Build 2019;183:1–16. https://doi.org/10.1016/j.enbuild.2018.10.035.
[59] Ravalico JK, Maier HR, Dandy GC, Norton J, Croke B. A comparison of sensitivity analysis techniques for complex models for environmental management. MODSIM05 Int Congr Model Simul Adv Appl Manag Decis Mak Proc 2016.
[60] King DM, Perera BJC. Morris method of sensitivity analysis applied to assess the importance of input variables on urban water supply yield - A case study. J Hydrol 2013. https://doi.org/10.1016/j.jhydrol.2012.10.017.
[61] Campolongo F, Braddock R. The use of graph theory in the sensitivity analysis of the model output: A second order screening method. Reliab Eng Syst Saf 1999;64(1):1-12. https://doi.org/10.1016/S0951-8320(98)00008-8.
[62] Zhang Y, Korolija I. Performing complex parametric simulations with jEPlus. in SET2010 9th International Conference on Sustainable Energy Technologies, Shanghai, China, 2010.
[63] Zhang Y, Korolija I. jEPlus - An EnergyPlus simulation manager for parametrics. jEPlus 1.7.2. 2016. Available online: http://sourceforge.net/projects/jeplus (accessed on 06 April 2019).
[64] R v.3.5.3. R Core Team. GNU GPL v2. Available online: https://www.r-project.org/ (accessed on 06 April 2019).
[65] Pujol G, Iooss B, with contributions from Khalid AJ, Veiga SD, Fruth J, Gilquin L, Guillaume J, Le-Gratiet L, Lemaitre P, Ramos B, Touati T, Weber F. R Package. Sensitivity: Global Sensitivity Analysis of Model Outputs. V 1.16.2. 2016. Available on: https://cran.r-project.org/web/packages/sensitivity/index.html (accessed 14 November 2019)
[66] Spanish Building Technical Code (Código Técnico de la Edificación). Construction Elements Catalogue (Catálogo de Elementos Constructivos). Instituto Eduardo Torroja, CEPCO y AICIA. 2011. Available online: https://itec.cat/cec/ (accessed on 14 November 2019).
[67] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. Taylor and Francis 2014; 2.
[68] Kennedy MC, O'Hagan A. Bayesian calibration of computer models. J R Stat Soc Ser B (Statistical Methodol 2001;63:425–464. https://doi.org/10.1111/1467-9868.00294.

[69] Menberg K, Heo Y, Choudhary R. Efficiency and Reliability of Bayesian Calibration of Energy Supply System Models in International Building Performance Simulation Association IBPSA, 2017. https://doi.org/10.26868/25222708.2017.315.
[70] Li Q, Augenbroe G, Brown J. Assessment of linear emulators in lightweight Bayesian calibration of dynamic building energy models for parameter estimation and performance prediction. Energy Build 2016;124:194–202. https://doi.org/10.1016/j.enbuild.2016.04.025.
[71] Lim H, Zhai ZJ. Comprehensive evaluation of the influence of meta-models on Bayesian calibration. Energy Build 2017;155:66–75. https://doi.org/10.1016/j.enbuild.2017.09.009.
[72] Higdon D, Kennedy M, Cavendish JC, Cafeo JA, Ryne RD. Combining field data and computer simulations for calibration and prediction. SIAM J Sci Comput 2005;26:448–466. https://doi.org/10.1137/S1064827503426693.
[73] Chong A, Lam KP, Pozzi M, Yang J. Bayesian calibration of building energy models with large datasets. Energy Build 2017;154:343–355. https://doi.org/10.1016/j.enbuild.2017.08.069.
[74] Neal RM. Probabilistic inference using markov chain monte carlo methods. 1993. Technical Report CRG-TR-93-1. Department of Computer Science. University of Toronto.
[75] Helton JC, Johnson JD, Sallaberry CJ, Storlie CB. Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliab Eng Syst Saf 2006;91:1175–1209. https://doi.org/10.1016/j.ress.2005.11.017.
[76] Loeppky JL, Sacks J, Welch WJ. Choosing the sample size of a computer experiment: A practical guide. Technometrics 2009;51:366–376. https://doi.org/10.1198/TECH.2009.08040.
[77] Annis J, Miller BJ, Palmeri TJ. Bayesian inference with Stan: A tutorial on adding custom distributions. Behav Res Methods 2017;49:863–886. https://doi.org/10.3758/s13428-016-0746-9.
[78] Ruiz G, Bandera C. Validation of Calibrated Energy Models: Common Errors. Energies 2017;10: 1587-1606. https://doi.org/10.3390/en10101587.
[79] ASHRAE (American Society of Heating, Ventilating, and Air Conditioning Engineers). *Guideline 14-2002, Measurement of Energy and Demand Savings; Technical Report; American Society of Heating, Ventilating, and Air Conditioning Engineers*, Atlanta, GA, USA, 2002.
[80] Chan ALS. Generation of typical meteorological years using genetic algorithm for different energy systems. Renew Energy 2016;90:1–13. https://doi.org/10.1016/j.renene.2015.12.052.
[81] Reinhart CF, Andersen M. Development and validation of a Radiance model for a translucent panel. Energy Build 2006;38:890–904. https://doi.org/10.1016/j.enbuild.2006.03.006.
[82] Yasin M, Scheidemantel E, Klinker F, Weinläder H, Weismann S. Generation of a simulation model for chilled PCM ceilings in TRNSYS and validation with real scale building data. J Build Eng 2019;22:372–382. https://doi.org/10.1016/j.jobe.2019.01.004.
[83] Martínez-Ibernón A, Aparicio-Fernández C, Royo-Pastor R, Vivancos JL. Temperature and humidity transient simulation and validation in a measured house without a HVAC system. Energy Build 2016;131:54–62. https://doi.org/10.1016/j.enbuild.2016.08.079
[84] Lee MD, Wagenmakers EJ. Bayesian cognitive modeling: A practical course. New York, NY: Cambridge University Press; 2014
[85] Kang Y, Krarti M. Bayesian-Emulator based parameter identification for calibrating energy models for existing buildings. Build Simul 2016:411–28. https://doi.org/10. 1007/s12273-016-0291-6.
[86] Sokol J, Cerezo Davila C, Reinhart CF. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. Energy Build 2017;134:11–24. https://doi.org/10.1016/j.enbuild.2016.10.050.
[87] Nagpal S, Mueller C, Aijazi A, Reinhart CF. A methodology for auto-calibrating urban building energy models using surrogate modelling techniques. J Build Perform Simul 2018:1–16. https://doi.org/10.1080/19401493.2018.1457722.
[88] Yuan J, Nian V, Su B. A Meta Model Based Bayesian Approach for Building Energy Models Calibration. Energy Procedia, 2017. https://doi.org/10.1016/j.egypro.2017.12.665.
[89] Kristensen MH, Choudhary R, Petersen S. Bayesian calibration of building energy models: Comparison of predictive accuracy using metered utility data of different temporal resolution. Energy Procedia, 2017. https://doi.org/10.1016/j.egypro.2017.07.322.
[90] Carstens H, Xia X, Yadavalli S. Low-cost energy meter calibration method for measurement and verification. Appl Energy 2017. https://doi.org/10.1016/j.apenergy.2016.12.028

Appendix A.

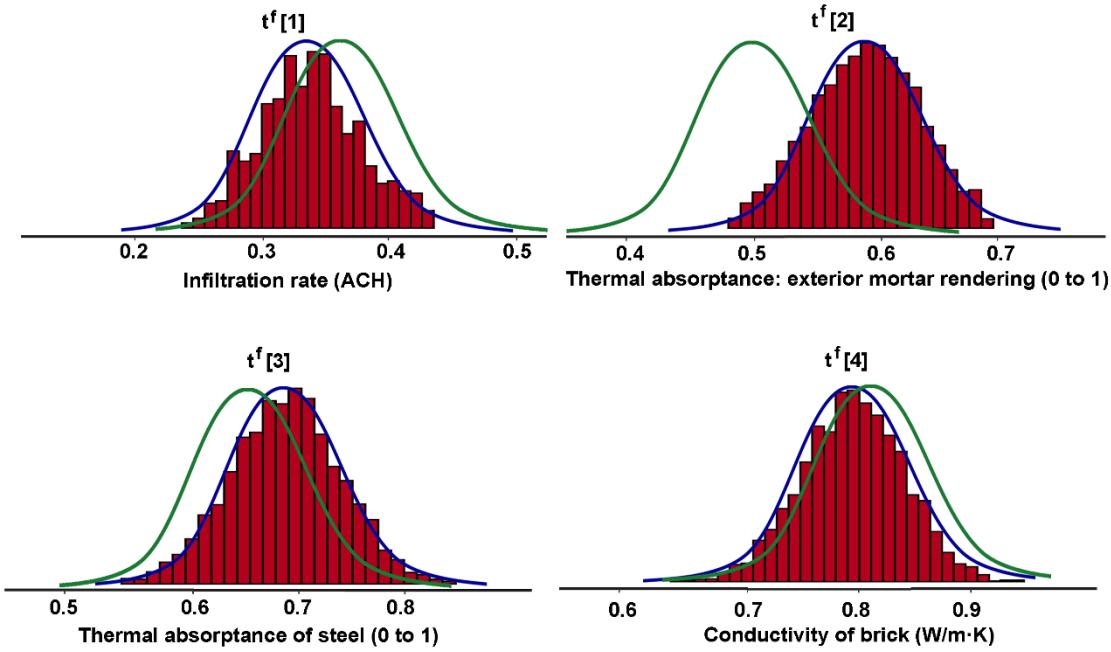


Figure A.1. Posterior distributions obtained in C4MVOFF (red histogram), determined from a normal distribution of calibration parameters (blue line). Green line refers to prior uncertainty distributions.

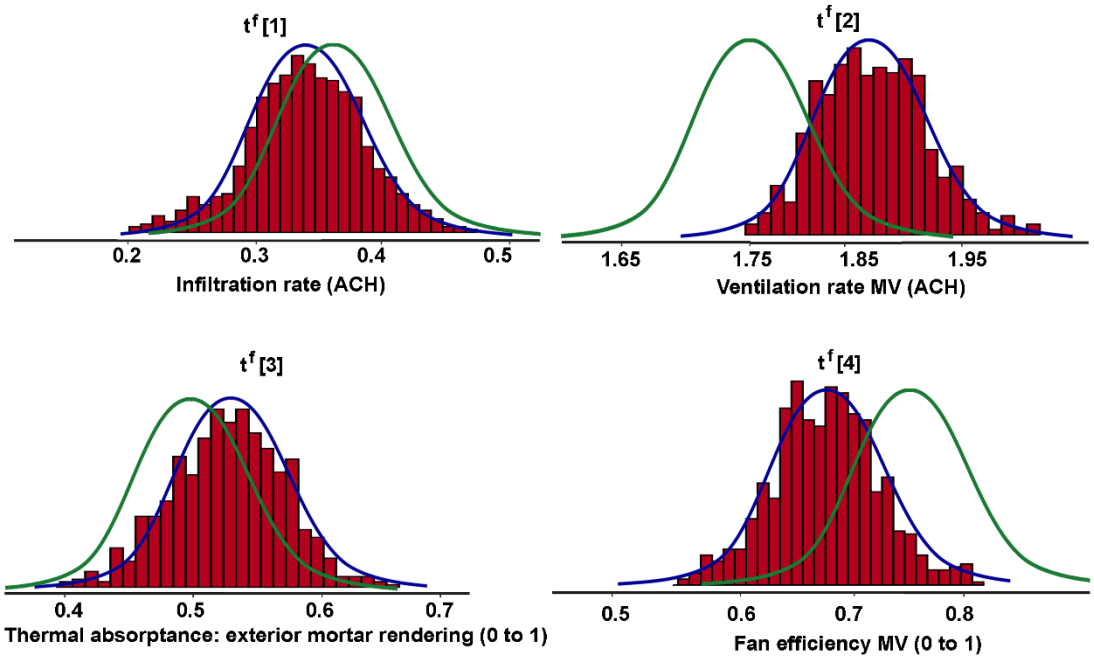


Figure A.2. Posterior distributions obtained in C4MVON (red histogram), determined from a normal distribution of calibration parameters (blue line). Green line refers to prior uncertainty distributions.

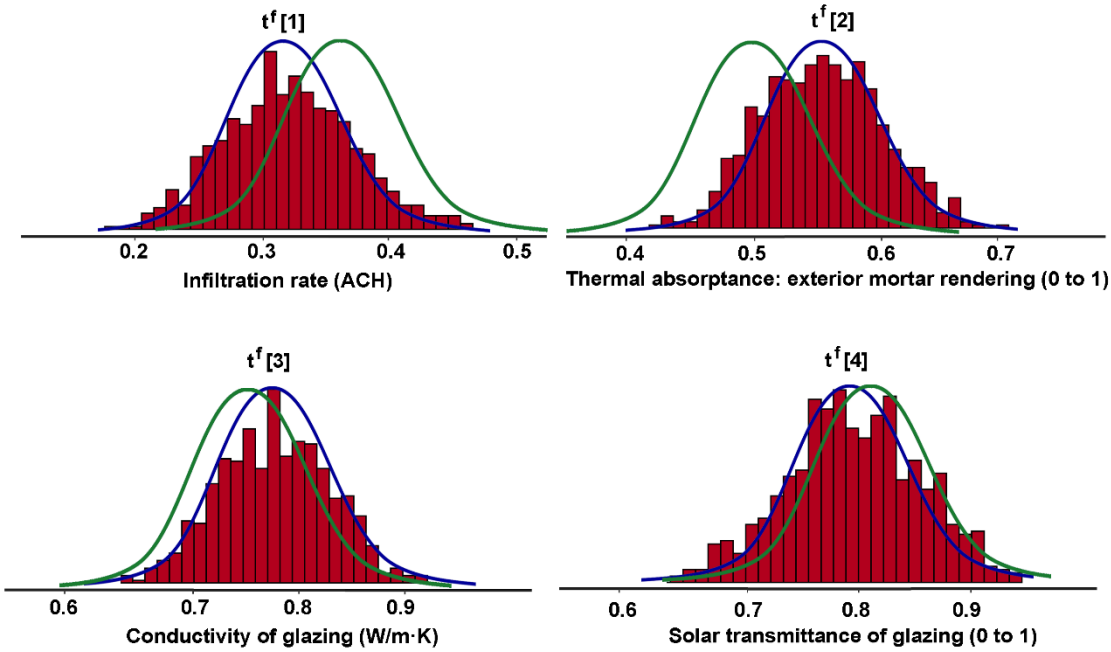


Figure A.3. Posterior distributions obtained in C3MVOFF50 (red histogram), determined from a normal distribution of calibration parameters (blue line). Green line refers to prior uncertainty distributions.

Table A.1 Normal prior distributions considered for each calibration parameter used.

ID	Parameter description	Protocols	Mean (μ)	Standard deviation (σ)
BRICKc	Conductivity of brick	C4MVOFF	0.80	0.03
FAN	Fan efficiency (MV)	C4MVON C3MVON50	0.70	0.05
FLOW	Ventilation rate (MV)	C4MVON C3MVON50	1.75	0.06
INFIL	Infiltration rate	C4MVOFF/ON C3MVOFF/ON50	0.35	0.05
RENDta	Thermal absorptance of exterior mortar rendering	C4MVOFF/ON C3MVOFF/ON50	0.45	0.05
STEELta	Thermal absorptance of steel	C4MVOFF	0.90	0.025
WDWc	Conductivity of glazing	C3MVOFF50	0.70	0.05
WDWst	Solar transmittance of glazing	C3MVOFF50	0.80	0.03

Appendix B.

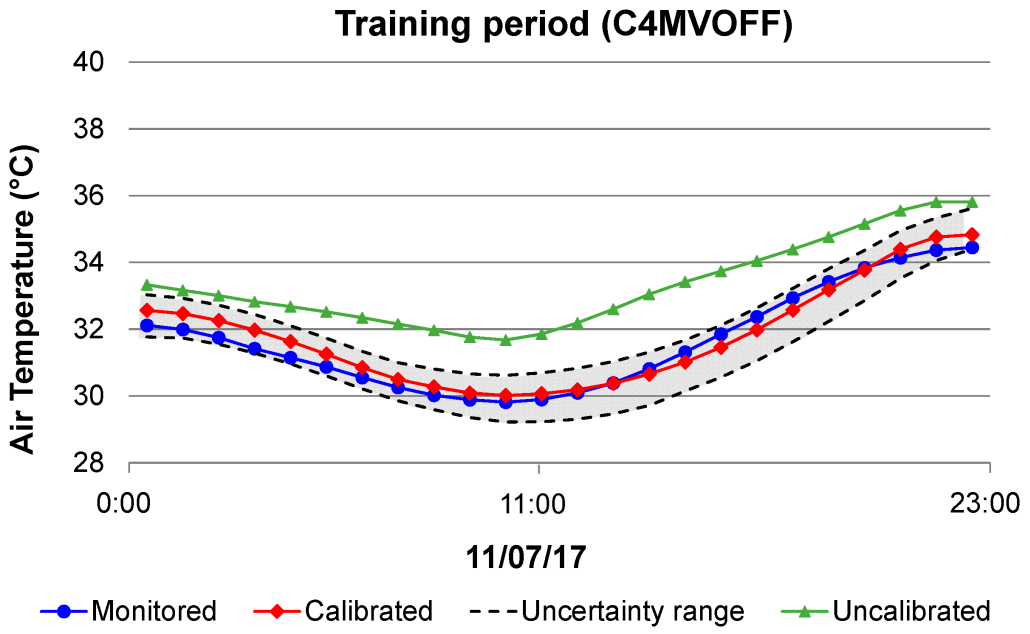


Figure B.1. Comparison between monitored and simulated data during the 24-hour training period of the C4MVOFF protocol. The uncertainty range related to the posterior distributions is shown in grey. The red line represents the best calibration results.

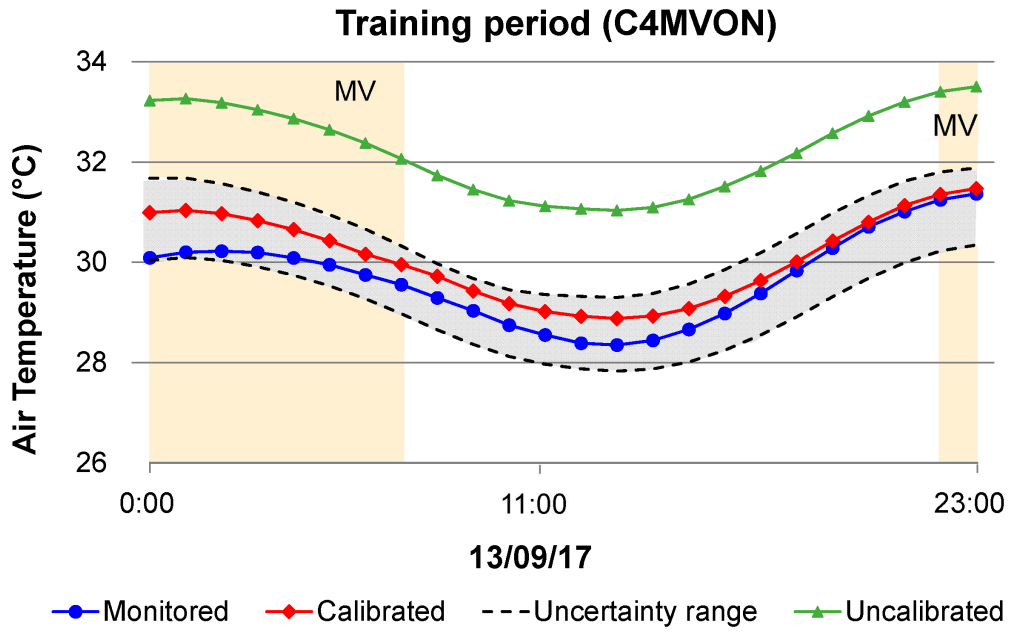


Figure B.2. Comparison between monitored and simulated data during the 24-hour training period of the C4MVON protocol. The uncertainty range related to the posterior distributions is shown in grey. The red line represents the best calibration results. MV means mechanical ventilation.

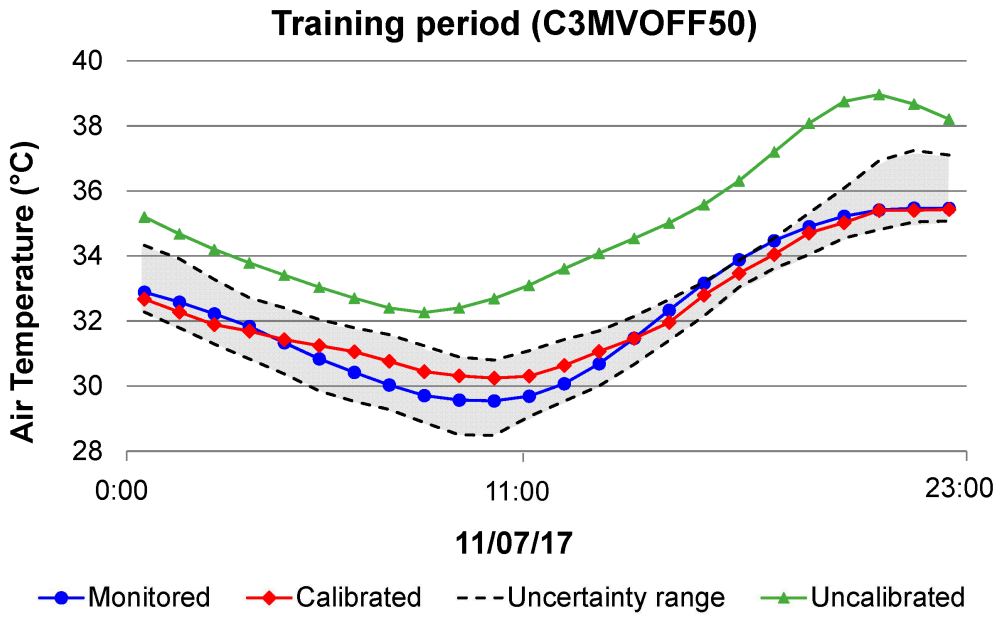


Figure B.3. Comparison between monitored and simulated data during the 24-hour training period of the C3MVOFF50 protocol. The uncertainty range related to the posterior distributions is shown in grey. The red line represents the best calibration results.

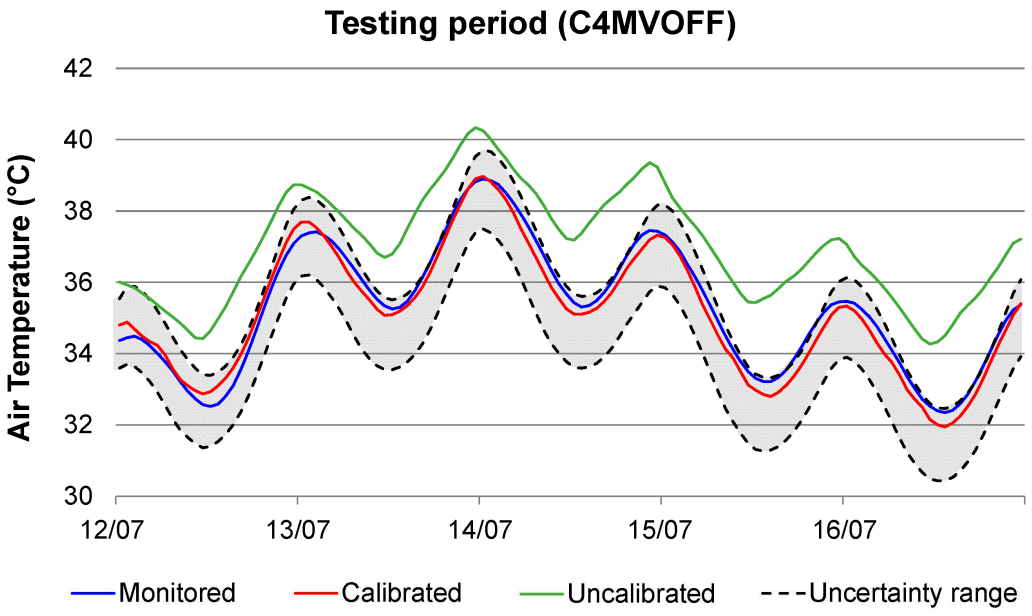


Figure B.4. Comparison between monitored and simulated data during the 120-hour testing period of the C4MVOFF protocol. The uncertainty range related to the posterior distributions is shown in grey. The red line represents the best calibration results.

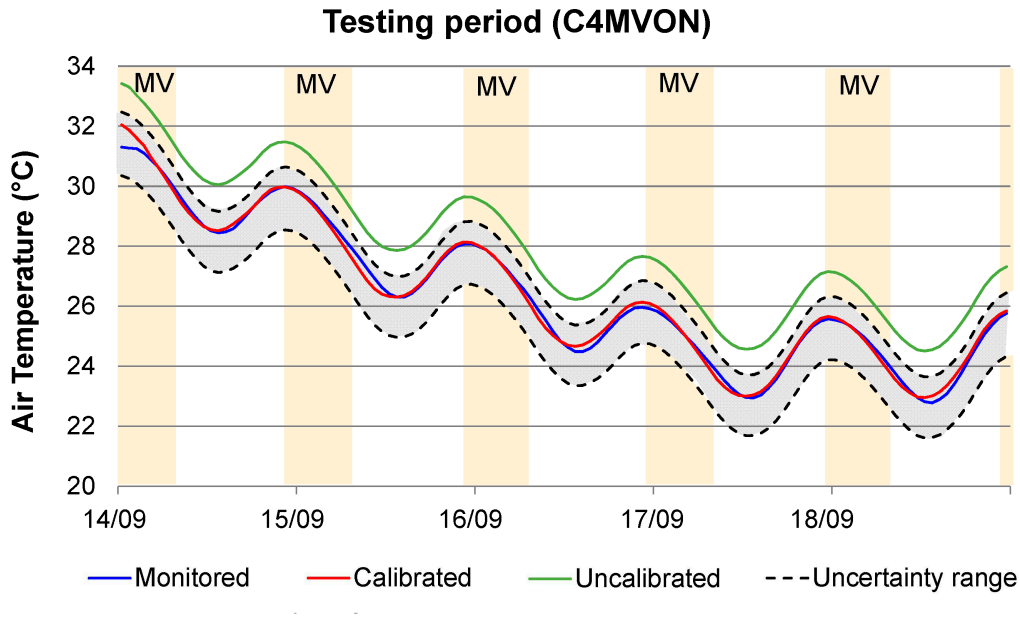


Figure B.5. Comparison between monitored and simulated data during the 120-hour testing period of the C4MVON protocol. The uncertainty range related to the posterior distributions is shown in grey. The red line represents the best calibration results. MV means mechanical ventilation.

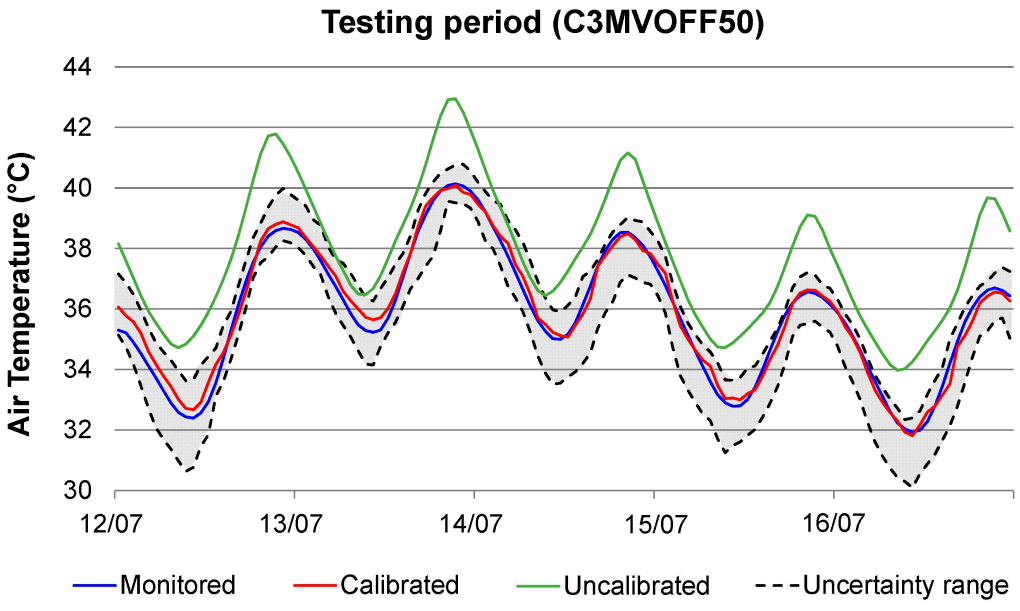


Figure B.6. Comparison between monitored and simulated data during the 120-hour testing period of the C3MVOFF50 protocol. The uncertainty range related to the posterior distributions is shown in grey. The red line represents the best calibration results.