Taylor & Francis
Taylor & Francis Group

Journal of Building Performance Simulation

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tbps20

Development of a Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature

Giorgos Petrou, Anna Mavrogianni, Phil Symonds, Zaid Chalabi, Kevin Lomas, Anastasia Mylona & Michael Davies

To cite this article: Giorgos Petrou, Anna Mavrogianni, Phil Symonds, Zaid Chalabi, Kevin Lomas, Anastasia Mylona & Michael Davies (04 Nov 2024): Development of a Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature, Journal of Building Performance Simulation, DOI: 10.1080/19401493.2024.2421330

To link to this article: https://doi.org/10.1080/19401493.2024.2421330

View supplementary material

Published online: 04 Nov 2024.

Submit your article to this journal

Article views: 17

View related articles

View Crossmark data

JOURNAL OF BUILDING PERFORMANCE SIMULATION
https://doi.org/10.1080/19401493.2024.2421330

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Development of a Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature

Giorgos Petrou [a], Anna Mavrogianni [a], Phil Symonds [a], Zaid Chalabi[a], Kevin Lomas [b], Anastasia Mylona[c] and Michael Davies [a]

[a]Institute for Environmental Design and Engineering, University College London, London, UK; [b]School of Architecture, Building and Civil Engineering, Loughborough University, Loughborough, UK; [c]Chartered Institution of Building Services Engineers, London, UK

ABSTRACT

Archetype-based housing stock models of summer indoor temperature can support the development of policies to manage the climate change-driven increase in cooling demand and heat-related health impacts. Calibration can reduce the performance gap of such models, however, work on this topic is limited. Motivated by the growing importance of this underexplored research area, this paper introduces a framework for the Bayesian calibration of archetype-based housing stock models of summer indoor temperature. The framework includes data-driven procedures to classify dwellings into homogeneous groups and specify prior probability distributions. To demonstrate its application, an established bottom-up model based on EnergyPlus was calibrated using data collected from 193 dwellings monitored during the 2009 4M survey in Leicester, England. Post-calibration, the root-mean-square error reduced from 2.5°C to 0.6°C and remaining uncertainties were quantified. The application of this modular framework may be extended to models of energy use and other indoor environmental parameters.

ARTICLE HISTORY
Received 18 March 2024
Accepted 18 October 2024

KEYWORDS
Bayesian calibration; archetype-based modelling; housing stock model; indoor temperature; uncertainty quantification; performance gap

1. Introduction

The scale of ongoing and future climate change, largely driven by the anthropogenic emission of greenhouse gases, is unprecedented; global surface temperature has increased faster since 1970 than in any other 50-year period in the last 2000 years (IPCC 2021). The past decade has seen the 10 hottest years since 1850, with 2023 being the warmest year on record with a mean temperature 1.35°C greater than the pre-industrial (1850–1900) average (NOAA 2024). The impact of heat on human health and wellbeing is substantial. For example, out of 85,000–145,000 fatalities in the European Economic Area linked to weather- and climate-related events between 1980 and 2022, 86–91% were due to heatwaves (European Environment Agency 2022). The effects of elevated temperatures extend to countries for which summer heat has not traditionally been a major public health concern (Taylor et al. 2023). One such country is the United Kingdom (UK), for which heat-related annual mortality is projected to grow in the absence of adaptation from a baseline of 2000 deaths in 2012 to 7,000–11,000 deaths in 2050 (Macintyre and Murage 2023).

The built environment plays an important role in adapting to rising temperatures. Data from seven European countries suggest that people typically spend over 90% of their time indoors (Schweizer et al. 2007), and evidence of the adverse effects of indoor heat on cognitive performance, productivity, sleep quality and mortality exist (Kovats and Brisley 2021). Indoor heat exposure can vary between buildings and occupants (Lomas et al. 2021), and its consequences will partly depend on occupant vulnerability (Macintyre and Murage 2023). Thus, effective climate change adaptation policies should take into consideration the diversity in building characteristics, occupant needs and desires.

An approach that can support effective policy development is building stock modelling (Oraiopoulos and Howard 2022). Building stock models can enable policymakers to assess how summer indoor temperatures would respond to different adaptation measures and under future climatic scenarios. Such models can be deployed at the local, regional and national level to inform a range of decision makers. A sub-class of building stock models are archetype-based housing stock models.

CONTACT Giorgos Petrou giorgos.petrou@ucl.ac.uk

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19401493.2024.2421330.

A *building archetype* may be defined as a notional building that represents a group of buildings with similar properties (Reinhart and Cerezo Davila 2016). A building archetype is specified by: (i) grouping the building stock according to a set of criteria (classifiers – *classification*); and (ii) defining each group's geometry, thermal properties, occupancy patterns and systems (*characterization*) (Reinhart and Cerezo Davila 2016).

A concern with building energy and indoor environmental performance modelling is the mismatch between the measured and predicted performance of a building, often referred to as the *performance gap* (de Wilde 2014). The performance gap arises in part due to our inability to accurately represent the real-world system as a mathematical model, and has been described as fundamentally an issue of verification, validation and calibration (de Wilde 2023). This inability results from the several sources of uncertainty that are integral to the modelling practice (Saltelli et al. 2008). To reduce model uncertainty, and the performance gap, one can use *calibration*, which typically refers to the process of learning the values of unknown model inputs using field observations of the model output (Kennedy and O'Hagan 2001). Several approaches to calibration exist that can be broadly categorized as *manual* or *automated* (Coakley, Raftery, and Keane 2014). While manual approaches predominantly rely on iterative pragmatic intervention by the modeller, automated approaches employ a computer-based mathematical approach to infer parameter values. Over the last decade, automated approaches have become the more popular choice in academic research, with several methods being proposed, including genetic algorithms and particle swarm optimization (Chong, Gu, and Jia 2021). One such method that has gained prominence within the built environment field, with several applications demonstrating its efficacy in reducing the performance gap, relies on Bayesian inference (Oraiopoulos and Howard 2022). Introduced by Kennedy and O'Hagan, Bayesian calibration aims to improve a computer model's predictive ability while simultaneously quantifying the uncertainty of: unknown model inputs (parametric uncertainty), the model's structure (model bias) and measurement error (Kennedy and O'Hagan 2001).

Building performance may refer to several characteristics, including thermal comfort, daylight levels or indoor air quality. Yet, the gap between actual and predicted energy performance has received far more attention, partly due to the greater availability of energy data (de Wilde 2014)**,** and partly due to the emphasis placed on lowering energy consumption in buildings (Jain et al. 2020). In a recent review on the Bayesian calibration of archetype-based housing stock models, all papers identified were concerned with the prediction of energy use (Petrou 2023). Further, the review revealed that: (i) there was limited discussion surrounding the choice of priors,[1] with uniform distributions often used despite their shortcomings in representing the best available evidence; (ii) classification, a source of uncertainty in archetype-based Bayesian calibration, was often implemented on an *ad hoc* basis and without a clear definition of homogeneity; (iii) in the limited examples where a data-driven approach to classification was used, it was unclear how this approach linked with the subsequent calibration process. An *ad hoc* approach will not necessarily result in a poor classifier selection. However, a data-driven approach that identifies classifiers based on their effect on the quantity being modelled may reveal useful insights, result in a more accurate classification process and, potentially, better-performing models (Sokol, Cerezo Davila, and Reinhart 2017).

The discrepancy between actual and modelled summer indoor temperature is a pertinent performance gap that has thus far received limited attention, despite its implications for thermal comfort, health, and cooling demand. Calibration efforts have more commonly focused on individual buildings (Baba et al. 2022) and test cells (Calama-González et al. 2021). A recent study that incorporated Bayesian calibration when modelling thermal comfort in the social housing stock of southern Spain indicates the growing interest in the accurately predicting summer indoor temperatures (Calama-González, Suárez, and León-Rodríguez 2022). As with most studies reviewed in Petrou (2023), the classification process was not informed by a data-driven process. Furthermore, the limited data available meant that the archetype model was calibrated against measurements from a single dwelling.

1.1. Aim and objectives

Recognizing the important role that modelling can play in adapting the housing stock, and with the ambition to improve existing practices, this paper aims to explore methods to quantify and reduce uncertainties of archetype-based housing stock models of summer indoor temperature. This is achieved through the following steps:

(1) Developing a Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature;
(2) Demonstrating the framework's application on the UK Housing Stock Model (UK-HSM);

(3) Quantifying the post-calibration improvement in UK-HSM predictions.

The key contributions of this work are:

- Developing a modular framework for the data-driven classification and Bayesian calibration of archetype-based housing models of summer indoor temperatures that relies on an explicit definition of homogeneity
- Employing an approach to identify prior probability distributions of model inputs depending on the data available
- Exploring the impact that outdoor temperatures have on the calibration of models of summer indoor temperature

This work is the outcome of a doctoral study, and more information can be found in Petrou (2023).

2. Framework

Contrary to most papers reviewed in Petrou (2023), the proposed framework relies on a clear and practical definition of homogeneity that can guide data-driven classification: a group of dwellings is considered

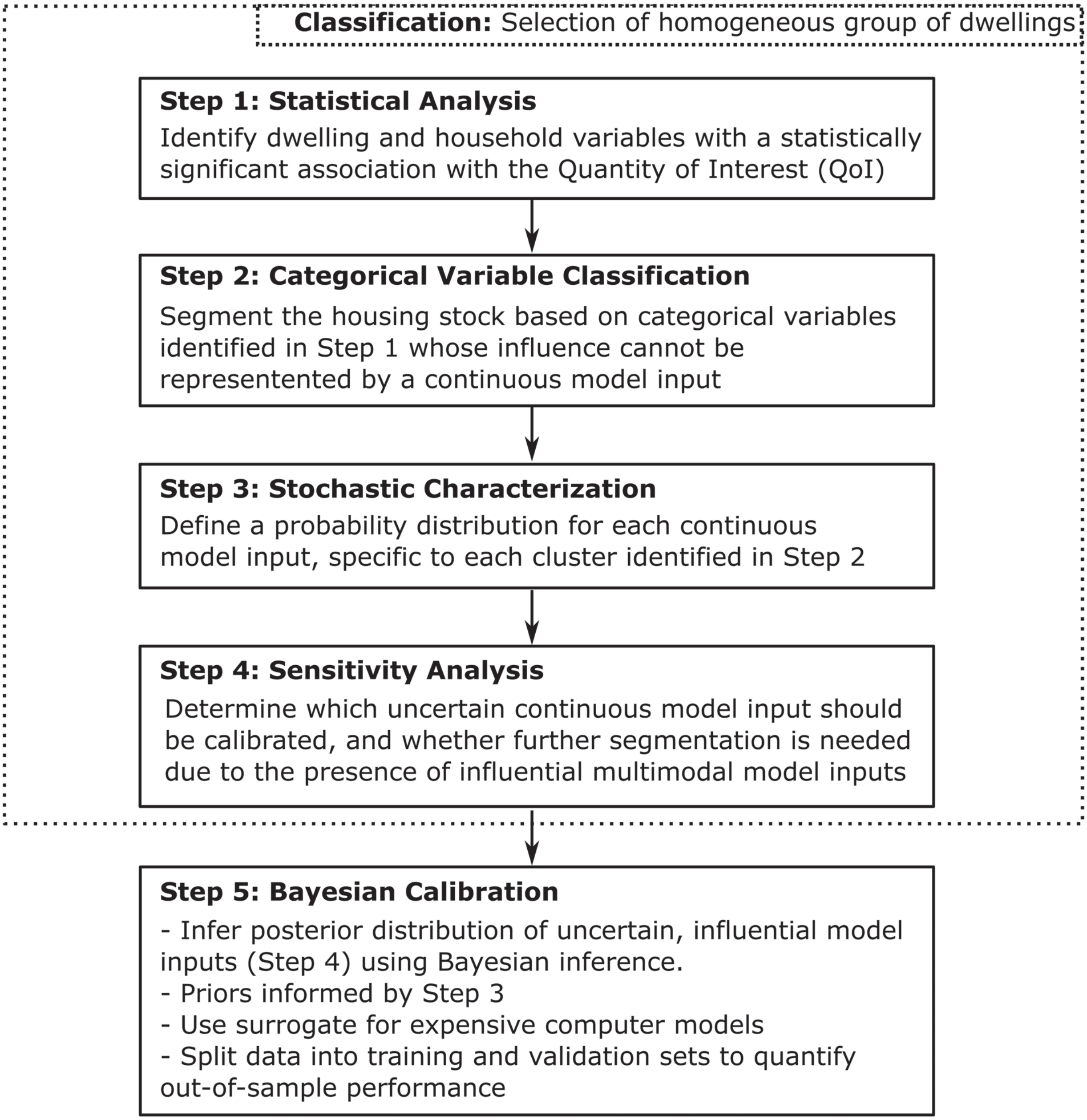

Classification: Selection of homogeneous group of dwellings
Step 1: Statistical Analysis
Identify dwelling and household variables with a statistically significant association with the Quantity of Interest (QoI)
Step 2: Categorical Variable Classification
Segment the housing stock based on categorical variables identified in Step 1 whose influence cannot be representented by a continuous model input
Step 3: Stochastic Characterization
Define a probability distribution for each continuous model input, specific to each cluster identified in Step 2
Step 4: Sensitivity Analysis
Determine which uncertain continuous model input should be calibrated, and whether further segmentation is needed due to the presence of influential multimodal model inputs
Step 5: Bayesian Calibration
- Infer posterior distribution of uncertain, influential model inputs (Step 4) using Bayesian inference.
- Priors informed by Step 3
- Use surrogate for expensive computer models
- Split data into training and validation sets to quantify out-of-sample performance

Figure 1. Workflow diagram for Bayesian calibration framework.

homogeneous *if the variability of influential building parameters could be described by unimodal distributions*.

Only the most influential parameters are considered to: (i) avoid the excessive segmentation of the housing stock based on parameters that would not largely influence the Quantity of Interest (QoI – desired model output), (ii) reduce computational cost (since a greater number of parameters would generally require a larger training dataset), and (iii) to reduce the risk of parameter non-identifiability[2] (Chong and Menberg 2018; Menberg, Heo, and Choudhary 2019).

The focus on unimodal distributions stems from practical and computational considerations. Multimodal distributions of building parameters are often the result of building characteristics that differ between modes but are shared amongst dwellings within modes (Petrou 2023). By identifying such characteristics, and grouping dwellings based on them, the analysis can be more informative to policymakers who are often interested in applying measures to groups of similar dwellings (Booth, Choudhary, and Spiegelhalter 2012). Further, the Markov Chain Monte Carlo (MCMC) algorithms often used in Bayesian model calibration, including Hamiltonian Monte Carlo, do not perform as well when faced with multimodal posteriors[3] (Yun et al. 2020).

Following from the definition of homogeneity, the Bayesian calibration framework is presented in Figure 1.

Step 1: Statistical analysis.

Bivariate and multivariate methods are used to analyse empirical observations and identify which variables (e.g. wall U-value) have a statistically significant association with the QoI. If the archetype-based model has yet to be developed, or further development is desired, Step 1 could inform this process.

Step 2: Categorical variable classification.

Considering the model structure, the housing stock is segmented based on statistically significant categorical variables (identified as such in Step 1) whose effect cannot be captured by a continuous model input. For example, assume *dwelling type* and *floor area bands* were both shown to be significantly associated with the QoI. If floor area can be specified as a continuous model input, it is not used as a classifier. On the other hand, the housing stock would need to be clustered based on the dwelling type, since this is not a continuous model input.

Step 3: Stochastic characterization.

For each cluster, a probability distribution is defined for each continuous model input (see Section 3.6 for the approach proposed). Where possible, the types of probability distribution functions (e.g. uniform, normal, lognormal) are best informed by empirical data. This process is not constrained to the dataset used in Steps 1 and 2.

Step 4: Sensitivity analysis.

This step has two main aims: (1) To determine which of the uncertain continuous model inputs are influential and should be calibrated, and (2) To determine whether further segmentation of the housing stock is needed due to influential model inputs being described by multimodal distributions. The first aim is common amongst Bayesian calibration studies in the built environment, with the Morris method most frequently being used (Hou, Hassan, and Wang 2021). The second aim is based on the proposed definition of homogeneity. Having identified the distribution that best describes each model input in Step 3, sensitivity analysis is used to rank model inputs based on their feasible range of values. Subsequently, if any influential variables are described by multimodal distributions, the housing stock is further segmented for each mode and Steps 3–4 are repeated for the newly formed clusters. Influential model inputs characterized by unimodal distributions can be calibrated. Fixed values may be used to describe non-influential model inputs.

Step 5: Bayesian calibration.

Bayes' theorem is used to infer the posterior distributions of calibration variables (and other parameters if specified) given empirical observations of the QoI and any prior knowledge about the distributional form of uncertain parameters. The implementation will depend on a few factors, such as:

- **Computational cost:** A surrogate model may replace a computationally expensive computer model (Higdon et al. 2004).
- **Posterior estimation method:** MCMC is the most frequently used approach (Hou, Hassan, and Wang 2021).
- **Data Aggregation and Likelihood:** Data aggregation from different dwellings, and the choice of likelihood,[4] are important considerations (Petrou 2023).

By sampling from the posterior distributions, the computer or surrogate model can be used for predictions under new settings that incorporate parameter and model inadequacy uncertainties. To quantify improvements in predictive performance post-calibration, part of the empirical observations should be reserved for validation.

3. Methods

Section 2 introduced the general form of the Bayesian calibration framework. In this section, the framework's application on a case study model is described. Sections 3.1 and 3.2 provide an overview of the model and key datasets used. The selection of QoI is discussed in Section 3.3, while the implementation of each step of the framework is detailed in Sections 3.4–3.8.

3.1. UK Housing Stock Model

UK-HSM is an established bottom-up energy and indoor environment model, that relies on EnergyPlus for dynamic thermal simulations. Its development and application has been discussed in-depth (Petrou 2023). Briefly, at the core of UK-HSM is a parametric tool written in Python with predefined material, construction, geometry and occupancy libraries, thought to be representative of the English housing stock (Oikonomou et al. 2018; Symonds et al. 2016). Following the specification of seventeen model inputs (Figure 2), an EnergyPlus Input Data File (.idf) is generated and simulated. For the present analysis, EnergyPlus version 8.8.0. was used, with six timesteps per hour.

3.2. Datasets

Numerous datasets were used when implementing the framework for the case study described in this paper. For brevity, this section will focus on the dataset used in Steps 2 and 5. The datasets used in Steps 1 and 3 have been described in previously published work (Petrou 2023; Petrou et al. 2019).

3.2.1. The 4M dataset

Steps 2 and 5 relied on data collected from face-to-face questionnaires and surveys as part of the 4M project (Lomas and Kane 2013). A stratified random sample of 575 homes, containing a mixture of dwelling types, located in Leicester, England, was used. In addition to data on the dwelling and occupant characteristics, hourly measurements of indoor air temperature were also collected from the living room and main bedroom. Further information on the 4M survey, including details on the temperature sensors used, is given in Lomas and Kane (2013). A subsample of 193 homes with adequate metadata were available to use in this study.

This paper focused on the period between 1st July and 31st August 2009 when indoor temperature measurements were available, and heating was assumed to be off for most homes. As noted by Lomas and Kane (2013), the summer of 2009 was relatively cool with average temperatures for July (16.2°C) and August (16.6°C) being 1.0°C and 0.5°C below the Leicester 10-year averages, respectively. The hottest period was between 28th of June and 2nd of July, with the average daily temperature exceeding 19°C and peaking on the 1st of July at 24.1°C.

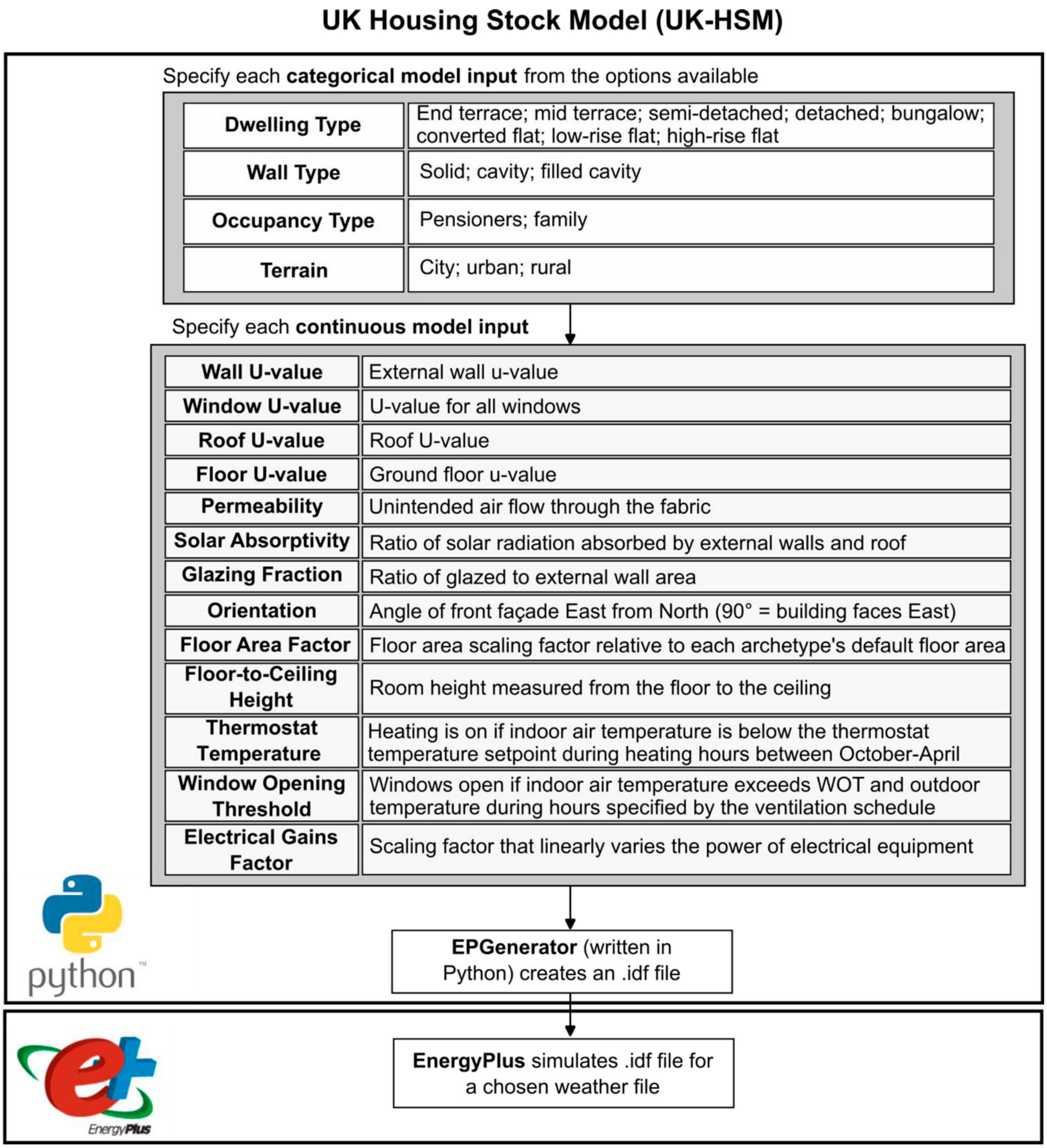
3.2.2. Weather data

Data from three weather stations (Figure 3), accessed through the Met Office Integrated Data Archive System (Met Office 2018), were used to construct a whole-year weather file for 2009. The stations were selected based on data availability and their proximity to the centre of Leicester. Hourly non-solar data were taken from the Cottesmore station. If a single hourly observation was missing, the mean of the hour before and after was used. There was one instance when data were missing for a continuous time period (13 consecutive hours). In that case, the hourly mean of same time period for the days before and after were used. Hourly solar data were based on recordings from the Sutton Bonington station. To replace missing solar data for 240 hours in December 2009, data recorded at the Church Lawford station were used. To estimate the solar components needed for the simulation, an in-house tool developed for the work described by Symonds et al. (2017) was used.

3.3. Quantity of interest

The chosen QoI was the archetype *Mean of the Daytime Living Room Temperature (MDLRT)*: The mean of the daytime (08:00–22:00) hourly living room temperatures, estimated daily and averaged across dwellings belonging to the same archetype.

The choice of QoI was based on previous and planned use of UK-HSM in modelling the effects of home energy efficiency and heat adaptation measures on heat-related mortality, with the purpose to inform policymakers on the potential costs and benefits of such measures (Taylor et al. 2018, 2021). Due to the lack of evidence on the relationship between indoor temperature and heat-mortality, previous work assumed the effect of daytime maximum living room temperature on heat-mortality to be proportional to that established for daily ambient maximum temperatures (Taylor et al. 2018). Since epidemiological relationships between heat-mortality and daily mean ambient temperatures also exist (Hajat et al. 2014), and measurements of daytime maximum living room temperature can be biased by short-term exposure to heat (e.g. direct sunlight), MDLRT was preferred.



UK Housing Stock Model (UK-HSM)

Specify each **categorical model input** from the options available

Dwelling Type	End terrace; mid terrace; semi-detached; detached; bungalow; converted flat; low-rise flat; high-rise flat
Wall Type	Solid; cavity; filled cavity
Occupancy Type	Pensioners; family
Terrain	City; urban; rural

Specify each **continuous model input**

Wall U-value	External wall u-value
Window U-value	U-value for all windows
Roof U-value	Roof U-value
Floor U-value	Ground floor u-value
Permeability	Unintended air flow through the fabric
Solar Absorptivity	Ratio of solar radiation absorbed by external walls and roof
Glazing Fraction	Ratio of glazed to external wall area
Orientation	Angle of front façade East from North (90° = building faces East)
Floor Area Factor	Floor area scaling factor relative to each archetype's default floor area
Floor-to-Ceiling Height	Room height measured from the floor to the ceiling
Thermostat Temperature	Heating is on if indoor air temperature is below the thermostat temperature setpoint during heating hours between October-April
Window Opening Threshold	Windows open if indoor air temperature exceeds WOT and outdoor temperature during hours specified by the ventilation schedule
Electrical Gains Factor	Scaling factor that linearly varies the power of electrical equipment

EPGenerator (written in Python) creates an .idf file

EnergyPlus simulates .idf file for a chosen weather file

Figure 2. UK housing stock model flowchart.

3.4. Step 1: statistical analysis

Step 1 for this work has been previously implemented and published (Petrou et al. 2019). The linked 2011 English Housing Survey and Energy Follow-up Survey were used to study the association of dwelling and household characteristics with summertime Standardized Indoor Temperature (SIT) in the English housing stock. A linear regression model was fitted for each home using observations of indoor living room temperature, and the daily mean of outdoor temperature ($\boldsymbol{T}_{out,mean}$) and global horizontal irradiance ($\boldsymbol{GHI}_{mean}$) (Petrou et al. 2019):

$$\boldsymbol{T}_{LR} = \beta_0 + \beta_1 \boldsymbol{T}_{out,mean} + \beta_2 \boldsymbol{GHI}_{mean} \qquad (1)$$

where $\boldsymbol{T}_{LR}$ is the daily daytime (08:00–22:00) mean indoor temperature estimated from hourly measurements of living room temperature and β_{0-2} are the fitted linear regression coefficients for each home. These models were subsequently used to predict each home's indoor temperature for the same outdoor conditions (i.e. the SIT), allowing for comparisons between dwellings in different

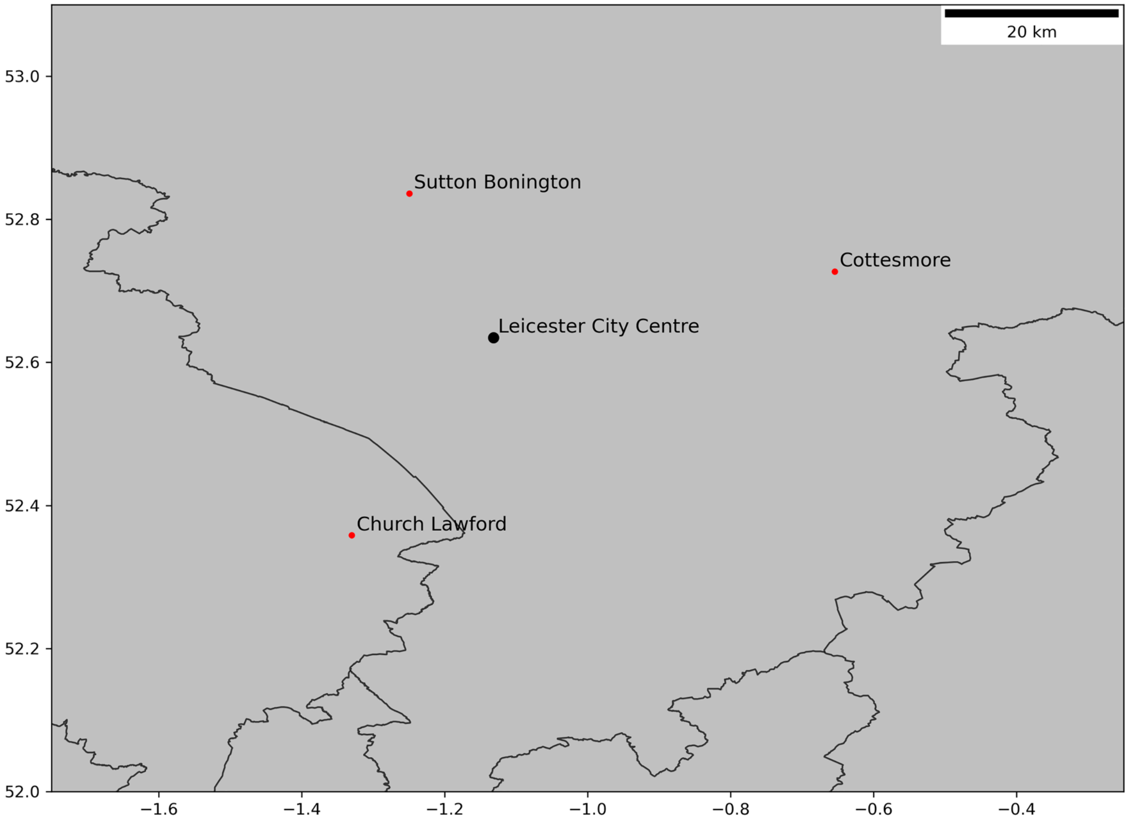


Figure 3. The location of the weather stations used to construct the 2009 weather file.

locations. Variables with a statistically significant association with SIT were identified using Kruskal–Wallis and Pairwise Mann–Whitney U-test for multiple comparisons with the False Discovery Rate (FDR) p-adjustment method.

3.5. Step 2: categorical variable classification

In accordance with the definition of homogeneity provided in Section 2, and by considering the UK-HSM model structure, a categorical variable shown to be statistically significant in Step 1 was:

- Used as a classifier if modelled explicitly (e.g. dwelling type).
- Not used as a classifier if it could be represented by a continuous UK-HSM model input (e.g. loft U-value representing loft insulation thickness).

Where empirical data are abundant, it is advisable to use all classifiers in the segmentation process. In the case study described in this paper, due to the relatively small number of homes monitored in the 4M project, not all classifiers were used to avoid excessive segmentation (see Section 4.2).

3.6. Step 3: stochastic characterization

The method used to identify each model input's probability distribution depended on the data available (Figure 4). For an empirical dataset whose tabulated values are available, a distribution-fitting method was implemented (Petrou et al. 2021). If empirical data were available only in a graphical form, information was extracted by overlaying the visualization onto a set of axes within the R package ggplot2 (Wickham 2016). Where this was applied to a barplot or histogram, a distribution was fitted to the extracted data. If empirical data were not available, the probability distributions were assumed based on judgement, experience and the information known about the uncertain variable (Mun 2012). These methods are described in detail within Section 3.1 of the Supplementary Material.

Where tabulated empirical data were available, distributions were derived from dwellings whose characteristics matched the classifiers identified in Step 2, capturing potential associations between the classifiers and model inputs. The quality of the fitted distributions was assessed using the Akaike Information Criterion (AIC) and Goodness-of-Fit plots, following the methodology in Petrou et al. (2021).

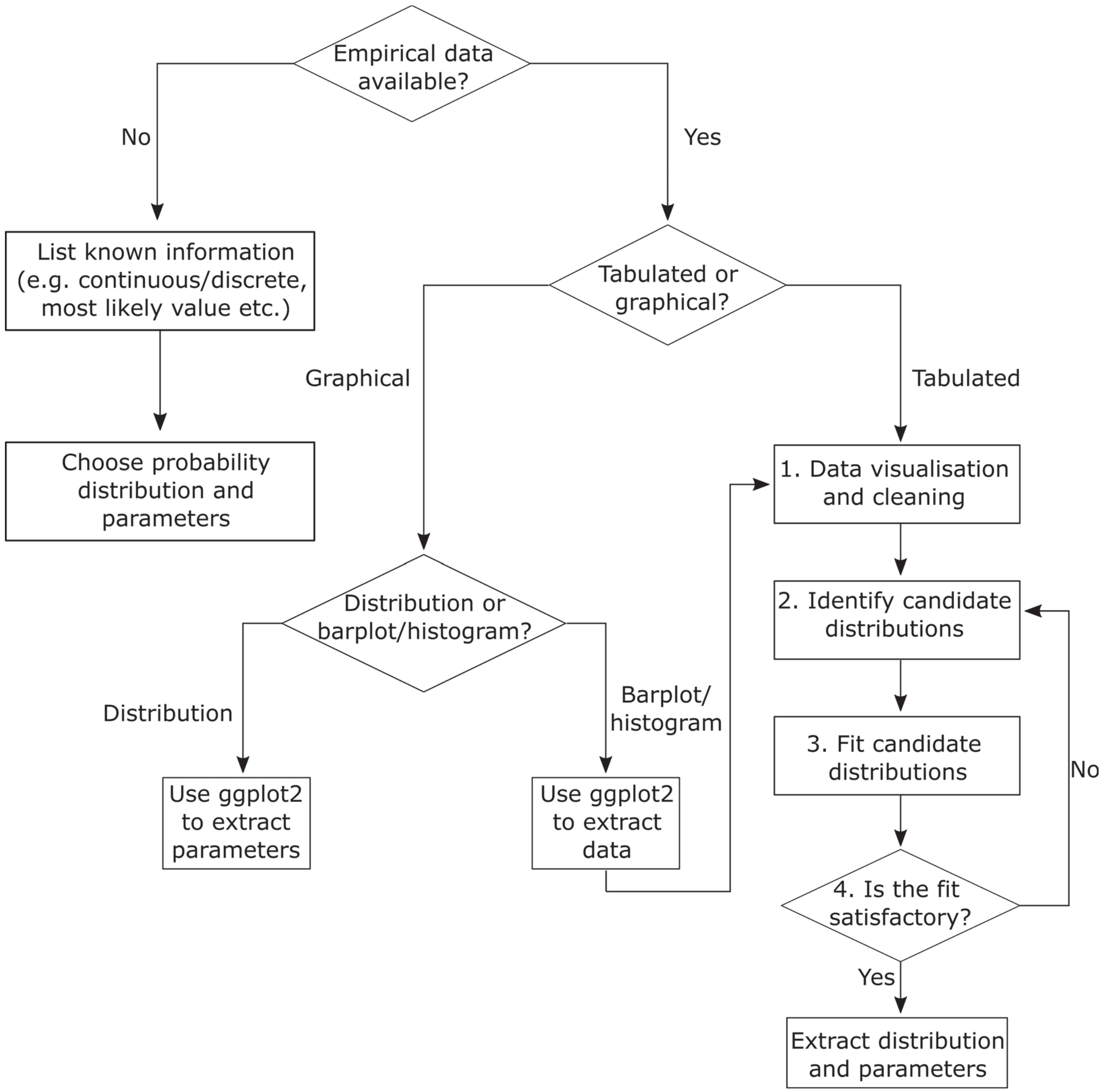

Empirical data available?
No
Yes
List known information (e.g. continuous/discrete, most likely value etc.)
Choose probability distribution and parameters
Tabulated or graphical?
Graphical
Tabulated
1. Data visualisation and cleaning
2. Identify candidate distributions
3. Fit candidate distributions
4. Is the fit satisfactory?
No
Yes
Extract distribution and parameters
Distribution or barplot/histogram?
Distribution
Barplot/ histogram
Use ggplot2 to extract parameters
Use ggplot2 to extract data

Figure 4. Workflow diagram for Step 3 (stochastic characterization).

3.7. Step 4: sensitivity analysis

The Morris method was used to carry out a two-stage sensitivity analysis using the Python package SALib (Herman and Usher 2017; Iwanaga, Usher, and Herman 2022). Informed by the work of Petersen, Kristensen, and Knudsen (2019), the number of levels was set at 12, while the trajectory number was set to 500 and simulations were run in batches until convergence was achieved after 1300 simulations. The summer-averaged (July–August 2009) MDLRT was used.

In the first stage, all continuous model inputs were sampled, with their bounds informed by Step 3 (Table 1). Categorical model inputs were specified as the most frequent value in the homogeneous cluster (Table 1). Through this process, variables to be calibrated or kept fixed were identified.

In the second stage, whether the Floor Area Factor (FAF), the only continuous model input available from the 4M survey, should be used as an explanatory variable was investigated. In this stage, FAF was allowed to vary within 0.9–1.1 (compared to 0.55–1.94 in the first stage). The Stage 2 interval represents the idea that the FAF of a typical house (FAF = 1) has a measurement error of 10%. By considering its influence on MDLRT in Stages 1 and 2, a decision on whether FAF should be kept fixed, calibrated or used as an explanatory variable was taken.

Table 1. Lower and upper bounds of model inputs sampled to train the surrogate model, and values of model inputs kept fixed.

Model input [unit]	Sensitivity analysis (Stage 1)	Type	Surrogate model training values
Wall U-value [W/(m^{2}K)]	0.26–1.48	Fixed	0.71
Window U-value [W/(m^{2}K)]	1.70–3.27	Fixed	2.5
Roof U-value [W/(m^{2}K)]	0.10–2.53	Fixed	0.51
Floor U-value [W/(m^{2}K)]	0.22–0.90	Fixed	0.70
Permeability [m^3/h/m^2 @ 50 Pa]	1.7–24.7	Fixed	11.29
Solar Absorptivity	0.16–0.96	Fixed	0.63
Glazing Fraction	0.12–0.48	Calib.	0.12–0.48
Orientation [°]	0.0–330.0	Calib.	0.0–330.0
Floor Area Factor	0.55–1.94	Explan	0.55–1.94
Floor-to-Ceiling Height [m]	2.32–2.77	Fixed	2.53
Window Opening Threshold [°C]	18.0–32.0	Calib.	18.0–32.0
Electrical Gains Factor	0.08–1.21	Calib.	0.08–1.21
Dwelling Type	–	Fixed	Semi-detached
Wall Type	–	Fixed	Filled cavity
Occupancy Type	–	Fixed	Pensioners
Terrain	–	Fixed	Urban

3.8. Step 5: Bayesian calibration

The calibration approach employed in this study was informed by the work of Kennedy and O'Hagan (2001); Booth, Choudhary, and Spiegelhalter (2012); and Kristensen et al. (2017)

A 'complete pooling' approach was selected, which assumes that all observations of daily indoor temperature come from a single distribution, the archetype distribution. Thus, all dwellings have an equal contribution to the estimation of the calibration parameters and model hyperparameters. The use of this method is supported by the idea that a homogeneous cluster has been identified, and influential calibration variables are modelled explicitly. Contrary to Booth, Choudhary, and Spiegelhalter (2012), the monitored data within the homogeneous cluster were not averaged across dwellings prior to the calibration, in order to quantify the level of unexplained variance that remained following the calibration. In addition, this implementation does not require the choice of an arbitrary cut-off point for calibration parameter values to discard, as per Cerezo et al. (2017). It includes a model discrepancy term which could potentially reveal shortcomings of UK-HSM. It also allows for the straightforward specification of non-normal and non-uniform priors for the calibration parameters, which will be shown to describe the calibration parameters best in Section 4.3.

3.8.1. Data generation and transformation

The monitored data consisted of hourly measurements of indoor temperature collected in dwellings belonging to the homogeneous cluster (Section 4.2). The simulation data were generated by sampling 50 times from uniform distributions assigned to the calibration and explanatory variables (Table 1). The commonly used Latin Hypercube Sampling (LHS) procedure was employed (Tian et al. 2018), following the suggestion of having at least ten samples per variable (Chong and Menberg 2018). Uniformly sampling, with the same lower and upper bounds as in Step 4, ensured that the surrogate model represented the computer model well across the entire range of input values. As per Higdon et al. (2004), the calibration and explanatory variables were standardized to be within the range [0, 1], while the observations (monitored and simulated MDLRT) were transformed to have a mean of 0 and variance of 1.

3.8.2. Statistical framework

Due to the large computational cost of UK-HSM simulations, a Gaussian process (GP) was trained as a surrogate model on simulated ($\mathbf{y}_c^{(S)}$) and monitored ($\mathbf{y}_c^{(M)}$) data (Higdon et al. 2004). Each monitored or simulated home is associated with D values of MDLRT. A subset of these days was used for the calibration ($D_c = 10$ days), while the remaining was used for validation ($D_v = 52$ days). With M monitored dwellings, the total number of monitored data points used for the calibration is $N_c^{(M)} = M \times D_c$. Similarly, with S computer simulations the total number of simulated data points used for calibration were $N_c^{(S)} = S \times D_c$. What differentiates each day in this statistical formulation is a set of weather variables. Day 1 ($d = 1$) is associated with weather variable values $\mathbf{w}_1$, day 2 ($d = 2$) is associated with weather variable values $\mathbf{w}_2$ and so on. What differentiates dwellings on the same day is the set of explanatory variables; in this paper, the only explanatory variable was FAF. Thus, monitored dwelling $m = 1$ is associated with explanatory variable value $x_{m=1}^{(M)}$, while simulated dwelling $s = 1$ is associated with explanatory variable value $x_{s=1}^{(S)}$. Note that $x_{m=1}^{(M)}$ and $x_{s=1}^{(S)}$ are not equivalent; $x_{m=1}^{(M)}$ came from measurements associated with the monitored dwelling $m = 1$ while $x_{s=1}^{(S)}$ was sampled probabilistically.

For the $N_c^{(M)}$ monitored data points used for the model calibration, the following statistical relationship was established:

$$y_{md}^{(M)} = y(\mathbf{x}_m^{(M)}, \mathbf{w}_d) = \eta(\mathbf{x}_m^{(M)}, \mathbf{w}_d, \boldsymbol{\theta}) + \delta(\mathbf{x}_m^{(M)}, \mathbf{w}_d) + \epsilon_{md}^{(M)} \tag{2}$$

where $y_{md}^{(M)}$ is the MDLRT for monitored home m on day d; $\eta(\cdot)$ is the surrogate model represented by a GP; $\delta(\cdot)$ is the discrepancy term (or model bias) represented by a GP; $\mathbf{w}_d$ are the weather-related variables corresponding to day d; $\mathbf{x}_m^{(M)}$ are all other explanatory variables associated with monitored dwelling m; $\boldsymbol{\theta}$ are the calibration parameters, and $\epsilon_{md}^{(M)}$ is the associated error term.

The error term, $\epsilon_{md}^{(M)}$, allows for different observations, $y(\mathbf{x}_m^{(M)}, \mathbf{w}_d)$, to exist for the same conditions and captures the measurement error and any residual variation (Higdon et al. 2004); this might include stochastic occupant behaviour and violations of the cluster homogeneity assumption (Kristensen et al. 2017). This is assumed to be normally distributed, with a mean of zero and a variance of $1/\lambda_\epsilon$ (Chong and Menberg 2018). For the $N_c^{(S)}$ model data points, the following statistical relationship was defined:

$$y_{sd}^{(S)} = y(\mathbf{x}_s^{(S)}, \mathbf{w}_d, \mathbf{t}_s) = \eta(\mathbf{x}_s^{(S)}, \mathbf{w}_d, \mathbf{t}_s) + \epsilon_{sd}^{(S)} \tag{3}$$

where $y_{sd}^{(S)}$ is the MDLRT for simulated home s on day d; $\mathbf{x}_s^{(S)}$ are explanatory variables associated with simulated dwelling s; $\mathbf{t}_s$ are sampled values of the calibration parameters and $\epsilon_{sd}^{(S)}$ is a simulation error (or noise) term.

The simulation error term $\epsilon_{sd}^{(S)}$ has been added for three reasons: (i) It ensures the numerical stability of the covariance function (Higdon et al. 2004), (ii) it allows for different values of $y(\mathbf{x}_s^{(S)}, \mathbf{w}_d, \mathbf{t}_s)$ for the same combination of $[\mathbf{x}_s^{(S)}, \mathbf{w}_d, \mathbf{t}_s]$, which in theory could occur due to the aggregation process, (iii) it allows the same set of UK-HSM simulations to be used in the parametric analysis discussed in Section 3.8.3, reducing the computational cost. The noise term is also assumed to be normally distributed, with a mean of zero and a variance of $1/\lambda_{sim}$. For both relationships defined in Equations (2–3), the measurement error of $\mathbf{x}_m^{(M)}$ and $\mathbf{w}_d$ was assumed to be negligible and was ignored.

As per Higdon et al. (2004), a single combined vector of monitored and simulation data (of length $N_c^{(M)} + N_c^{(S)}$) was constructed $\mathbf{z} = [\mathbf{y}_c^{(M)}, \mathbf{y}_c^{(S)}]$.[5] By making the commonly used assumption that the error terms are *independently and identically distributed* (*iid*) – they come from the same distribution and are mutually independent (Smith 2013) – the resulting likelihood function was defined as (Higdon et al. 2004):

$$L(z|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\xi}) \propto |\mathbf{K}_z|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \mathbf{K}_z^{-1}(\mathbf{z} - \boldsymbol{\mu})\right\} \tag{4}$$

where $\mathbf{K}_z$ is the combined covariance matrix, $\boldsymbol{\mu}$ is the mean function defined as a vector of zeros, and $\boldsymbol{\xi}$ represents the hyperparameters of the surrogate model, model bias and error terms (please refer to Section 1.5.3 of the Supplementary Material for further information). $|\mathbf{K}_z|$ and $\mathbf{K}_z^{-1}$ represent the determinant and inverse of the combined covariance matrix, respectively.

The calibration parameter priors were based on the distributions identified in Step 3 of the Bayesian calibration framework. However, since the calibration parameters were standardized to be in the interval [0, 1], the prior distributions had to be reparametrized to be on the same scale. Further information on the re-parametrization process, and the complete set of priors is provided in Sections 1.5.4–1.5.5 of the Supplementary Material.

3.8.3. The choice of variables: parametric calibration

The MDLRT observed on day d ($y_{m,d}^{(M)}$) for a free-running building[6] is more likely to be similar to the value of the previous day $y_{m,d-1}^{(M)}$, than that of the previous week $y_{m,d-7}^{(M)}$. This is supported by the AutoCorrelation Function plot (Figure 5), where the cluster's mean MDLRT is autocorrelated for up to four days. For the purposes of statistical modelling and calibration, while $y_{md}^{(M)}$ (and $\epsilon_{md}^{(M)}$) are not independent variables, they can be conditionally independent given the right selection of predictors $\mathbf{x}_m^{(M)}$ and $\mathbf{w}_d$. In turn, this satisfies the assumption of the error terms to be *iid*. The explanatory parameters ($\mathbf{x}_m^{(M)}$) do not influence the autocorrelation observed in Figure 5 since they do not vary between days. However, the choice of weather variables is expected to have an effect. To evaluate this effect, in addition to the use of daily mean outdoor temperature and global horizontal irradiance, the use of lag components[7] of the daily mean outdoor temperature was explored, with the rationale that the indoor temperature will be affected from the ambient conditions of the previous days as a result of the building and surrounding environment's thermal mass. The focus on lag components of outdoor temperature was informed by an exploratory analysis that preceded the parametric analysis. For brevity, this analysis is included in the Supplementary Material.

While the calibration variables were selected using the Morris method, it was not possible to know whether parameter identifiability issues would arise prior to the calibration (Chong and Menberg 2018), nor what their effect would be on the model's predictive performance. To determine this and the effect of using one or two lag components of outdoor temperature, a parametric calibration analysis was conducted (Table 2). Due to its dominance during the sensitivity analysis, the Window Opening Threshold (WOT) was included in all calibration runs. The calibrations were run for 500 MCMC iterations using the No-U-Turn Sampler (NUTS) MCMC algorithm, shown to perform better than other commonly used MCMC algorithms (Chong et al. 2017).

3.8.4. Training and validation

Due to the strong association between MDLRT and Outdoor Temperature (OT) (see Section 2.2.1 of the Supplementary Material), 10 days of observations (16.1%) that provided good coverage of the OT variation over the monitored period (62-day) were selected for training.

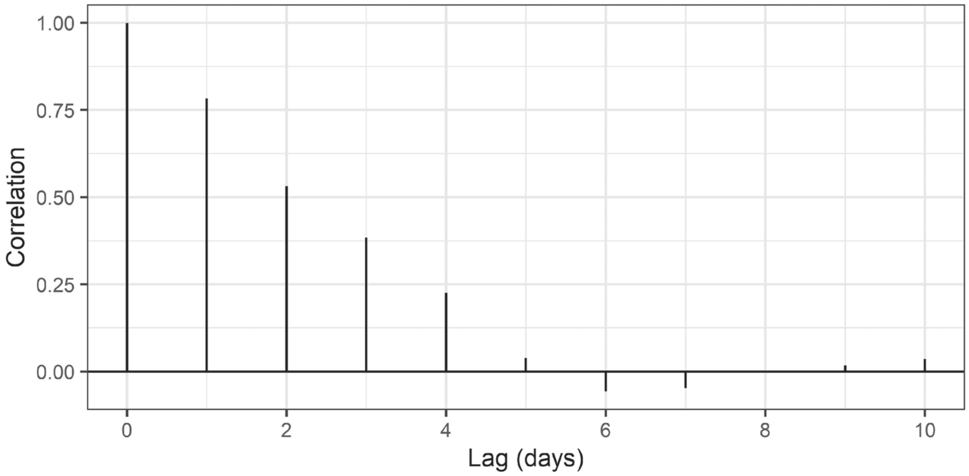


Figure 5. AutoCorrelation function plot of the mean of the mean daytime living room temperature for the homogeneous cluster of dwellings.

Table 2. Summary of variable combinations assessed during the parametric calibration.

Experiment	Weather	Explanatory	Calibration
EXP1	OT, GHI	FAF	WOT, Orientation, GF, EGF
EXP2	OT, GHI	FAF	WOT, Orientation, GF
EXP3	OT, GHI	FAF	WOT, Orientation, EGF
EXP4	OT, GHI	FAF	WOT, GF, EGF
EXP5	OT, GHI	FAF	WOT, GF
EXP6	OT, GHI	FAF	WOT, EGF
EXP7	OT, GHI	FAF	WOT, Orientation
EXP8	OT, GHI	FAF	WOT
EXP1L1	OT, GHI, OTL1	FAF	WOT, Orientation, GF, EGF
EXP2L1	OT, GHI, OTL1	FAF	WOT, Orientation, GF
EXP3L1	OT, GHI, OTL1	FAF	WOT, Orientation, EGF
EXP4L1	OT, GHI, OTL1	FAF	WOT, GF, EGF
EXP5L1	OT, GHI, OTL1	FAF	WOT, GF
EXP6L1	OT, GHI, OTL1	FAF	WOT, EGF
EXP7L1	OT, GHI, OTL1	FAF	WOT, Orientation
EXP8L1	OT, GHI, OTL1	FAF	WOT
EXP1L2	OT, GHI, OTL1, OTL2	FAF	WOT, Orientation, GF, EGF
EXP2L2	OT, GHI, OTL1, OTL2	FAF	WOT, Orientation, GF
EXP3L2	OT, GHI, OTL1, OTL2	FAF	WOT, Orientation, EGF
EXP4L2	OT, GHI, OTL1, OTL2	FAF	WOT, GF, EGF
EXP5L2	OT, GHI, OTL1, OTL2	FAF	WOT, GF
EXP6L2	OT, GHI, OTL1, OTL2	FAF	WOT, EGF
EXP7L2	OT, GHI, OTL1, OTL2	FAF	WOT, Orientation
EXP8L2	OT, GHI, OTL1, OTL2	FAF	WOT

Note: OT = Outdoor Temperature; GHI = Global Horizontal Irradiance, OTL1/2 = Outdoor Temperature with Lag of 1/2 *d*(s); GF = Glazing Fraction; FAF = Floor Area Factor; WOT = Window Opening Threshold; EGF = Electrical Gains Factor.

These days were chosen because their OT values corresponded to regular intervals across the range of OT values observed. The remaining 52 days of observations (83.9%) were used for validation. As the computational cost of GP-based Bayesian calibration scales rapidly with data ($\sim \mathcal{O}(N^3)$) (Chong et al. 2017), and initial calibration runs demonstrated substantial improvement in performance with 10 training days, the chosen length of training was thought to offer an appropriate balance between computational cost and calibration performance.

To capture the uncertainty in the calibrated parameters and hyperparameters, 500 posterior samples were used to predict the MDLRT of each day. To quantify the calibrated model's predictive performance, the mean predicted MDLRT was estimated for each day, resulting in a vector of predictions $\overline{\mathbf{y}_v^{(P)}} = [\overline{y_1^{(P)}}, \overline{y_2^{(P)}}, \cdots, \overline{y_{D_v}^{(P)}}]$. The averaged predictions were compared against the daily mean values of the monitored data during the same unseen period ($\overline{\mathbf{y}_v^{(M)}} = [\overline{y_1^{(M)}}, \overline{y_2^{(M)}}, \cdots, \overline{y_{D_v}^{(M)}}]$) using a set commonly used set of validation metrics: Root Mean Square Error (RMSE), Mean Bias Error (NMBE) and Coefficient of Determination (R^2). The monitored data were also compared against the MDLRT predicted by UK-HSM, providing a baseline for the improvement in predictive performance following the calibration. The computer simulations and calibrations were run on UCL's High Performance Computing facilities utilizing Intel(R) Xeon(R) Gold 6140 CPU (2.30 GHz) and Intel(R) Xeon(R) Gold 6240 CPU (2.60 GHz) processors.

4. Results

4.1. *Step 1: statistical analysis*

The first step relies on the analysis carried out by Petrou et al. (2019). Statistically significant differences were found in the living room SIT for the following variables: dwelling type, dwelling age, floor area, storey, construction, presence of double glazing, nature of area, main heating system and SAP 09,[8] age band of oldest person, extended tenure of household, occupant with illness or disability, household income (5 bands) and occupant on means tested or certain disability-related benefits. From these variables, floor area is modelled as continuous model input. The choice of classifiers based on the

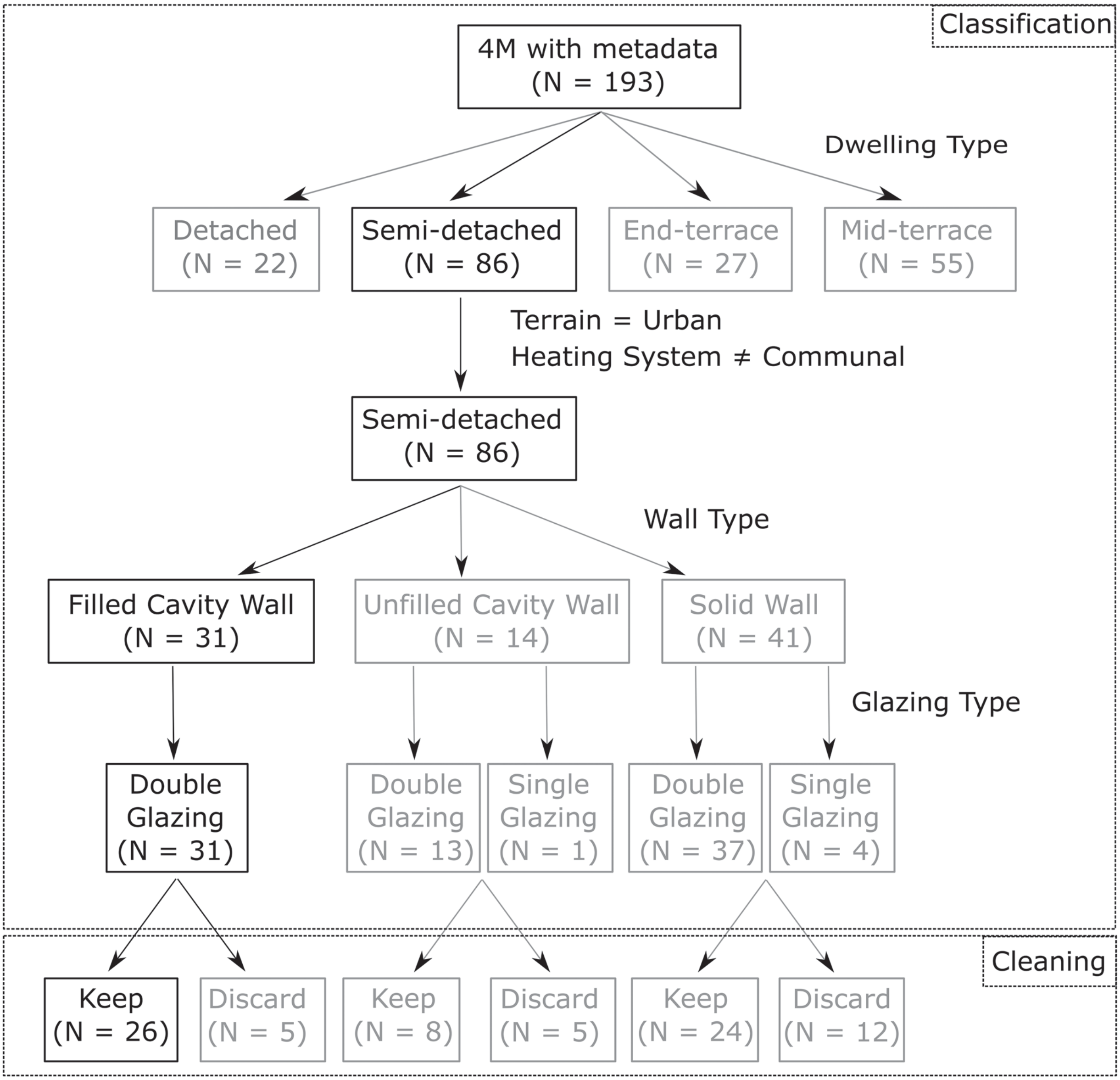


Figure 6. Flowchart of the first stage of the classification process and the subsequent cleaning.

remaining variables, and the UK-HSM model structure, is discussed in the next section.

4.2. Step 2: categorical variable classification

The 4M dataset was first segmented according to the *dwelling type*, a categorical model input in UK-HSM (Figure 6). Due to its larger sample size compared to others, the rest of the classification focused on the group of semi-detached dwellings, which is the most frequently occurring dwelling type in the UK (DLUHC 2021). Since all dwellings were located in Leicester, the same *terrain* (Urban) was assumed.

The *wall type* and *glazing type* were both associated with the living room SIT and were thus used as classifiers. Of the dwelling characteristics identified as potentially important in Step 1, *number of storeys*, *dwelling age* and *SAP 09* were not used as classifiers. This was a pragmatic decision since further classification could result in groups of dwellings that were too small, and where extreme values (outliers) within these groups could significantly influence the calibration process. The number of storeys was thought to have a small effect since the calibration focused on living rooms which are most commonly located on the ground floor. The effects of the dwellings' age and SAP rating are expected to be at least partly captured through the wall and glazing type.

Classification based on household variables was not performed. The only household variable that may be partly captured in UK-HSM is the *household composition*

Table 3. Model input distributions identified for the chosen cluster of dwellings.

Parameter	Distribution	Resources
Wall U-value	gamma(shape = 9.5, rate = 13)	[1]
Window U-value	norm(mean = 2.5, sd = 0.3)	[2]
Roof U-value	Multimodal	[2, 3]
Floor U-value	Multimodal	[2, 4]
Permeability	weibull(shape = 2.6, scale = 13)	[3, 5, 6, 7]
Solar Absorptivity	beta(shape1 = 4, shape2 = 2.5)	[8]
Glazing Fraction	gamma(shape = 14, rate = 53)	[4]
Orientation	unif(min = 0, max = 330)	[4]
Floor Area Factor	invweibull(shape = 5.5, scale = 0.74)	[3]
Floor-to-Ceiling Height	lnorm(meanlog = 0.93, sdlog = 0.034)	[4]
Window Opening Threshold	logis(location = 23.6, scale = 1.85)	[9]
Electrical Gains Factor	gamma(shape = 4.3, rate = 9.5)	[10]

Sources: [1] Hulme and Doran (2014); [2] BRE (2019); [3] 4M; [4] DLUHC (2021); [5] Stephen (2000); [6] BRE (2004); [7] Pan (2010); [8] CIBSE (2015); [9] Rijal et al. (2007); [10] Intertek (2012).

which was not shown to be significantly associated with summer living room indoor temperatures. Following data cleaning – see Petrou (2023) – the rest of this paper will concentrate on the largest cluster: semi-detached dwellings with filled cavity walls and double glazing.

4.3. Step 3: stochastic characterization

Through the application of the methods described in Section 3.6, a unimodal probability distribution was identified for ten out of the twelve continuous model inputs. Further discussion on selecting these distributions, and their goodness-of-fit, is provided in Petrou (2023). The two model inputs described by multimodal distributions are the Roof U-value and Floor U-value (Table 3). Whether further segmentation is recommended based on the modes of these two model inputs was determined by the sensitivity analysis in Section 4.4.

4.4. Step 4: sensitivity analysis

Floor U-value and Roof U-value, the two model inputs described by a multimodal distribution in Step 3 were found to be non-influential, thus, further segmentation based on their modes was not required (Table 4). Assuming an uncertainty of ±10% around the FAF value of 1.0 resulted in a comparatively small $\mu^* = 0.16$, almost 2.5 times smaller than the next largest μ^* and 43.5 times smaller than the most influential parameter. Therefore, variation within this bound is relatively unimportant and this parameter may be used as an explanatory variable. In both stages of the sensitivity analysis, WOT was the dominant model input and was selected for calibration, together with the Orientation, Glazing Fraction and Electrical Gains Factor (EGF).

Table 4. Summary of the rank and absolute mean of elementary effects (μ^*) for each parameter, in ascending order of Stage 2 rank.

Parameter	Stage 1 Rank (μ^*)	Stage 2 Rank (μ^*)	Type
Window Opening Threshold	**1 (7.20)**	**1 (6.96)**	**Calib.**
Orientation	**3 (1.21)**	**2 (1.10)**	**Calib.**
Glazing Fraction	**2 (1.40)**	**3 (0.97)**	**Calib.**
Electrical Gains Factor	**4 (1.14)**	**4 (0.86)**	**Calib.**
Permeability	6 (0.74)	5 (0.71)	Fixed
Wall U-value	7 (0.68)	6 (0.63)	Fixed
Window U-value	8 (0.57)	7 (0.58)	Fixed
Solar Absorptivity	9 (0.46)	8 (0.37)	Fixed
Floor Area Factor	5 (0.91)	9 (0.16)	Explan
Roof U-value	10 (0.11)	10 (0.08)	Fixed
Floor U-value	11 (0.09)	11 (0.08)	Fixed
Floor-to-Ceiling Height	12 (0.08)	12 (0.08)	Fixed

Notes: Type corresponds to how each parameter will be treated at the calibration step. Calibration parameters are in bold.

4.5. Step 5: Bayesian calibration

The daily mean MDLRT ($\overline{\mathbf{y}_{\mathbf{v}}^{(\mathbf{M})}}$) of the calibrated, bias-corrected, models ($\eta(x, w, t) + \delta(x, w)$) is compared against the uncalibrated model in Section 4.5.1. An in-depth exploration of the calibration results for a single model is presented in Section 4.5.2.

4.5.1. Parametric analysis

For the models where the outdoor temperature and GHI were the only weather variables used (EXP1–8), the only experiment that did not converge is EXP3 (Table 5). For the models that converged, their out-of-sample predictive performance was higher than the uncalibrated model according to RMSE, and MBE, but lower according to R^2. Specifically, RMSE reduced by 60.5–62.4% from a baseline of 2.53°C, while MBE decreased from −2.44°C by 93.1–93.9%. However, R^2 also reduced from 0.79 to 0.41–0.45.

Following the addition of a one-day lag component of the outdoor temperature (EXP1L1–EXP8L1), five out of the eight experiments converged (Table 5). The performance of these experiments is comparable, and has improved compared to the first set of experiments (EXP1–8); RMSE and MBE range between 0.64–0.70°C and 0.03–0.05°C, respectively. The R^2 for the calibrated models is lower (0.70–0.74) than that of the uncalibrated model (0.79), but higher than for EXP1–8 (0.41–0.45).

The addition of a second outdoor temperature lag component (EXP1L2–EXP8L2) resulted in further improvement in predictive performance across most metrics. RMSE and R^2 for EXP1L2–EXP8L2 have marginally improved when compared against EXP1L1–EXP8L1. The magnitude of MBE is comparable for the two sets of models and less than 0.1°C, but the sign differs. Two experiments (EXP1L2, EXP7L2) did not converge within 500 iterations.

Table 5. Validation metrics calculated over a 52-day period for the parametric calibration experiments.

	RMSE [°C] (ΔRMSE [%])	MBE [°C] (ΔMBE [%])	R^2	Time [hrs]
Uncalb.	2.53 (0)	−2.44 (0)	0.79	–
EXP1	0.98 (−61.2)	−0.16 (−93.5)	0.43	1.77
EXP2	1.00 (−60.6)	−0.16 (−93.3)	0.41	1.02
EXP3		–		
EXP4	1.00 (−60.5)	−0.15 (−93.9)	0.41	1.14
EXP5	0.96 (−61.9)	−0.15 (−93.9)	0.43	0.99
EXP6	0.96 (−62.1)	−0.16 (−93.3)	0.44	1.2
EXP7	0.95 (−62.3)	−0.17 (−93.1)	0.44	1.32
EXP8	0.95 (−62.4)	−0.17 (−93.2)	0.45	0.73
EXP1L1		–		
EXP2L1	0.64 (−74.5)	0.05 (−101.9)	0.73	1.35
EXP3L1		–		
EXP4L1	0.65 (−74.2)	0.04 (−101.8)	0.73	1.25
EXP5L1	0.70 (−72.5)	0.05 (−101.9)	0.7	1.36
EXP6L1	0.64 (−74.8)	0.03 (−101.1)	0.74	1.29
EXP7L1		–		
EXP8L1	0.65 (−74.2)	0.03 (−101)	0.72	1.2
EXP1L2		–		
EXP2L2	0.59 (−76.5)	−0.02 (−99.3)	0.77	1.53
EXP3L2	**0.58 (−76.9)**	**−0.04 (−98.2)**	**0.77**	**1.95**
EXP4L2	0.60 (−76.2)	−0.05 (−98.1)	0.76	2.13
EXP5L2	0.64 (−74.6)	−0.04 (−98.2)	0.74	1.68
EXP6L2	**0.59 (−76.8)**	**−0.05 (−98)**	**0.77**	**1.67**
EXP7L2		–		
EXP8L2	0.60 (−76.5)	−0.04 (−98.3)	0.77	1.66

Notes: RMSE = Root-mean-square error, ΔRMSE = Percentage change in RMSE post-calibration, MBE = mean bias error, ΔMBE = Percentage change in MBE post-calibration, R^2 = coefficient of determination, Time = Calibration computing time. Experiments EXP3, EXP1L1, EXP3L1, EXP7L1, EXP1L2 and EXP7L2 did not converge. Best performing models are in bold.

A common characteristic amongst all experiments that did not converge (EXP3, EXP1L1, EXP3L1, EXP7L1, EXP1L2, EXP7L2) was the use of Orientation and WOT as calibration parameters. Since other calibration experiments that included WOT did converge, including EXP8, EXP8L1 and EXP8L2 where WOT was the only calibration parameter, the use of Orientation may have contributed to the lack of convergence. A clear pattern regarding the time taken to complete 500 MCMC was not observed.

Within the first set of experiments (EXP1–8), the best-performing model is EXP8, WOT being the only calibration parameter. Amongst the models with a single lag component, the best-performing model is EXP6L1, where EGF was calibrated together with WOT. The best-performing model across all experiments is EXP3L2, with WOT, EGF and Orientation being calibrated, although its performance was only marginally better than EXP6L2 where Orientation was not used. Given the lack of convergence in other models that included Orientation, and the small difference in predictive performance between the two models, subsequent analysis will concentrate EXP6L2.

4.5.2. *Detailed analysis*

Figure 7 reveals that, in general, the calibrated model without bias-correction performs better (RMSE of 0.96°C) than the uncalibrated model (2.53°C) but worse than the calibrated model with bias-correction (RMSE = 0.59°C). For 34 out of the 52 days, the absolute difference between monitored data and predictions of the calibrated model with bias-correction is less than 0.5°C, while for 18 days the differences are less than 0.2°C.

The largest discrepancy occurred on the 3rd of July, when the mean calibrated prediction deviated from the mean MDLRT by 2.6°C (Figure 7). On the three days following the 3rd of July (4th–6th of July), the absolute discrepancies were within 0.3°C. According to Bastos and O'Hagan (2009), a single extreme point could indicate a local problem that might be addressed with the addition of more data points. Further, the extreme discrepancy of July 3rd is largely responsible for the marginally lower R^2 of the calibrated model (0.77) compared to the uncalibrated model (0.79). If R^2 were to be recalculated after excluding the 3rd of July, R^2 would improve as a result of the calibration from 0.81 to 0.86.[9] Thus, the calibrated model represents day-to-day fluctuations better than the uncalibrated model for most days.

The spread in the posterior distribution of WOT is smaller than its prior, with a posterior median of 21.8°C, and a 90% credible interval of 20.7–22.9°C (Figure 8). A *credible interval* contains a specified amount of posterior probability, in this case the central 90% probability. The prior and posterior distributions of EGF are similar, with a small shift in the median being observed post-calibration. The posterior median of EGF is 0.46 (3878 kWh/year) with a 90% credible interval of 0.19–0.93 (1602–7841 kWh/year). The hyperparameter posteriors can be found in Section 2.2.3.1 of the Supplementary Material.

With a median model bias of −0.5°C and a mean of −0.38°C, despite the use of a prior with a mean of zero, the computer model is more likely to over-predict even after the model calibration. Themodel biaswaspositively correlated with outdoor temperature and its lag components (see Section 2.2.3 of the Supplementary Material for further discussion).

5. Discussion

Bayesian calibration offers the potential to quantify and reduce uncertainties in archetype-based models. Following calibration, the gap between model outputs and real-world observations is expected to reduce, increasing the level of trust that may be placed on the model.

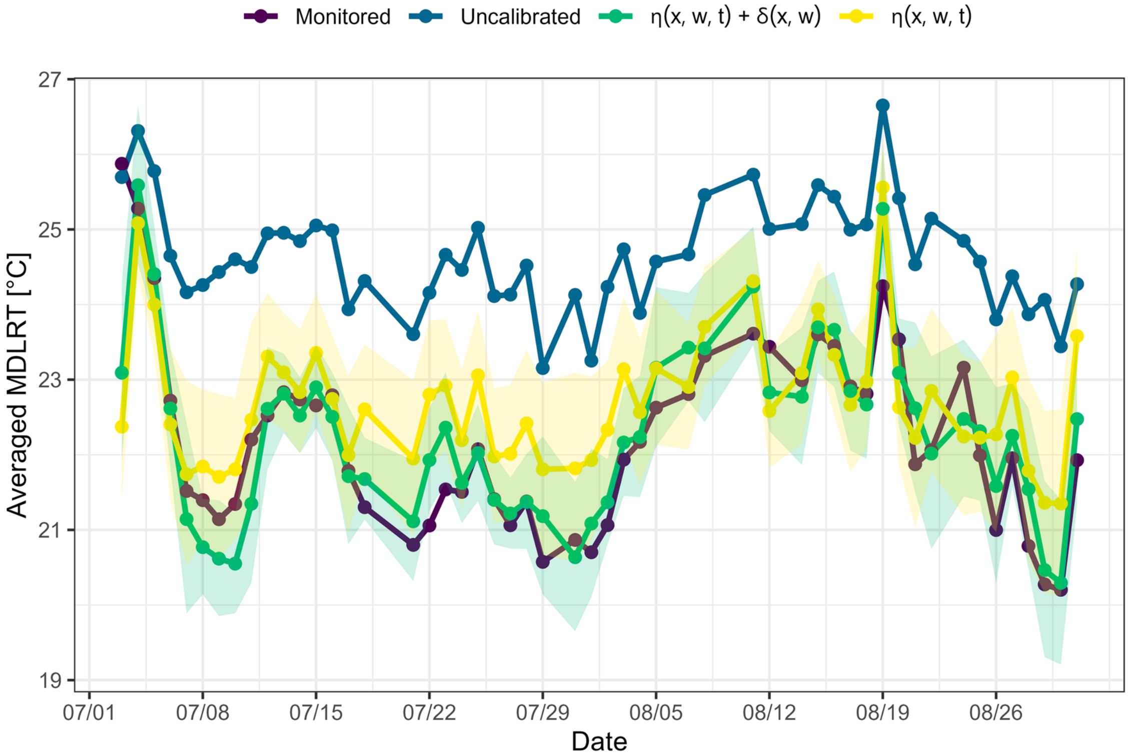


Figure 7. Timeseries plot of the mean daytime living room temperature, averaged per day across: (i) Monitored dwellings in the homogeneous cluster, (ii) Simulated (uncalibrated) dwellings, (iii) the bias-corrected calibrated model predictions ($\eta(x, w, t) + \delta(x, w)$), and (iv) the calibrated model predictions without model bias ($\eta(x, w, t)$). The shaded region represents an uncertainty of $\pm 1.96\sigma$ around the mean.

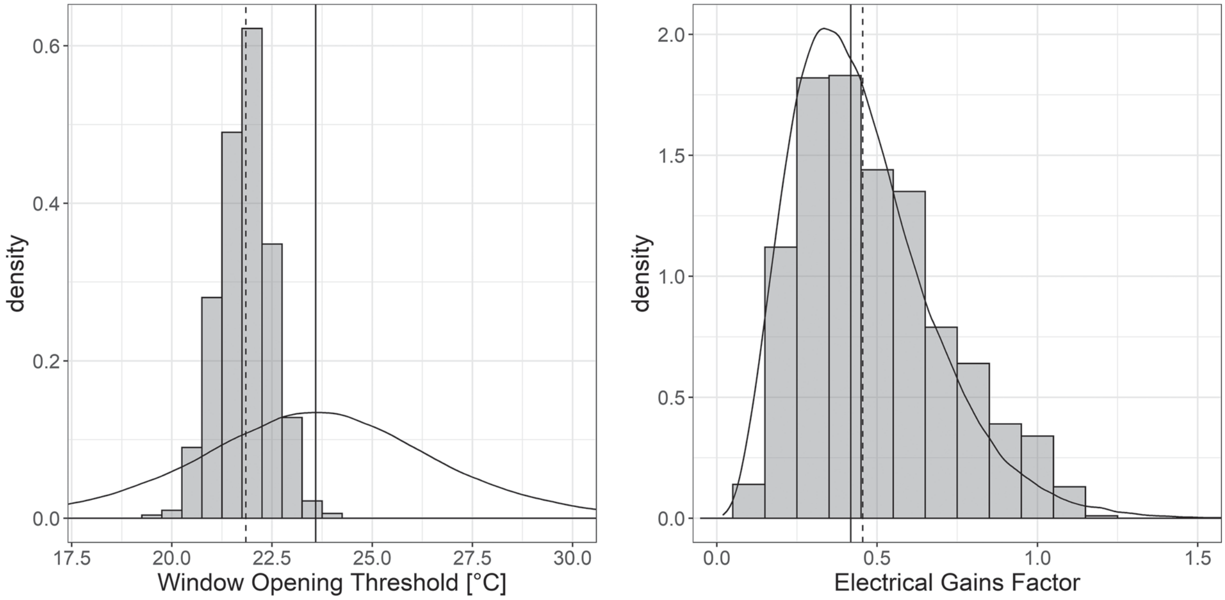


Figure 8. Density plot lines and histograms for the prior and posterior distributions of the calibration parameters, respectively. The vertical solid (dashed) line indicates the median of the prior (posterior) distribution. Area under each plot is unity.

Existing work on Bayesian calibration has largely focused on the reduction of the energy performance gap in archetype-based housing models, with limited work on models of summer indoor temperature (Calama-González, Suárez, and León-Rodríguez 2022). Furthermore, previous studies seldomly considered the classification of the housing stock into homogeneous groups in tandem with the calibration procedure, with a clear definition of homogeneity often lacking, and an *ad hoc* approach being commonly used (Petrou 2023). In addition, the choice of prior distributions for the calibration parameters – an essential ingredient in Bayesian calibration – has received limited attention to date. Building on existing work with the purpose of improving current practices, this paper presents a framework for the Bayesian calibration of archetype-based models of summer indoor temperature. The framework relies on a clear definition of homogeneity and proposes data-driven approaches to classify the housing stock into groups, to identify the calibration parameters and their corresponding prior probability distributions. To demonstrate its use, the framework was used to calibrate a previously developed archetype-based model (UK-HSM) and quantify the improvement in its out-of-sample predictive performance.

5.1. Main findings

By drawing on ten data sources, probability distributions were identified for all twelve continuous UK-HSM model inputs (Table 3). In all but two cases, non-normal and non-uniform distributions were found to best describe the possible values of the model inputs. Informed by these distributions, the sensitivity analysis revealed that Window Opening Threshold was the dominant UK-HSM model input, followed by the Glazing Fraction, Orientation and Electrical Gains Factor. This result provides further evidence of the importance of window opening in determining summer indoor temperatures in UK dwellings.

The parametric calibration revealed that including at least one lag component of outdoor temperature led to substantial improvements in predictive performance compared to a calibration with no lag components; the addition of a second lag component further improved out-of-sample predictions. Some parametric experiments did not converge, all of which included orientation as a calibration parameter. A possible reason for the lack of convergence in these models could be multi-modality in the posterior distributions of the orientation parameter that results in poor MCMC sampling.

For the calibrations that converged, RMSE reduced post-calibration from 2.53°C to 0.58–0.70°C. Such performance exceeds what has been previously achieved (RMSE of 0.94–1.73°C) for the semi-detached archetype of UK-HSM (Symonds et al. 2017). Even more encouraging is the fact that MBE for the calibrated models was less than 0.1°C, comparable to values obtained from the calibration of a test cell (Calama-González et al. 2021), suggesting that the calibrated model does not have a tendency to under- or over-predict. This contrasts with the tendency of the uncalibrated UK-HSM to overpredict (Figure 7). The calibration process came at a computational cost of 0.73–2.13 h (Table 5), a cost thought to be warranted given the improvement in RMSE and MBE.

A comparison between prior and posterior distributions for one of the best-performing models revealed the EGF distributions to be similar, indicating confirmation of the prior knowledge, or parameter non-identifiability (Menberg, Heo, and Choudhary 2019). Further calibration using a different prior for EGF provided further evidence to the lack of identifiability (Section 2.2.4 of Supplementary Material). The improvement in out-of-sample prediction despite the suspected lack of parameter identifiability is not surprising, and it has been previously observed (Arendt, Apley, and Chen 2012). On the interpretation of the posterior distributions as real-world physical quantities, caution should be applied as highlighted by (Booth, Choudhary, and Spiegelhalter 2012; Kennedy and O'Hagan 2001). The extent to which the posterior distributions are representative of the physical quantities can seldomly be verified, since a ground truth[10] about the physical quantities does not exist, unless a study is concerned with synthetic or test cell data.

The association between the model bias and the lag components of outdoors temperature may indicate limitations in the modelling of thermal mass by UK-HSM. However, further investigation is required to understand this finding.

5.2. Strengths and contributions

A key contribution of this study to academic research is the development and application of a Bayesian calibration framework that has several potential uses within the field of building modelling. The proposed framework is modular and flexible; a modeller can choose what methods to use in each step depending on the data available, their model and preference. It was demonstrated that the framework application on UK-HSM substantially improved its predictive performance while simultaneously apportioning uncertainties to different sources. While developed with models of summer indoor temperature in mind, it is expected that the framework can be tailored to the use of different types of archetype-based building stock models, such as those of winter indoor temperature, energy use or fuel poverty. Although the framework could be applied with a relatively small amount of data, it is likely to be most beneficial and suitable for applications with large datasets.

The methodology used for identifying prior probability distributions depending on the data available can inform future Bayesian calibration work. Further, this approach may be of value in other analyses beyond Bayesian calibration where identifying analytical distributions is required.

A further novel contribution is the set of learnings derived from the first application of Bayesian calibration on archetype-based models of free-floating summer indoor temperature. One such learning relates to the importance of outdoor temperature lag components, which have not been previously used or discussed in published work on archetype-based Bayesian calibration.

Through the potential for improvement in modelling practice offered by this work, there could be indirect benefits to the construction industry, energy and public health policy-making where that may be informed by archetype-based models.

5.3. *Limitations*

The use of a parametric approach highlighted the impact that weather variables, and specifically the lag components of outdoor temperature, can have on the calibration of MDLRT. While informative, the parametric experiment was not exhaustive. For instance, the inclusion of a third outdoor temperature lag component could have led to further model improvement, although any further improvement was expected to be marginal.

The choice of MDLRT was informed by previous research utilizing UK-HSM to evaluate heat-related mortality changes due to home energy efficiency measures. While findings from this research will guide the future use of UK-HSM, their generalizability to other models of summer indoor temperature remains uncertain, especially when modelling buildings at different temporal resolutions, with heating or air conditioning. Nevertheless, the framework may still be applied to calibrate such models.

This study did not attempt to optimize the choice of hyperparameter priors but has instead relied on published recommendations (Chong and Menberg 2018; Menberg, Heo, and Choudhary 2019). Recent work suggests that refining the hyperparameter priors can be of benefit (Wang et al. 2022), although it is expected to only offer a small improvement in comparison to the overall improvement.

In practice, the framework's implementation will often be impacted by the limited availability of empirical data. As a result, classifiers may not all be identified or used, and the priors may be poorly defined. In turn, the attribution of uncertainties, along with parameter inference and predictive performance may be affected. The framework's users should attempt to utilize all evidence available and carefully apply expert judgment where needed to reduce the impact of limited data. Moreover, effort is required to collect and make openly available large-scale datasets of linked indoor temperature, dwelling and household characteristics; such actions would enable further development and comprehensive calibration of building stock models.

Finally, it was assumed that explanatory and weather variables did not have measurement error (or that its influence was negligible). This is a common (Booth, Choudhary, and Spiegelhalter 2012; Chong and Menberg 2018), yet simplifying assumption, whose impact was not investigated.

6. Conclusions

This paper developed and applied a modular Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature. The framework relies on a practical definition of homogeneity and covers the steps of data-driven classification, stochastic characterization, sensitivity analysis, and calibration through Bayesian inference. The framework's use was demonstrated using data collected from 193 dwellings located in Leicester, monitored as part of the 4M project, and the UK Housing Stock Model (UK-HSM). The calibration was carried out using the Mean of the Daytime Living Room Temperature for the chosen cluster of semi-detached dwellings. To assess the impact that the choice of calibration parameter and weather variables can have on the outcome, the calibration was carried out for 18 different combinations of variables.

This work revealed that the inclusion of lag components of outdoor temperature resulted in improved performance, and the temperature at which windows open was the dominant model input for UK-HSM. Following calibration, the model's root-mean-square error reduced by ∼77% (to ∼0.6°C), and the uncertainties from different sources were quantified. The framework offers the potential to reduce the performance gap in various modelling applications, increasing the trust that can be placed on archetype-based models, and thus their utility as tools to inform policy-making.

Acknowledgements

We would like to acknowledge Professor David Allinson for his support regarding the provision and use of the 4M dataset.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was made possible by support from the EPSRC Centre for Doctoral Training in Energy Demand (LoLo) [grant numbers EP/L01517X/1 and EP/H009612/1], the National Institute for Health and Care Research (NIHR) Health Protection

Research Unit in Environmental Change and Health [grant number NIHR200909] and the Chartered Institution of Building Services Engineers.

Data availability statement

Data available on reasonable request from the authors.

Notes

1. A prior distribution captures how plausible each value of a parameter is, according to the modeller's subjective opinion, before observing the data (Bolstad and Curran 2017).
2. Non-identifiability, indicated by posterior distributions that are weak (uninformative) or mirror the priors, arises when a unique combination of calibration parameters does not exist or cannot be determined by the currently available data (Chong and Menberg 2018; Menberg, Heo, and Choudhary 2019).
3. A *posterior* distribution represents the relative weights of belief for each parameter value, estimated after the application of Bayes' theorem (Bolstad and Curran 2017): $Posterior = \frac{Likelihood \times Prior}{Average\ probability\ of\ the\ data}$.
4. Assuming the observations are fixed, the *likelihood* represents the relative weights of belief for the observed data for different values of the unknown parameters (Bolstad and Curran 2017).
5. The combined vector $\mathbf{z}$ is a subset of the combined vector of all observations, $[\mathbf{y}^{(M)}, \mathbf{y}^{(S)}]$, which has been normalised to have a mean of 0 and variance of 1.
6. A *free-running building* does not make use of mechanical heating or cooling. This is the case for the homes considered in this study during the summer period.
7. For day d, with weather observations $\mathbf{w}_d$, the one-day lag components are the weather observations of the previous day ($\mathbf{w}_{d-1}$).
8. This refers to the SAP rating estimated using the 2009 version of the Standard Assessment Procedure (SAP), and it is a measure of the floor area adjusted energy costs (BRE 2011).
9. The R^2 of the uncalibrated model would also improve if this point was excluded in the validation procedure.
10. Ground truth in this context refers to information provided by direct observation as opposed to information provided by inference (*Ground truth definition and meaning | Collins English Dictionary*, n.d.).

ORCID

Giorgos Petrou http://orcid.org/0000-0002-1524-2681
Anna Mavrogianni http://orcid.org/0000-0002-5104-1238
Phil Symonds http://orcid.org/0000-0002-6290-5417
Kevin Lomas http://orcid.org/0000-0001-5792-0762
Michael Davies http://orcid.org/0000-0003-2173-7063

References

Arendt, P. D., D. W. Apley, and W. Chen. 2012. "Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability." *Journal of Mechanical Design* 134 (10): 100908. https://doi.org/10.1115/1.4007390.

Baba, F. M., H. Ge, R. Zmeureanu, and L. (Leon) Wang. 2022. "Calibration of Building Model Based on Indoor Temperature for Overheating Assessment Using Genetic Algorithm: Methodology, Evaluation Criteria, and Case Study." *Building and Environment* 207:108518. https://doi.org/10.1016/j.buildenv.2021.108518.

Bastos, L. S., and A. O'Hagan. 2009. "Diagnostics for Gaussian Process Emulators." *Technometrics* 51 (4): 425–438. https://doi.org/10.1198/TECH.2009.08019.

Bolstad, W. M., and J. M. Curran. 2017. *Introduction to Bayesian Statistics*. 3rd ed. Hoboken, NJ: Wiley.

Booth, A. T., R. Choudhary, and D. J. Spiegelhalter. 2012. "Handling Uncertainty in Housing Stock Models." *Building and Environment* 48 (Supplement C): 35–47. https://doi.org/10.1016/j.buildenv.2011.08.016.

BRE. 2004. *Assessment of Energy Efficiency Impact of Building Regulations Compliance*.

BRE. 2011. *The Government's Standard Assessment Procedure for Energy Rating of Dwellings. 2009 Edition Incorporating RdSAP* 2009. Building Research Establishment (BRE) on Behalf of the Department of Energy and Climate Change (DECC).

BRE. 2019. *Appendix S: Reduced Data SAP for Existing Dwellings*.

Calama-González, C. M., R. Suárez, and Á. L. León-Rodríguez. 2022. "Thermal Comfort Prediction of the Existing Housing Stock in Southern Spain Through Calibrated and Validated Parameterized Simulation Models." *Energy and Buildings* 254:111562. https://doi.org/10.1016/j.enbuild.2021.111562.

Calama-González, C. M., P. Symonds, G. Petrou, R. Suárez, and Á. L. León-Rodríguez. 2021. "Bayesian Calibration of Building Energy Models for Uncertainty Analysis Through Test Cells Monitoring." *Applied Energy* 282:116118. https://doi.org/10.1016/j.apenergy.2020.116118.

Cerezo, C., J. Sokol, S. AlKhaled, C. Reinhart, A. Al-Mumin, and A. Hajiah. 2017. "Comparison of Four Building Archetype Characterization Methods in Urban Building Energy Modeling (UBEM): A Residential Case Study in Kuwait City." *Energy and Buildings* 154:321–334. https://doi.org/10.1016/j.enbuild.2017.08.029.

Chong, A., Y. Gu, and H. Jia. 2021. "Calibrating Building Energy Simulation Models: A Review of the Basics to Guide Future Work." *Energy and Buildings* 253:111533. https://doi.org/10.1016/j.enbuild.2021.111533.

Chong, A., K. P. Lam, M. Pozzi, and J. Yang. 2017. "Bayesian Calibration of Building Energy Models with Large Datasets." *Energy and Buildings* 154:343–355. https://doi.org/10.1016/j.enbuild.2017.08.069.

Chong, A., and K. Menberg. 2018. "Guidelines for the Bayesian Calibration of Building Energy Models." *Energy and Buildings* 174:527–547. https://doi.org/10.1016/j.enbuild.2018.06.028.

CIBSE. 2015. *Guide A, Environmental Design*. London: Chartered Institute of Building Services Engineers.

Coakley, D., P. Raftery, and M. Keane. 2014. "A Review of Methods to Match Building Energy Simulation Models to Measured Data." *Renewable and Sustainable Energy Reviews* 37:123–141. https://doi.org/10.1016/j.rser.2014.05.007.

de Wilde, P. 2014. "The Gap Between Predicted and Measured Energy Performance of Buildings: A Framework for Investigation." *Automation in Construction* 41:40–49. https://doi.org/10.1016/j.autcon.2014.02.009.

de Wilde, P. 2023. "Building Performance Simulation in the Brave New World of Artificial Intelligence and Digital Twins: A Systematic Review." *Energy and Buildings* 292:113171. https://doi.org/10.1016/j.enbuild.2023.113171.

DLUHC. 2021. *English Housing Survey*. Accessed February 26, 2022. https://www.gov.uk/government/collections/english-housing-survey.

European Environment Agency. 2022. *Economic Losses and Fatalities from Weather- and Climate-Related Events in Europe.* LU: Publications Office. Accessed January 18, 2024. https://data.europa.eu/doi/10.2800530599.

Ground Truth Definition and Meaning | Collins English Dictionary. n.d. *Collins English Dictionary*. Accessed January 8, 2023. https://www.collinsdictionary.com/dictionary/english/ground-truth.

Hajat, S., S. Vardoulakis, C. Heaviside, and B. Eggen. 2014. "Climate Change Effects on Human Health: Projections of Temperature-Related Mortality for the UK During the 2020s, 2050s and 2080s." *Journal of Epidemiology and Community Health* 68 (7): 641–648. https://doi.org/10.1136/jech-2013-202449.

Herman, J., and W. Usher. 2017. "SALib: An Open-Source Python Library for Sensitivity Analysis." *The Journal of Open Source Software* 2 (9): 97. https://doi.org/10.21105/joss.00097.

Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. 2004. "Combining Field Data and Computer Simulations for Calibration and Prediction." *SIAM Journal on Scientific Computing* 26 (2): 448–466. https://doi.org/10.1137/S1064827503426693.

Hou, D., I. G. Hassan, and L. Wang. 2021. "Review on Building Energy Model Calibration by Bayesian Inference." *Renewable and Sustainable Energy Reviews* 143:110930. https://doi.org/10.1016/j.rser.2021.110930.

Hulme, J., and S. Doran. 2014. *In-Situ Measurements of Wall U-Values in English Housing*. 290-102. Building Research Establishment (BRE) on Behalf of the Department of Energy and Climate Change (DECC).

Intertek. 2012. *Household Electricity Survey: A Study of Domestic Electrical Product Usage*.

IPCC. 2021. "Summary for Policymakers." In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Iwanaga, T., W. Usher, and J. Herman. 2022. "Toward SALib 2.0: Advancing the Accessibility and Interpretability of Global Sensitivity Analyses." *Socio-Environmental Systems Modelling* 4:18155. https://doi.org/10.18174/sesmo.18155.

Jain, N., E. Burman, S. Stamp, D. Mumovic, and M. Davies. 2020. "Cross-Sectoral Assessment of the Performance Gap Using Calibrated Building Energy Performance Simulation." *Energy and Buildings* 224:110271. https://doi.org/10.1016/j.enbuild.2020.110271.

Kennedy, M., and A. O'Hagan. 2001. "Bayesian Calibration of Computer Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (3): 425–464. https://doi.org/10.1111/1467-9868.00294.

Kovats, S., and R. Brisley. 2021. "Health, Communities and the Built Environment." In *The Third UK Climate Change Risk Assessment Technical Report*, edited by R. A. Betts, A. B. Haward, and K. V. Pearson. London.

Kristensen, M. H., R. Choudhary, R. H. Pedersen, and S. Petersen. 2017. "Bayesian Calibration of Residential Building Clusters Using a Single Geometric Building Representation." In *15th International Conference of the International Building Performance Simulation Association, Building Simulation 2017*, San Francisco, CA, 2251–2260.

Lomas, K. J., and T. Kane. 2013. "Summertime Temperatures and Thermal Comfort in UK Homes." *Building Research & Information* 41 (3): 259–280. https://doi.org/10.1080/09613218.2013.757886.

Lomas, K. J., S. Watson, D. Allinson, A. Fateh, A. Beaumont, J. Allen, H. Foster, and H. Garrett. 2021. "Dwelling and Household Characteristics' Influence on Reported and Measured Summertime Overheating: A Glimpse of a Mild Climate in the 2050s." *Building and Environment* 201:107986. https://doi.org/10.1016/j.buildenv.2021.107986.

Macintyre, H., and P. Murage. 2023. "Chapter 2. Temperature Effects on Mortality in a Changing Climate." In *Health Effects of Climate Change (HECC) in the UK: 2023 Report*.

Menberg, K., Y. Heo, and R. Choudhary. 2019. "Influence of Error Terms in Bayesian Calibration of Energy System Models." *Journal of Building Performance Simulation* 12 (1): 82–96. https://doi.org/10.1080/19401493.2018.1475506.

Met Office. 2018. *Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (1853–Current)*. Accessed January 15, 2019. http://catalogue.ceda.ac.uk/uuid/220a65615218d5c9cc9e4785a3234bd0.

Mun, J. 2012. "Understanding and Choosing the Right Probability Distributions." In *Advanced Analytical Models*, edited by Johnathan Mun, 899–917. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781119197096.app03.

NOAA. 2024. *2023 was the World's Warmest Year on Record, by Far*. National Oceanic and Atmospheric Administration. Accessed January 18, 2024. https://www.noaa.gov/news/2023-was-worlds-warmest-year-on-record-by-far.

Oikonomou, E., A. Mavrogianni, R. Raslan, J. Taylor, and M. Davies. 2018. *English Archetypes*. Accessed June 22, 2022. https://www.ucl.ac.uk/energy-models/models/english-archetypes.

Oraiopoulos, A., and B. Howard. 2022. "On the Accuracy of Urban Building Energy Modelling." *Renewable and Sustainable Energy Reviews* 158:111976. https://doi.org/10.1016/j.rser.2021.111976.

Pan, W. 2010. "Relationships Between Air-Tightness and Its Influencing Factors of Post-2006 New-Build Dwellings in the UK." *Building and Environment* 45 (11): 2387–2399. https://doi.org/10.1016/j.buildenv.2010.04.011.

Petersen, S., M. H. Kristensen, and M. D. Knudsen. 2019. "Prerequisites for Reliable Sensitivity Analysis of a High Fidelity Building Energy Model." *Energy and Buildings* 183:1–16. https://doi.org/10.1016/j.enbuild.2018.10.035.

Petrou, G. 2023. "Development of a Bayesian Calibration Framework for Archetype-Based Housing Stock Models of Summer Indoor Temperature." University College London. https://discovery.ucl.ac.uk/id/eprint/10163978/.

Petrou, G., A. Mavrogianni, P. Symonds, and M. Davies. 2021. "Beyond Normal: Guidelines on How to Identify Suitable Model Input Distributions for Building Performance Analysis." In *2021 Building Simulation Conference*, Bruges, Belgium, 1421–1428. https://doi.org/10.26868/25222708.2021.30333.

Petrou, G., P. Symonds, A. Mavrogianni, A. Mylona, and M. Davies. 2019. "The Summer Indoor Temperatures of the English Housing Stock: Exploring the Influence of Dwelling and Household Characteristics." *Building Services Engineering Research and Technology* 40 (4): 492–511. https://doi.org/10.1177/0143624419847621.

Reinhart, C. F., and C. Cerezo Davila. 2016. "Urban Building Energy Modeling – a Review of a Nascent Field."

Building and Environment 97 (Supplement C): 196–202. https://doi.org/10.1016/j.buildenv.2015.12.001.

Rijal, H. B., P. Tuohy, M. A. Humphreys, J. F. Nicol, A. Samuel, and J. Clarke. 2007. "Using Results from Field Surveys to Predict the Effect of Open Windows on Thermal Comfort and Energy Use in Buildings." *Energy and Buildings* 39 (7): 823–836. https://doi.org/10.1016/j.enbuild.2007.02.003.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, eds. 2008. *Global Sensitivity Analysis: The Primer*. Chichester: John Wiley.

Schweizer, C., R. D. Edwards, L. Bayer-Oglesby, W. J. Gauderman, V. Ilacqua, M. Juhani Jantunen, H. K. Lai, M. Nieuwenhuijsen, and N. Künzli. 2007. "Indoor Time–Microenvironment–Activity Patterns in Seven Regions of Europe." *Journal of Exposure Science & Environmental Epidemiology* 17 (2): 170–181. https://doi.org/10.1038/sj.jes.7500490.

Smith, R. C. 2013. *Uncertainty Quantification: Theory, Implementation, and Applications*. Philadelphia: Society for Industrial and Applied Mathematics (Computational Science and Engineering Series).

Sokol, J., C. Cerezo Davila, and C. F. Reinhart. 2017. "Validation of a Bayesian-Based Method for Defining Residential Archetypes in Urban Building Energy Models." *Energy and Buildings* 134 (Supplement C): 11–24. https://doi.org/10.1016/j.enbuild.2016.10.050.

Stephen, R. 2000. *Airtightness in UK Dwellings*. Bracknell: BRE Press.

Symonds, P., J. Taylor, Z. Chalabi, A. Mavrogianni, M. Davies, I. Hamilton, S. Vardoulakis, C. Heaviside, and H. Macintyre. 2016. "Development of an England-Wide Indoor Overheating and Air Pollution Model Using Artificial Neural Networks." *Journal of Building Performance Simulation* 9 (6): 606–619. https://doi.org/10.1080/19401493.2016.1166265.

Symonds, P., J. Taylor, A. Mavrogianni, M. Davies, C. Shrubsole, I. Hamilton, and Z. Chalabi. 2017. "Overheating in English Dwellings: Comparing Modelled and Monitored Large-Scale Datasets." *Building Research & Information* 45 (1–2): 195–208. https://doi.org/10.1080/09613218.2016.1224675.

Taylor, J., R. McLeod, G. Petrou, C. Hopfe, A. Mavrogianni, R. Castaño-Rosa, S. Pelsmakers, and K. Lomas. 2023. "Ten Questions Concerning Residential Overheating in Central and Northern Europe." *Building and Environment* 234:110154. https://doi.org/10.1016/j.buildenv.2023.110154.

Taylor, J., P. Symonds, C. Heaviside, Z. Chalabi, M. Davies, and P. Wilkinson. 2021. "Projecting the Impacts of Housing on Temperature-Related Mortality in London During Typical Future Years." *Energy and Buildings* 249:111233. https://doi.org/10.1016/j.enbuild.2021.111233.

Taylor, J., P. Symonds, P. Wilkinson, C. Heaviside, H. Macintyre, M. Davies, A. Mavrogianni, and E. Hutchinson. 2018. "Estimating the Influence of Housing Energy Efficiency and Overheating Adaptations on Heat-Related Mortality in the West Midlands, UK." *Atmosphere* 9 (5): 190. https://doi.org/10.3390/atmos9050190.

Tian, W., Y. Heo, P. de Wilde, Z. Li, D. Yan, C. S. Park, X. Feng, and G. Augenbroe. 2018. "A Review of Uncertainty Analysis in Building Energy Assessment." *Renewable and Sustainable Energy Reviews* 93:285–301. https://doi.org/10.1016/j.rser.2018.05.029.

Wang, Y., Y. Shangguan, Z. Wang, and Y. Xue. 2022. "The Influence and Adjust Method of Hyperparameters' Prior Distributions in Bayesian Calibration for Building Stock Energy Prediction." *Energy and Buildings* 273:112413. https://doi.org/10.1016/j.enbuild.2022.112413.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org.

Yun, J., M. Shin, I. H. Jin, and F. Liang. 2020. "Stochastic Approximation Hamiltonian Monte Carlo." *Journal of Statistical Computation and Simulation* 90 (17): 3135–3156. https://doi.org/10.1080/00949655.2020.1797031.