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

Giorgos Petrou

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Bartlett School of Energy, Environment and Resources

University College London

29th January 2023

I, Giorgos Petrou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Adverse effects to health and wellbeing from increased exposure to heat at home has been repeatedly identified as a major climate change adaptation risk in the United Kingdom by the Climate Change Committee and others. Despite recent progress, policy gaps in the adaptation of the housing stock exist. The development of such policies can be guided by housing stock models, that enable the assessment of the impact of climate change adaptation and energy efficiency measures on building performance under different climate scenarios. To ensure well-informed decision-making, uncertainties in these models should be considered. Motivated by the lack of work on this topic, this thesis aims to quantify and reduce uncertainties of archetype-based housing stock models of summer indoor temperature through a Bayesian calibration framework.

The framework includes the data-driven classification of dwellings into homogeneous groups, the characterisation of model input uncertainty in the form of probability distributions – which can be used as calibration priors – and their reduction through Bayesian inference. The framework's implementation was demonstrated using the 'UK Housing Stock Model' (a bottom-up model based on EnergyPlus), the 2011 English Housing Survey and Energy Follow-Up Survey (EHS-EFUS), and the 2009 4M survey in Leicester. The model's root-mean-square error reduced from $2.5\,^{\circ}\mathrm{C}$ (pre-calibration) to $0.6\,^{\circ}\mathrm{C}$ (post-calibration), while input and structural uncertainties were quantified.

This work offers several novel contributions, including a modular framework that can be adapted for the improvement of other archetype-based housing stock models, an open-source method for identifying model input probability distributions,

and an alternative formulation of Gaussian processes that substantially reduces the computational cost of Bayesian calibration. Learnings from this first calibration of its type can inform future academic research. Finally, the analysis of 2011 EHS-EFUS provides evidence to building designers and policymakers on the dwelling and household characteristics associated with high summer indoor temperatures.

Impact Statement

This thesis describes the research carried out in response to the aim of quantifying and reducing uncertainties in archetype-based housing stock models of summer indoor temperature. The contributions of this study to academic research and knowledge are manifold. A modular Bayesian calibration framework was developed, which can be used to quantify and reduce uncertainties, and improve the predictive performance of archetype-based models of summer indoor temperature. In addition, the framework can be adapted for other archetype-based models, such as those of winter indoor temperature, energy use, ventilation or indoor air quality. Through the framework's application on the bottom-up UK Housing Stock Model, the first Bayesian calibration of an archetype-based housing stock model of summer indoor temperature, insights were generated that can inform future research. Furthermore, a novel and open-source method for identifying probability distributions that adequately describe empirical data has been introduced, which can find multiple applications within the field of building energy modelling. An alternative formulation of Gaussian processes, proposed in this work, offers a substantial reduction in the computational cost of Bayesian calibration.

The outcomes of this thesis also have implications for industry practitioners. Findings from the statistical analysis of summer indoor temperatures, monitored during the 2011 English Housing Survey Energy Follow-Up Survey (EHS-EFUS), provide further evidence to building designers on the factors associated with indoor overheating. The differences between stated thermal discomfort and quantified overheating for the 2011 EHS-EFUS, using metrics defined in the Chartered Institution of Building Services Engineers (CIBSE) Technical Memorandum 59 (TM59), suggest

a need to review and potentially refine overheating criteria. The discrepancy between modelled and monitored indoor temperature, and its reduction through calibration, highlight the importance of adopting uncertainty quantification and reduction techniques within industry. This is especially pertinent given the use of Dynamic Thermal Simulation (DTS) tools to demonstrate compliance with the building regulations (Part O) and for assessing indoor overheating. The methodological advancements offered by this thesis on surrogate modelling, uncertainty quantification and reduction, can be integrated in DTS tools, and enable such work to be undertaken at a more manageable computational cost.

A direct impact on policymaking is the provision of evidence on the factors associated with high summer indoor temperature, which can inform building regulations and national climate change adaptation plans. For example, households with at least one occupant on means tested, or certain disability-related benefits were associated with higher summer temperatures. Such findings should be taken into consideration when devising an indoor overheating adaptation strategy. An indirect, yet important, contribution to policy follows from the advancement of uncertainty quantification and reduction methods and knowledge within the field of built environment research. The use of calibrated housing stock models of summer indoor temperature that provide a quantified estimate of uncertainty encourages more robust and better-informed decision making.

Publications

Part of the work presented in this thesis has been published in a peer-reviewed journal and conference paper:

Petrou, G., Symonds, P., Mavrogianni, A., Mylona, A., & Davies, M. (2019). The summer indoor temperatures of the English housing stock: Exploring the influence of dwelling and household characteristics. *Building Services Engineering Research and Technology*. doi:10.1177/0143624419847621

Petrou, G., Mavrogianni, A., Symonds, P., & Davies, M. (2021). Beyond Normal: Guidelines on How to Identify Suitable Model Input Distributions for Building Performance Analysis. *Proceedings of Building Simulation 2021*: 17th Conference of IBPSA. doi:10.26868/25222708.2021.30333

I have also led or contributed to several other related publications within the field of built environment research since the start of my doctoral work:

Petrou, G., Hutchinson, E., Mavrogianni, A., Milner, J., Macintyre, H., Phalkey, R., Hsu, S. C., Symonds, P., et al. (2022). Home energy efficiency under net zero: time to monitor UK indoor air. *BMJ*, 377, e069435. doi:10.1136/bmj-2021-069435

Tsoulou, I., Jain, N., Oikonomou, E., **Petrou, G.**, Howard, A., Gupta, R., Mavrogianni, A., Milojevic, A., et al. (2021). Assessing the Current and Future Risk of Overheating in London's Care Homes: The Effect of Passive Ventilation. *Proceedings of Building Simulation 2021*: 17th Conference of IBPSA. doi:10.26868/25222708.2021.30677

Petrou, G., Mavrogianni, A., Symonds, P., & Davies, M. (2021). A case study on the impact of fixed input parameter values in the modelling of indoor overheating. *Journal of Physics: Conference Series.* doi:10.1088/1742-6596/2069/1/012137

Calama-González, C. M., Symonds, P., **Petrou, G.**, Suárez, R., & León-Rodríguez, Á. L. (2021). Bayesian calibration of building energy models for uncertainty analysis through test cells monitoring. *Applied Energy*, 282. doi:10.1016/j.apenergy.2020.116118

Ibbetson, A., Milojevic, A., Mavrogianni, A., Oikonomou, E., Jain, N., Tsoulou, I., **Petrou G.**, Gupta R., Davies M., Wilkinson, P. (2021). Mortality benefit of building adaptations to protect care home residents against heat risks in the context of uncertainty over loss of life expectancy from heat. *Climate Risk Management*, 32. doi:10.1016/j.crm.2021.100307

Gustin, M., McLeod, R. S., Lomas, K. J., **Petrou, G.**, & Mavrogianni, A. (2020). A high-resolution indoor heat-health warning system for dwellings. *Building and Environment*, 168. doi:10.1016/j.buildenv.2019.106519

Vallejo Espinosa, A. L., Symonds, P., & **Petrou, G.** (2019). Modelling the Influence of Layout On Overheating Risk of London Flats. *Proceedings of Building Simulation 2019: 16th Conference of IBPSA. IBPSA.* doi:10.26868/25222708.2019.210891

Petrou, G., Mavrogianni, A., Symonds, P., Mylona, A., Dane, V., Raslan, R., & Davies, M. (2018). Can the choice of building performance simulation tool significantly alter the level of predicted indoor overheating risk in London flats?. *Building Services Engineering Research and Technology.* doi:10.1177/0143624418792340

Petrou, G., Mavrogianni, A., Symonds, P., Korolija, I., Mylona, A., Raslan, R., Dane, V., Davies, M. (2018). What are the implications of algorithm choice on the overheating risk assessment?. *Proceedings of the Building Simulation and Optimization 2018.* University of Cambridge.

Petrou, G., Mavrogianni, A., Mylona, A., Raslan, R., Virk, G., & Davies, M. (2017). Inter-model comparison of indoor overheating risk prediction for English dwellings. 38th Air Infiltration and Ventilation Centre (AIVC) Conference.

Acknowledgements

In writing this section, I was struck by the number of people who have contributed, inspired or in some way helped me through my doctoral studies. I will not name all of them, but I am thankful to everyone who has played a part.

I would first like to thank my primary supervisor, Prof. Anna Mavrogianni, for the continuous and unwavering support over the last few years. I started this PhD with limited understanding of built environment research, and Anna has patiently and enthusiastically guided me throughout this journey. My subsidiary supervisors, Dr Phil Symonds and Prof. Mike Davies have always been happy to discuss my progress, provide the right advice at the right time, challenge me and encourage me when needed. Dr Anastasia Mylona, my industry-based supervisor (Chartered Institution of Building Services Engineers), has consistently provided me with timely and invaluable feedback. Further, I would like to extend my gratitude to Dr Zaid Chalabi for his helpful feedback on the mathematical sections of this thesis. I would also like to thank Dr Dane Virk for his willingness to guide me in the initial stages of my PhD.

I've had a great time being part of the LoLo CDT, and for this I would like to thank all my fellow CDT friends and the CDT management team, especially Mae Oroszlany, Dr Jenny Crawley and Prof. Cliff Elwell.

Although we haven't been in contact since my undergraduate studies, I would like to thank my personal tutor Dr Marco Polin. When I mentioned to Marco – on the day before the LoLo CDT recruitment deadline – that I was interested in this PhD but had decided not to apply due to being overwhelmed with exam preparation and thinking that I did not stand a good chance, he helped me see the wood for the

trees and convinced me to apply.

I can genuinely say that I have enjoyed the full support of my parents, Petros and Maro, and my brother Stelios, throughout my life, and words cannot express how grateful I am. I would also like to thank my friends who have supported me, in their own way, through this journey. I'm confident that many of them are still not sure what my PhD is about, and some of them are still wondering why I chose to research indoor overheating when air-conditioning exists.

Finally, I would like to thank Jess for a myriad of things, including proofreading the entire thesis, for patiently listening to me go on about Gaussian processes and Bayesian calibration for hours while I was trying to make sense of them, and for ensuring that I take breaks (when I would otherwise stubbornly work through the weekend). Her unrivalled encouragement, especially when I was writing the thesis while working full time, has been vital.

Contents

List of Figures

List of Tables

Nomenclature

Chapter 2

i index for observations, known conditions and errors $i = 1, 2, \ldots, n$

n number of observations and associated known conditions and observation errors

$p(\mathbf{y})$ average probability of the data

$p(\mathbf{y} \mid \boldsymbol{\theta})$ probability of the data (likelihood)

$p(\boldsymbol{\theta})$ prior probability distributions

$p(\boldsymbol{\theta} \mid \mathbf{y})$ posterior probability distributions

$\mathbf{t}$ vector of calibration parameters

$\mathbf{y}$ vector of observations (measurements) of the physical system

$\mathbf{w}_i$ known conditions or settings associated with observation y_i

$\delta(\cdot)$ model discrepancy (or bias) term

ε_i observation (or measurement) error associated with field measurement y_i

$\zeta(\cdot)$ physical system

$\eta(\mathbf{w}_i, \boldsymbol{\theta})$ model outputs at conditions $\mathbf{w}_i$ and with calibration parameters $\boldsymbol{\theta}$

$\boldsymbol{\theta}$ vector of calibration parameters when their values result in a computer model that can effectively simulate the physical system

Chapter 4

A ratio, $h_c/(h_c+h_r)$, which depends on the surface heat transfer coefficient of the clothed body by convection (hc) and radiation (hr), used in the estimation of T_{op}

GHI_{mean} daily-mean global horizontal irradiance

SIT_{room} standardised indoor temperature for a given room

T_{air} air temperature

T_{od-1} daily mean outdoor ambient temperature of the previous day

T_{op} operative temperature

$T_{out,mean}$ daily-mean outdoor temperature

T_{max} maximum acceptable temperature

T_{rad} radiant temperature

T_{rm} exponentially weighted running mean of outdoor ambient temperature

T_{rm-1} exponentially weighted running mean of outdoor ambient temperature of the previous day

α constant used in the approximation of T_{rm}

β_{0-2} regression coefficients used in the standardisation of indoor temperatures

Chapter 5

AIC Akaike Information Criterion

AICc corrected Akaike Information Criterion

AIC_{min} minimum Akaike Information Criterion amongst candidate distributions

$f(x_i \mid \boldsymbol{\phi})$ probability of observing data point x_i given distribution parameters $\boldsymbol{\phi}$ for probability distribution function $f(\cdot \mid \cdot)$

i index for observations

j index for distributions

K number of distribution parameters

$\mathcal{L}(\boldsymbol{\phi} \mid \mathbf{x})$ likelihood of distributional parameter having a value $\boldsymbol{\phi}$ given the observed data $\mathbf{x}$

N number of observations

R number of candidate distributions

w_j Akaike weight for distribution j

$\mathbf{x}$ vector of observation $\mathbf{x} = x_{i=1}, \ldots, x_n$

Δ_j information loss (difference in AIC) when distribution j is selected instead of the best candidate distribution

Chapter 6

EE_{it} elementary effect for model input i in trajectory t

p total number of levels

r total number of trajectories

μ mean of elementary effects

μ^* mean of absolute elementary effects

σ standard deviation of elementary effects

Chapter 7

D number of daily values in the dataset ($D = D_c + D_v$)

D_c number of daily values used for calibration

D_v number of daily values used for validation

d index for day

$\mathbf{K}_z$ is the combined covariance matrix of the surrogate model and discrepancy term

$\mathcal{L}(\cdot \mid \cdot)$ likelihood function

L number for posterior samples

l index for posterior sample

M number of monitored dwellings in the homogeneous cluster

m index for monitored dwelling

MDLRT mean daytime living room temperature

$N_c^{(M)}$ number of monitored data points used for calibration ($N_c^{(M)} = M \times D_c$)

$N_c^{(S)}$ number of simulation data points used for calibration ($N_c^{(S)} = S \times D_c$)

N_c number of data points used for calibration ($N_c = N_c^{(M)} + N_c^{(S)}$)

$N_v^{(M)}$ number of monitored data points used for validation ($N_v^{(M)} = M \times D_v$)

$N_v^{(S)}$ number of simulation data points used for validation ($N_v^{(S)} = S \times D_v$)

S number of simulation runs

s index for simulated dwelling

$\mathbf{w}_d$ vector of weather variables associated with day d

$\mathbf{x}_m^{(M)}$ explanatory variables associated with monitored dwelling m

$\mathbf{x}_s^{(S)}$ explanatory variables associated with simulated dwelling s

$\mathbf{Y}_v^{(P)}$ $L \times D_v$ matrix of posterior predictions

$\mathbf{y}_c^{(M)}$ vector of monitored data used in the calibration. It consists of daily mean indoor temperatures from dwellings in homogeneous cluster

$\mathbf{y}_c^{(S)}$ vector of simulation data from archetype model. It consists of daily mean indoor temperatures

$\overline{y_d^{(M)}}$ the mean, averaged across M dwellings, monitored MDLRT for day = d

$\overline{y_v^{(M)}}$ the mean, averaged across M dwellings and D_v days, monitored MDLRT for the validation period

$\overline{y_d^{(P)}}$ the mean, averaged across $l = 500$ posterior MCMC draws, calibrated prediction of the MDLRT for day = d

$\mathbf{z}$ combined monitoring and simulation data vector used for calibration $\mathbf{z} = [\mathbf{y}_c^{(M)}, \mathbf{y}_c^{(S)}]$

$\delta(\cdot)$ is the discrepancy term (or model bias) represented by a Gaussian process

$\varepsilon_{md}^{(M)}$ Error term associated with the monitored data

$\varepsilon_{sd}^{(S)}$ Error term associated with the simulated data

$\eta(\cdot)$ is the surrogate model represented by a Gaussian process

$\boldsymbol{\theta}$ vector of calibration parameters when their values result in a computer model that can effectivelly simulate the physical system

$\boldsymbol{\mu}$ mean function of the covariance matrix defined as a vector of zeros

$\boldsymbol{\xi}$ represents the hyperparameters of the surrogate model

Chapter 1

Introduction

1.1 Research Context

1.1.1 Anthropogenic Climate Change

In their latest contribution to the Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report (AR6), Working Group I (WGI) concluded that human influence on warming the atmosphere, ocean and land is "unequivocal" (IPCC, 2021). The observed increase in well-mixed greenhouse gas (GHG) emissions since the 1750s is caused by human activities, and it is the main driver behind a global surface temperature increase of 1.09 [0.95 to 1.20][1] °C in 2011–2020 compared to 1850–1900. The scale of recent climatic changes is unprecedented, over many centuries to many thousands of years, with the increase in global surface temperature being faster since 1970 than in any other 50-year period in the last 2000 years. The changing climate has already caused widespread adverse impacts to nature and people, including, but not limited to, human mortality and morbidity due to extreme heat, increased flooding, climate-related food-borne and water-borne diseases, and the increased incidence of vector borne diseases (IPCC, 2022). Unless decisive mitigation and adaptation actions are taken now and in the near term (until 2040), the mid to long-term (2041-2100) impacts are expected to be multiple times higher than what has been observed (IPCC, 2022).

Mitigation actions aim to reduce emissions or enhance the sinks of GHG, while

[1] 90 % confidence interval

adaptation is the process of adjustment to the changing climate and its effects in order to moderate harm or exploit beneficial opportunities (IPCC, 2018). Global negotiations on mitigation culminated at the 21st United Nations Climate Change Conference of Parties (COP21) in the Paris Agreement, adopted by 195 nations in December 2015, whose central aim is to limit global temperature rise to well below 2 °C relative to pre-industrial levels and pursue efforts to limit warming to 1.5 °C (UNFCCC, 2015; IPCC, 2018). To track global action, the Paris Agreement requests countries to submit their *Nationally Determined Contributions* (NDCs), which outline each country's mitigation efforts, to the United Nations Framework Convention on Climate Change (UNFCCC) secretariat, and new or updated NDCs should be submitted every five years (UNFCCC, 2015). The 26th Conference of Parties (COP26) marked the end of the first five-year cycle. While a warming of just under 2 °C might be achieved if all ambitions announced at COP26 are materialised, climate policies at the time of writing would likely result in a temperature rise of 2.7 °C compared to pre-industrial levels (CCC, 2021a). Such a level of warming is expected to have *high* to *very high* impacts on terrestrial, freshwater and ocean ecosystems (IPCC, 2022).[2] Further, under any emission scenario considered by WGI in their AR6 contribution, global surface temperature is expected to continue increasing until at least mid-century (IPCC, 2021). As adverse impacts of climate change escalate with every increment of global warming (IPCC, 2022), and since further warming is unavoidable (IPCC, 2021), it is imperative that ambitious mitigation actions are coupled with strong adaptation efforts to minimise current and future impacts of climate change.

1.1.2 Climate Change and the United Kingdom

The United Kingdom (UK) is also experiencing the effects of climate change. Average UK land temperature has risen by around 1 °C since 1850–1900, and the ten warmest years since 1884 have all occurred since 2002 (Kendon et al., 2020; Met Office, 2021). 2020 was the third warmest, fifth wettest and eighth sunniest year on UK record; the only year to be in the top-10 for all three variables (Kendon et al., 2021).

[2] Please see IPCC (2022) for a detailed description of these impacts at a global and regional level.

The wettest February, April, June, November and December recorded in monthly series from 1862, have all occurred since 2009 (Kendon et al., 2021). Climate change is also increasing the frequency and severity of extreme events (IPCC, 2021). In August 2020, southern England experienced one of the most significant heatwaves of the last 60 years, while in February 2020 storms Ciara and Dennis, arriving a week apart, caused severe and widespread flooding across the UK (Kendon et al., 2021). In July 2022, a new record maximum temperature was recorded for Wales (37.1 °C), Scotland (34.8 °C) and England (and UK, 40.3 °C) (Met Office, 2022a).

By the end of the 21st century, all areas of the UK are projected to be warmer, with the level of warming depending on future emissions (Figure 1.1). The 50th percentile annual mean temperature anomaly for 2080–2099, compared to 1981–2000, is 1.4 °C for the lowest emissions Representative Concentration Pathway (RCP2.6), but 3.9 °C for RCP8.5, the highest emissions scenario (Lowe et al., 2018). The probability of observing summers as hot as 2018, which was the joint-warmest summer in the UK (together with 1976, 2003 and 2006) and had severe implications for several sectors, has already increased due to climate change and is projected to reach approximately 50 % by mid-century (CCC, 2021b; Met Office, 2021). Further, projections show a continued sea level rise until 2100 regardless of the emission pathway, an increase in average winter precipitation and a significant increase in hourly precipitation extremes (Met Office, 2021).

Recognising the need to decarbonise, the UK was the first major economy to commit, through the *The Climate Change Act 2008 (2050 Target Amendment) Order 2019*, to a legally-binding target of reaching net zero GHG emissions by 2050. The UK has also set an ambitious NDC target of reducing its GHG emissions by at least 68 % on 1990 levels by 2030 (HMG, 2021c). In addition to the NDC and in accordance to the 2008 Climate Change Act, the process of reducing GHG emissions in the UK is also guided by the Carbon Budgets that restrict UK's GHG emissions over five-year periods (HMG, 2021c). The sixth carbon budget, set in law by the UK government, requires that the UK GHG emissions should be reduced by 78 % compared to the 1990 levels by 2035 (HMG, 2021c). Achieving these targets

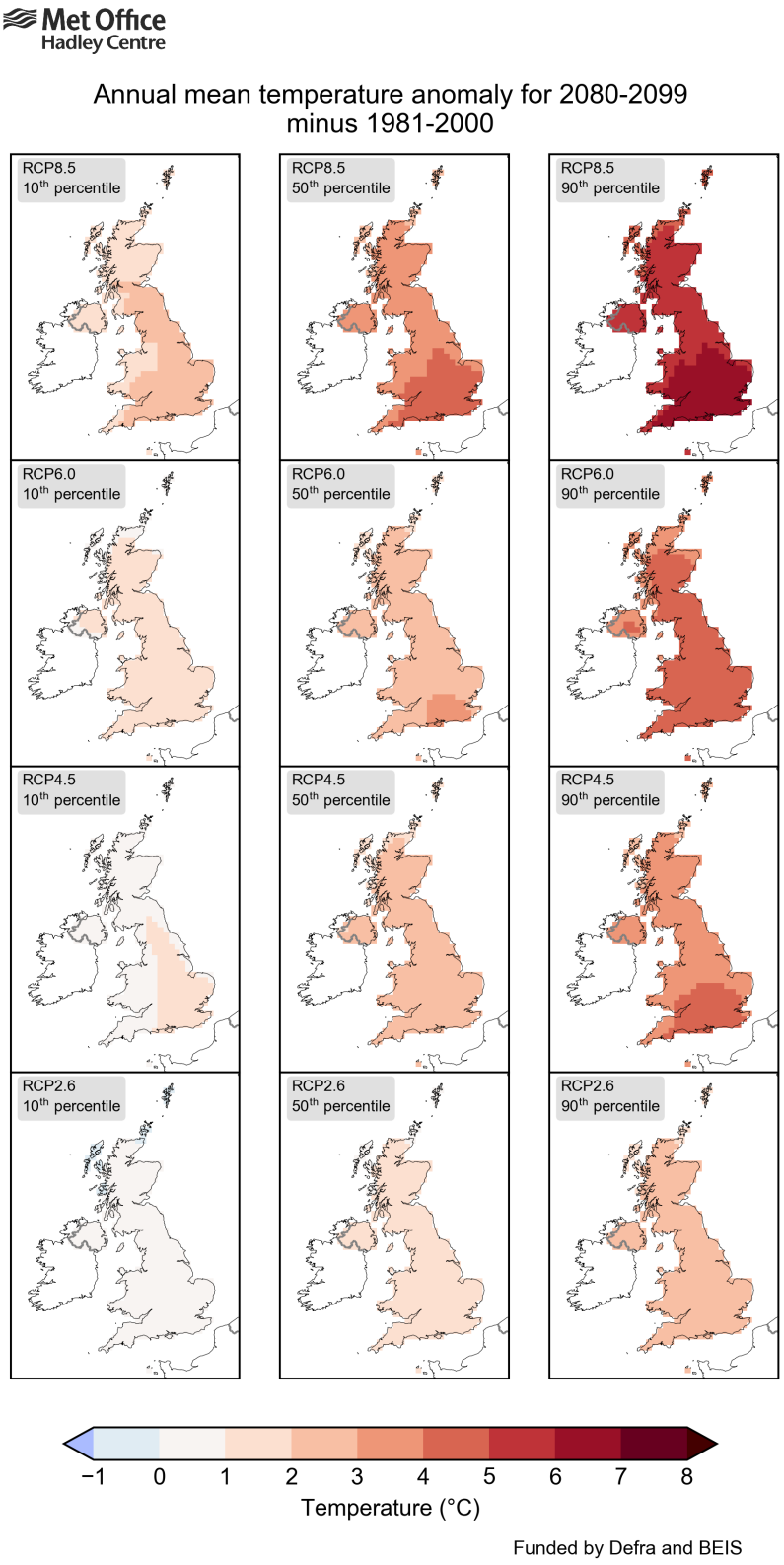


Figure 1.1: Probabilistic 20-year mean temperature anomalies over land based on the 2018 UK Climate Projections, generated for four Representative Concentration Pathways (RCP2.6–8.5) using the Met Office web tool (Met Office, 2018a).

requires action across all sectors, including housing, which was responsible for around a fifth of UK territorial CO_2 emissions in 2020 (BEIS, 2021a). Such actions may translate to urgent and large-scale changes in the housing stock. According to the UK government's Heat and Building Strategy, new homes should be built to a very high standard of thermal efficiency and airtightness, whereas in existing homes that fall below the government's energy standard a "fabric-first" approach will prioritise building envelope improvements (HMG, 2021b). In conjunction with improved thermal efficiency, the widespread adoption of heat pumps and the use of heat networks and hydrogen, where appropriate, will be responsible for the majority of GHG emissions reduction from homes (HMG, 2021c). While their contribution to the UK's mitigation efforts is important, especially for the near-term emissions reduction, buildings in general and housing in particular, also play an important part in adaptation.

1.2 Study Motivation

1.2.1 The Importance of Adaptation

In their Independent Assessment of UK Climate Risk, the Climate Change Committee (CCC) identified "risks to human health, wellbeing and productivity from increased exposure to heat in homes and other buildings" as one of eight risks with the highest priority for adaptation (CCC, 2021b). The UK's relatively mild winters and temperate summer conditions meant that *indoor overheating*, which qualitatively describes the state at which occupants feel uncomfortably warm due to the indoor environment (CIBSE, 2013), has not traditionally been a concern, and since 1965 efforts have instead focused on reducing heat loss in cold weather (Lomas and Porritt, 2017). While the UK housing stock remains thermally inefficient compared to neighbouring heating-dominated countries (ACE, 2015), increased levels of building thermal insulation and airtightness, along with the use of better-performing boilers facilitated the reduction of its GHG emissions by 22 % in 2015 compared to 1990s levels, despite a 25 % increase in the number of homes (BEIS, 2017).

Despite the improvements in the housing stock's thermal efficiency, fuel poverty

and the health implications of low indoor winter temperatures remain a major concern in the UK (Armstrong et al., 2018). At the time of writing, in 2022, the percentage of homes in fuel poverty may increase for the first time in more than a decade due to the steep increase in cost of fuel (Hinson and Bolton, 2022). However, a "perfect storm of interacting factors" has resulted in indoor overheating also being a major concern in the UK and other heat-dominated countries (Lomas and Porritt, 2017): a warming planet with more frequent and severe heatwaves, urbanization and the associated urban heat island (UHI) effect, an ageing population with little experience of warm weather and a housing stock adapted primarily to cold winter conditions have culminated to a growing risk of indoor overheating with wide-ranging implications. Cognitive performance, reduced productivity, sleep quality and overall quality of life are all thought to be adversely affected by indoor overheating (Lan et al., 2011; Okamoto-Mizuno and Mizuno, 2012), which is thought to occur in approximately 19 %[3] of existing homes (Lomas et al., 2021). Further, studies on morbidity and mortality during periods of high outdoor temperature suggest that such conditions contribute to significant excess mortality in England, and provide indirect evidence on the importance of indoor overheating as determinant of heat-health (Kovats and Brisley, 2021; WHO, 2018). A notable recent example was the summer of 2020, where the total cumulative all-causes excess mortality (taking out the effects of Covid-19) during its three heatwaves in England (2556 deaths) was comparable to the heatwaves of 2003 (2234 deaths) and 2006 (2323 deaths) (Kovats and Brisley, 2021). In the absence of adaptation, Hajat et al. (2014) estimated a 257 % increase in heat-related deaths by 2050, from an annual baseline of around 2000 excess deaths in the 2000s. While such forecasts are not without uncertainties, the trend of increasing heat-related mortality and other adverse consequences caused by the lack of adequate adaptation to a warming planet is clear.

To limit the impact of indoor overheating on the occupants' health and wellbeing, clear and strong policy is required. However, as the CCC noted in June 2021, "policies to address overheating risks in buildings are still missing despite it being

[3]Such estimates depend on multiple factors, including the method used to quantify indoor overheating risk.

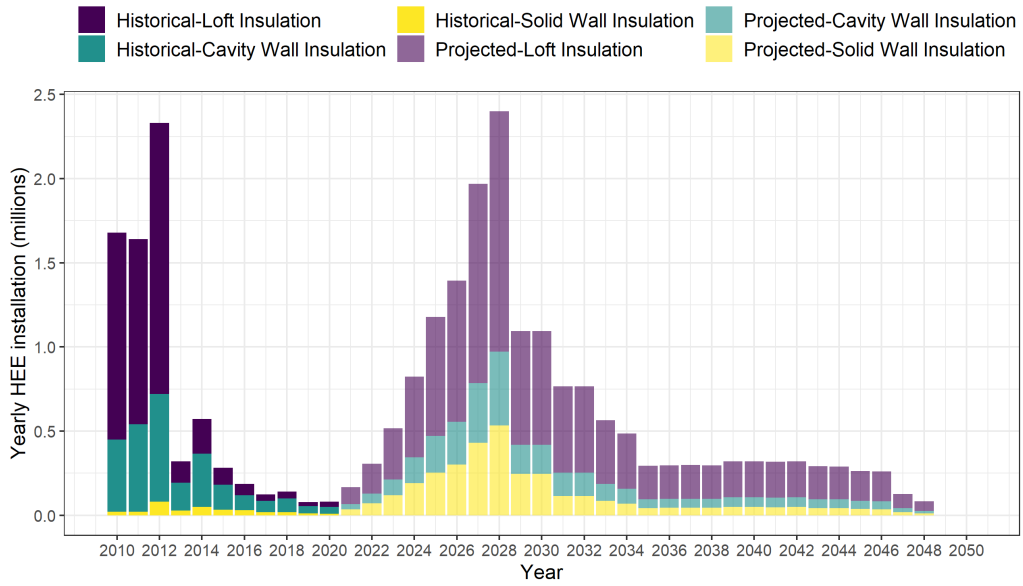


Figure 1.2: Historical (BEIS, 2021d; CCC, 2019) and projected (CCC, 2020) uptake of Home Energy Efficiency (HEE) measures.

one of the top risks in all UK climate risk assessments published to date" (CCC, 2021b). This changed in December 2021, when Part O Overheating was introduced as an amendment to the Building Regulations 2010 (*The Building Regulations Etc. (Amendment) (England) Regulations 2021*). Part O, which came into force on the 15th June 2022, signals a step forward in tackling indoor overheating. It requires that "reasonable provision" is made to limit unwanted solar gains in summer and provide adequate means to remove heat from the indoor environment of any building that contains at least one room for residential purposes, other than a room in a hotel[4]. Compliance with Part O can be demonstrated through the use of one of two methods described in Approved Document O (ADO) (HMG, 2021a): (1) a simplified method, and (2) a dynamic thermal modelling method which is largely based on Technical Memorandum 59 (TM59) released by the Chartered Institution of Building Services Engineers (CIBSE) (CIBSE, 2017).

While the Part O amendment and the release of ADO are positive steps in the effort to adapt against a warming climate (CCC, 2022), it is important to highlight that

[4]It is worth highlighting that Part O requires that in meeting this obligation, "account must be taken of the safety of any occupant, and their reasonable enjoyment of the residence; and mechanical cooling may only be used where insufficient heat is capable of being removed from the indoor environment without it." (*The Building Regulations Etc. (Amendment) (England) Regulations 2021*)

Part O, and ADO, apply only to *new* residential buildings (HMG, 2021a). While new homes are generally considered to be at a greater risk of indoor overheating, evidence suggests that a significant percentage of the existing housing stock already overheats (Lomas et al., 2021). Further, there are concerns that home energy efficiency (HEE) measures can, in some circumstances, exacerbate the risk of indoor overheating (Shrubsole et al., 2014; Taylor et al., 2021). Since the uptake of HEE measures is expected to increase due to climate change mitigation actions (Figure 1.2), and hot summers like 2018 are expected to further increase in frequency (Met Office, 2021), policies should aim to reduce indoor overheating risk in the existing housing stock whilst also reducing its environmental footprint.

1.2.2 Building Stock Models as Tools to Support Policymaking

Devising effective mitigation and adaptation policies is challenging. To identify solutions that offer the greatest benefit for the smallest cost, it is important to quantify each option's effectiveness and investigate any unintended consequences. This may be achieved through *modelling*; the process of creating an abstraction of a natural system governed by certain (often mathematical) rules, used to estimate an output of interest or as a way to understand the natural system better (Saltelli et al., 2008). As Oraiopoulos and Howard (2022) describe, models have been used to guide energy policy since the 1970s. Initially the capabilities of such models in representing buildings was limited, however, research in the decades that followed has resulted in their advancement and maturity. Building Performance Simulation (BPS) tools enable a mathematical model of a building and its systems to be constructed and simulated at a given level of abstraction (Mantesi et al., 2018). Several BPS tools exist with different capabilities[5], which often rely on a series of heat and mass transfer processes linked through a heat balance equation (Crawley et al., 2008). To construct and simulate a BPS model, numerous inputs must be provided, including: the building's geometry, thermophysical characteristics, electrical equipment, heating ventilation and air conditioning (HVAC) system, occupant presence and actions, and

[5]The terms "Building Energy Model (BEM)" or "Dynamic Thermal Simulation (DTS)" are also commonly used to describe BPS tools or models depending on their characteristics.

information on the local weather. Outputs from such simulations can include heating, cooling and electricity demand, indoor temperature and pollutant concentration. As it is not uncommon for policies to target thousands to millions of homes, that might share some similarities but will differ in many ways, there has been a growing interest in building stock modelling over the last decade (Oraiopoulos and Howard, 2022; Reinhart and Cerezo Davila, 2016).

Building stock modelling (BSM) is the development and implementation of mathematical representations of a group of buildings, used to investigate one or multiple quantities of interest. What constitutes as a *group* will depend on the application, and the scale typically ranges from buildings at a local or regional level to that of a national level (Reinhart and Cerezo Davila, 2016). One approach to BSM, referred to as "bottom-up engineering" or "bottom-up building physics", relies on the use of BPS tools to predict the quantity of interest for each member, or a representative sample, of the selected group of buildings (Kavgic et al., 2010; Lim and Zhai, 2017b). Thus, the same calculation engine (i.e. the BPS tool) might be used to model a single building or a large group of buildings. However, in the case of building stock modelling, a set of inputs must be defined for each building being modelled, and appropriate techniques are needed to analyse the outputs. Therefore, bottom-up buildings stock models require extensive databases of empirical data to support the detailed modelling of each building (Kavgic et al., 2010). Compiling these databases requires substantial time, effort, and financial investment, while it can also take a considerable amount of time to construct and simulate the models (Lim and Zhai, 2017b; Reinhart and Cerezo Davila, 2016). Since a complete dataset of the inputs needed to model each building in a stock is not available, some inputs are often inferred based on the building characteristics (e.g. typology and approximate age of construction) that are available (Mavrogianni et al., 2012; Taylor et al., 2015). Another commonly used simplification in building stock modelling relies on the concept of *archetypes*; building definitions that represent a group of dwellings with similar properties (Reinhart and Cerezo Davila, 2016). This method reduces the amount of data and model simulations required. Generating archetypes requires

the *classification* (or *segmentation*) of the building stock into *homogeneous* groups (i.e. groups of dwellings with similar properties), and the *characterisation* of each group, the specification of all model inputs associated with each group of dwellings (Reinhart and Cerezo Davila, 2016). One such model is the archetype-based UK Housing Stock Model (UK-HSM); a model developed by UCL researchers that uses EnergyPlus, an open-source BPS tool developed by the United States Department of Energy (DOE, 2016), as its calculation engine. A key aim of UK-HSM is to guide heat-related adaptation efforts in UK homes, and has so far been used to assess the impact of home energy efficiency retrofit and occupant actions on summer indoor temperatures (Mavrogianni et al., 2012; Mavrogianni et al., 2014) and for heat-related mortality modelling in London and the West Midlands (Taylor et al., 2015; Taylor et al., 2018b).

Archetype-based building (or housing) stock models can be an asset to policymakers, allowing for the consideration of multiple policy options on groups of dwellings representative of the building stock. However, the level of trust in their results should be limited until they are validated against empirical data and the associated uncertainties, which are at the core of any modelling process, are quantified and reduced. Uncertainties may arise from multiple sources, including (Kennedy and O'Hagan, 2001): a lack of knowledge about the true values of some model inputs (parameter uncertainty), the inherent variability of the system being modelled (residual variability), and the inadequacy of the model in fully representing the real system (model inadequacy). All aforementioned sources of uncertainty are expected to affect UK-HSM, and an empirical validation study by Symonds et al. (2017) sought to quantify its predictive performance against the indoor temperatures monitored in 823 dwellings during the 2011 Energy Follow-Up Survey (Hulme et al., 2013a). The study revealed that predictions were generally better when comparing aggregate groups of dwellings than in dwelling-by-dwelling comparison. For the group of semi-detached dwellings, the most frequently occurring building typology in England, the archetype-level Root Mean Square Error for the mean of daily maximum temperatures in the summer period ranged between 0.94 – 1.73 °C depending

on location, with the differences deemed statistically significant. To improve the predictive performance of UK-HSM, a method to quantify the different sources of uncertainty and reduce parameter uncertainties in archetype-based models is required; one such method that is popular within the building modelling community relies on Bayesian inference (Hou et al., 2021; Oraiopoulos and Howard, 2022).

1.2.3 A Bayesian Solution to the Problem of Uncertainties

The use of Bayesian inference for the calibration of computer models, introduced by Kennedy and O'Hagan (2001), has inspired the calibration of several archetype-based building stock models (Hou et al., 2021; Oraiopoulos and Howard, 2022). The key ingredients to this process are: (1) a model with at least one uncertain input and an output of interest, (2) empirical (also referred to as *field* or *monitored*) data of the output of interest, and (3) a set of probability distributions, referred to as *priors*, representing the modeller's assumptions of the possible values that the uncertain model inputs might take. Through the use of Bayes' Theorem, a set of updated distributions of the uncertain model inputs, referred to as *posteriors*, can be obtained.

The main benefits of Bayesian calibration over other approaches relate to its ability to not only improve predictive performance, but also quantify uncertainty and apportion it to different sources (Booth et al., 2012). Many other calibration procedures are deterministic, resulting in precise estimates of calibration parameters that might be inaccurate if the empirical data used for the calibration are insufficient or otherwise limited, and ignore uncertainties that naturally arise in modelling procedures (Hou et al., 2021).

A review by Hou et al. (2021) revealed that Bayesian calibration of building stock models has concentrated on energy-related outputs (for example, heat demand or electricity consumption).[6] An attempt at calibrating archetype-based models of indoor temperature using Bayesian inference could not be identified, despite the important contribution that such models can have in guiding adaptation policy. To undertake Bayesian calibration on archetype-based models, a necessary prerequisite

[6]The only exception in the Hou et al. (2021) review was the work of Braulio-Gonzalo et al. (2016) who looked at discomfort hours, in addition to energy demand. However, this work did not involve any field data and relied only on simulation data.

is the classification of the housing stock into homogeneous groups of dwellings in order to accurately apportion uncertainty. Despite the importance of this step, it is often not discussed in detail, nor is it clear how the classification performed takes into consideration the calibration process (see Section 2.4.2). Another crucial step in the Bayesian calibration process is the choice of priors, especially when field data are limited. In some cases, priors are defined as uniform distributions, which are unlikely to represent the existing knowledge about the model inputs (see Section 2.4.5.1). Where non-uniform distributions are assigned, there is often little discussion as to why these distributions were selected.

1.3 Research Aim, Objectives and Scope

As argued in Section 1.2, there is an urgent need to adapt the housing stock in ways that safeguard occupants from the risks to health, wellbeing and productivity associated with elevated ambient temperature, and more frequent and severe heatwaves. Archetype-based housing stock models of indoor temperature can support such efforts, by enabling policymakers to evaluate and compare the impact of different interventions. However, uncertainties are at the core of any modelling process, and it is crucial for such uncertainties to be quantified and minimised in order to allow for well-informed decision-making. Yet, efforts in calibrating such models are lacking. This research gap has motivated this doctoral study, whose aim is:

> *to quantify and reduce uncertainties of archetype-based housing stock models of summer indoor temperature.*

To achieve this aim, and by examining published work on model calibration, three research objectives have been identified:

1. To develop a Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature;

2. To quantify the uncertainty of the UK Housing Stock Model inputs with the greatest influence on summer indoor temperature for a homogeneous group of dwellings;

3. To quantify the level of improvement in the predictive ability of the UK Housing Stock Model following application of the Bayesian calibration framework and reduce model input uncertainty for a homogeneous group of dwellings.

In response to the first research objective, a Bayesian calibration for archetype-based housing stock models of indoor temperature will be introduced. The focus on an archetype approach is due to the lack of detailed model input data that make it challenging to model individual buildings within the housing stock. With increasing data availability and computational resources, the prevalence of the individual building modelling approach of BSM is expected to grow, yet the archetype-based approach remains a key method within the field, with many recent examples utilising it (e.g. Tardioli et al., 2020; Wang et al., 2020). Advancements made within this thesis are expected to contribute to the Bayesian calibration of building stock models that take an individual building and archetype-based approach. The multistep framework will consider classification and characterisation in addition to model calibration. The process will rely on a definition of homogeneity that will be provided together with the proposed framework. It is expected that the framework could be easily adapted for most archetype-based models of indoor temperature. For this thesis, however, the framework's application will focus on UK-HSM, which acts as a case study for the proposed framework.

To achieve the second research objective, the framework's classification and characterisation steps will be applied to UK-HSM. Statistical analysis of the 2011 English Housing Survey and Energy Follow-Up Survey will be considered together with the UK-HSM model structure to identify a homogeneous group of dwellings (Hulme et al., 2013a). Model input uncertainty will be quantified using empirically-based probability distributions, where possible. Sensitivity analysis will be used to select the model inputs with the greatest influence on summer indoor temperature.

Having quantified UK-HSM model input uncertainty for a homogeneous group of dwellings in the second research objective, the third research objective aims to reduce this uncertainty using Bayesian calibration and quantify the improvement in predictive ability. The process will be informed by the probability distributions

of the most influential variables identified in the second research objective, and the empirical data of summer indoor temperature collected during the 4M project in Leicester in 2009 (Lomas and Kane, 2013). The data will be split into a training and validation set, and the improvement in out-of-sample prediction will be quantified using commonly used validation metrics.

1.4 Novel Contributions

The novel contributions of this thesis are:

1. A modular Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature that relies on a practical definition of homogeneity. The framework explicitly describes the classification and prior elicitation steps in detail.

2. The first application of Bayesian calibration on an archetype-based model of summer indoor temperature. Lessons learned from this process, such as the importance of outdoor temperature lag components, can inform research in the field of indoor temperature modelling.

3. The study of associations between dwelling and household characteristics and standardised summer indoor temperature using large-scale empirical data, published as a paper (Petrou et al., 2019b). This provides further evidence to the field of indoor overheating on the factors associated with high indoor temperature.

4. The introduction of an innovative method for identifying suitable probability distributions from empirical data. This technique can find many uses within the building modelling field, including the characterisation of model inputs or the analysis of model output, improving existing modelling practices. A paper provides guidance on the method's application, while publicly available R code facilitates its adoption by the wider modelling community (Petrou et al., 2021b).

5. An advancement in the implementation of Gaussian processes as surrogate

models in the process of Bayesian calibration that can greatly reduce computational cost. This alternative approach can allow calibration to be implemented even if computational resources are relatively limited. In addition, it may improve current metamodelling practices by enabling the use of more training data, where previously the processing time would have been prohibitive.

1.5 Structure of the Thesis

This thesis consists of eight main chapters, including this introductory chapter (Figure 1.3). Following the main body, appendices are available and provide supporting information, a more detailed coverage of the relevant theory and additional results. The following paragraphs provide an outline of each chapter.

Chapter 2 begins with a review on the topic of indoor overheating in residential settings in temperate climates, placing emphasis on the UK. The review covers the main causes of indoor overheating, its implications on occupants' health and wellbeing, along with ways for quantifying and reducing the overheating risk. The chapter continues with an overview of building stock modelling, and explains why uncertainties are an intrinsic component of the modelling process. Approaches to model calibration are described and compared, and the theory of Bayesian inference for model calibration is introduced. Published research on the Bayesian calibration of archetype-based housing stock models is reviewed, and the chapter concludes with a summary of the key research gaps.

In Chapter 3, the five-step Bayesian calibration framework is introduced, together with a definition of homogeneity specific to this work. This is in response to the first research objective set out in Section 1.3 and relies on theory and literature covered in Chapter 2. The purpose and importance of each step is explained, while a more detailed description of the methods associated with each step is included in Chapters 4–7. The two main datasets used within this study are described, and modelling details regarding UK-HSM are provided.

In Chapter 4, a group of dwellings assumed to be homogeneous is identified by applying Steps 1 and 2 of the Bayesian calibration framework – whether further

segmentation of the group is required is assessed in Step 4 (described in Chapter 6). Chapter 4 begins with a detailed description of the methods relating to the framework's first two steps, followed by the relevant results. In Step 1, a statistical analysis of the 2011 EFUS is carried out and variables that are significantly associated with summer indoor temperatures are identified. In Step 2, a subset of the statistically significant variables are used as classifiers to select a group of dwellings from the 4M dataset suspected to be homogeneous. The chapter concludes with a discussion on limitations and a summary.

Chapter 5 focuses on Step 3 of the Bayesian calibration framework. This step requires the identification of appropriate probability distributions for each continuous model input of UK-HSM. Methods for accomplishing this task are first detailed, before their application is presented. The findings from this component of the work and its limitations are discussed.

Chapter 6 covers Step 4 of the Bayesian calibration framework, the sensitivity analysis, that aimed to identify influential model inputs to be calibrated. In addition, through this process, it is determined whether the group of dwellings suspected to be homogeneous following Step 2 (Chapter 4) should be further segmented. The limitations associated with this process are discussed. Jointly, Chapters 4–6 address the second research objective identified in Section 1.3.

Chapter 7 details the calibration of UK-HSM using the 4M data for a homogeneous group of dwellings. The chapter begins with a description of methods specific to this component of the work. Details regarding the parametric calibration experiment that was conducted are provided, and the findings are presented and discussed. The results from this chapter include a quantified estimate of the improvement in out-of-sample predictive performance, and the reduction of model input uncertainties; thus, addressing the third research objective. The limitations of this work are also discussed.

Chapter 8, begins by providing a summary of the key conclusions of this work in Section 8.1. Section 8.2 revisits the research aim and objectives set out in Section 1.3 and evaluates whether they have been achieved. The novel contributions

and limitations of this thesis are discussed, and an outline of future work that could follow on from this study is presented.

Further information that might be of interest to the readers on topics such as the mathematical background of Gaussian Processes, floor plans and schedules of the UK-HSM archetypes, or additional results are provided within the Appendices (Chapters A–F).

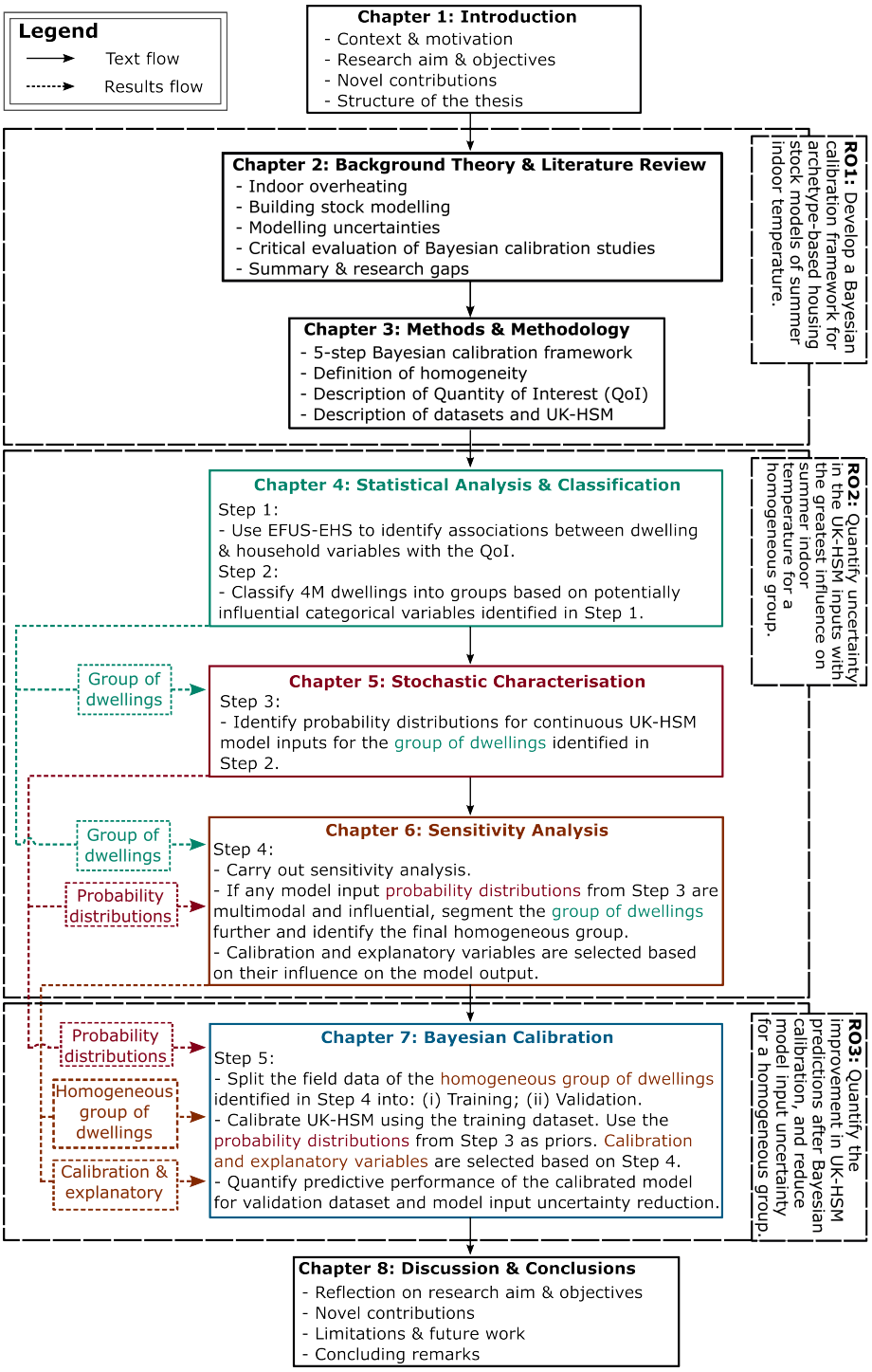


Figure 1.3: Flowchart depicting the text and information flow of the thesis. RO1-3 are shortened versions of Research Objective 1-3. UK-HSM stands for UK Housing Stock Model, QoI for Quantity of Interest, EFUS for Energy Follow-Up Survey and EHS for English Housing Survey.

Chapter 2

Background Theory and Literature Review

The previous chapter established the link between *climate change*, driven by the anthropogenic increase in greenhouse gases (IPCC, 2021), and *indoor overheating*, qualitatively defined as the state at which occupants feel uncomfortably warm due to the indoor environment (CIBSE, 2013). The importance of indoor overheating as an adaptation risk in the United Kingdom (UK) was highlighted (CCC, 2021b), and its implications for occupant health and wellbeing were briefly discussed. As argued, while the introduction of Approved Document O is a step in the right direction in tackling indoor overheating in new homes, there is currently a gap in policy for existing homes (CCC, 2022). A valuable tool for policymakers, that could inform policies on the adaptation of existing homes and enhance policies on new homes, is *building stock modelling*; an approach with a long history of guiding energy policy (Oraiopoulos and Howard, 2022). An established model that has been used to assess the impact of climate change, home energy efficiency and adaptation on indoor overheating and health is the archetype-based UK Housing Stock Model (UK-HSM) (Taylor et al., 2021). Validation work on UK-HSM, that recognised the need for uncertainty quantification and calibration (Symonds et al., 2017), and the lack of such work on archetype models of summer indoor temperature have motivated this research (Section 1.2).

The purpose of this chapter is to provide the theoretical foundation that will

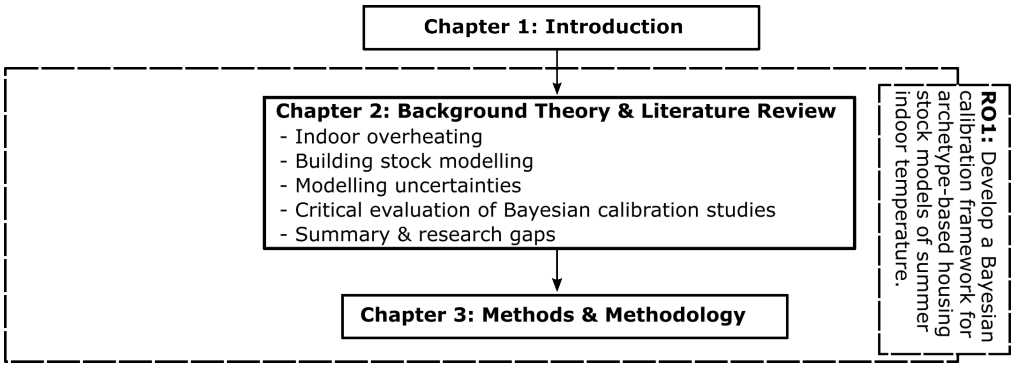


Figure 2.1: Chapter 2 flowchart. This is an abridged version of Figure 1.3, listing the five sections of Chapter 2. RO1 is a shortened version of Research Objective 1.

be used to address the research aim and objectives outlined in Section 1.3. This will be achieved through a critical review of the relevant literature, and a detailed consideration of the theory governing model uncertainties and calibration. It consists of four core sections and concludes with a summary of the literature discussed, and the key research gaps that this thesis aims to address (Figure 2.1).

In the first segment, the topic of indoor overheating is explored in more detail (Section 2.1). The causes of indoor overheating, are first considered, followed by a description of its implications on occupants and society. In the last part of this segment, established approaches for assessing thermal discomfort and quantifying excess heat-related mortality are discussed.

The second segment of this chapter focuses on building stock modelling (Section 2.2). Following a brief introduction to the various approaches of building stock modelling, the steps for archetype-based model development are discussed in more detail. The segment concludes with a review of the literature describing the development, application and validation of the archetype-based UK-HSM.

Building on the previous section, the third segment focuses on the topic of modelling uncertainties (Section 2.3). The segment begins with a discussion on why modelling and uncertainties are inseparable, before the various classes and sources of uncertainty are defined. Examples of where uncertainties can arise during building stock modelling are provided, and methods for quantifying and reducing them are introduced. One such approach is *model calibration*. Following a brief description

of the various approaches to calibration, emphasis is placed on statistical – more specifically *Bayesian* – calibration, for reasons discussed in Sections 2.3.5–2.3.6. The segment concludes with the key steps in Bayesian calibration.

In the fourth part of this chapter, a critical evaluation of published studies on the Bayesian calibration of archetype-based housing stock models is presented (Section 2.4). Eight studies are compared and contrasted according to thirteen categories, with similarities, strengths and weaknesses discussed. A common denominator across all studies is the focus on energy-related outputs.

For the first three parts, the literature was identified through an iterative process and over several years that largely relied on the use of appropriate keywords in *Google Scholar* to identify relevant studies (*pearl growing*), and examining publications that were either cited in or cited by them (*snowballing*). To identify the studies on the Bayesian calibration for archetype-based housing stock models discussed in the fourth segment of this chapter, a review was carried out in April 2018. Boolean searching was used in "ScienceDirect", with the terms "Building AND Calibration AND Simulation" in titles, abstracts and keywords. Following a review of titles and abstracts, and the application of the snowballing technique, four relevant studies were identified. Since then, another four studies were added based on monitoring of the literature and the comprehensive reviews by Hou et al. (2021) and Oraiopoulos and Howard (2022).

2.1 Indoor Overheating

Part of the challenge in mitigating indoor overheating is its complex nature; it is a sociotechnical problem that may be influenced by several physical parameters, human behaviour and the occupants' perception of the thermal environment (Kougionis, 2018). Under the same conditions, satisfaction with the indoor environment can vary between occupants (CIBSE, 2013), making it difficult to accurately quantify and minimise overheating risk. Progress in tackling this challenge has been made, and the following sections summarise the current state of knowledge with regard to the causes, implications and approaches to quantifying indoor overheating risk.

2.1.1 Causes of Indoor Overheating

While the increase in indoor overheating risk in UK homes is largely driven by the warming effect of greenhouse gas emissions, many other factors are thought to further exacerbate it. The naturally occurring spatial variation of the climate contributes to a greater overheating risk potential for homes located in the southern regions of the UK (Taylor et al., 2014). Another spatial variation in overheating risk potential arises from the impact of the local urban form on a home's microclimate. Ambient temperatures are higher in urban areas than in suburban or rural areas due to the use of materials that increase the absorption and retention of heat (e.g. concrete) or the creation of dense urban street canyons that trap solar radiation and waste heat from anthropogenic activities; a phenomenon referred to as the *Urban Heat Island* (UHI) effect (Mavrogianni et al., 2010; Heaviside et al., 2017). A home's overheating risk potential can also be influenced by its interaction with neighbouring buildings, trees and other vegetation that may provide shade from the sun but also form a wind barrier (Pisello et al., 2018).

Beyond the impact of local environmental characteristics, building design itself can increase or decrease the risk of indoor overheating; when heat gains from the outdoor or indoor environment are greater than heat losses through the building envelope, the accumulated heat results in a rise of indoor temperature. Building characteristics can alter this heat balance, modifying the risk of indoor overheating. The level of overheating risk is known to vary between building typologies, with top floor flats, mid-terraced dwellings and bungalows shown to overheat more (Lomas and Kane, 2013; Taylor et al., 2016). However, variations within a dwelling type can be greater than between dwelling types and may be the result of factors such as orientation, glazing ratio, fabric thermal resistance and thermal mass (Mavrogianni et al., 2012).

Transmission of solar radiation through non-opaque surfaces can elevate indoor overheating risk, especially for homes with highly-glazed and unshaded façades facing south, south-west or west (ZCH, 2015). The heat conducted through opaque surfaces can also be important, and may be reduced through increasing the albedo or

thermal resistance of external surfaces. Evidence suggests that roof insulation can reduce overheating risk, especially for the top floor rooms (Mavrogianni et al., 2012; Taylor et al., 2014). The effect of wall insulation is less clear as studies suggest an increase, decrease or no significant difference on indoor overheating risk, depending on where the insulation is placed, on whether purge ventilation is used and on the building's location (Fosas et al., 2018; Lomas et al., 2021; Mavrogianni et al., 2012; Peacock et al., 2010). Heavyweight constructions are associated with a lower risk of indoor overheating, especially when accompanied by night cooling, due to the dampening effect of thermal mass on the indoor temperature profile (Peacock et al., 2010; Hacker et al., 2008).

Ventilation provision can also be crucial. For mechanically ventilated buildings it is important to ensure that the design, installation and operation of ventilation systems maintains good air quality but also reduces indoor temperatures where necessary (McLeod and Swainson, 2017). For naturally ventilated homes, a building design that can only provide single-sided ventilation or offers limited window opening area will be more prone to indoor overheating (McLeod and Swainson, 2017; Petrou et al., 2019a).

Occupant behaviour is fundamental to how a building performs, and it is crucial to understand what factors may influence the occupants' actions in order to design and retrofit homes in a manner that limits indoor overheating risk. For example, even where there is – in theory – capacity for effective ventilation, factors such as noise, poor outdoor air quality, lack of mobility or security concerns can all result in a less than optimal use of windows from an overheating risk perspective (Mavrogianni et al., 2016). High levels of heat generated indoors (internal gains) can also have a significant contribution to indoor overheating risk (McLeod and Swainson, 2017; ZCH, 2015). Internal gains are generated from common activities such as cooking or using electrical equipment (ZCH, 2015). In some cases, large internal gains may arise from a poor building design, for instance as a by-product of poorly insulated hot water pipes.

2.1.2 Implications of Indoor Overheating

Physiologically, the human body tries to maintain a core body temperature of approximately 37 °C through the process of thermoregulation (Bouchama and Knochel, 2002; WHO, 2009). Heat is gained from the environment, produced by the metabolism, and dissipated by radiation, convection, conduction and perspiration (WHO, 2009). When exposed to elevated temperature, the human body may experience *heat stress*, defined as the "perceived discomfort and physiological strain associated with exposure to a hot environment" (Bouchama and Knochel, 2002). If the level, or the rate, of heat stress experienced is greater than the human body can adapt to during the exposure period, then thermoregulatory failure can take place. Heat would no longer be dissipated effectively, leading to heat exhaustion, mild-to-moderate illness with numerous symptoms, and may eventually develop into a heat stroke (Bouchama and Knochel, 2002). Heat stroke is characterised by a core body temperature greater than 40 °C with possible impairments to the central nervous system that may be fatal if left untreated. Since evidence suggests that people in several countries spend on average 56-66 % of their time at home, and more than 90 % indoors (Schweizer et al., 2007), exposure to heat indoors contributes a major part of overall exposure.

The relative risk of heat-related mortality has been shown to increase with ambient temperature above a threshold, and is thus greatest during periods of extreme heat (Armstrong et al., 2010). An example where such deadly consequences occurred at a large scale was the 2003 European heatwave, which resulted in 70,000 additional deaths compared to the reference period of 1998-2002 amongst 16 European countries (Robine et al., 2008). Over a 10-day period, an excess of more than 2000 deaths were reported within England and Wales during the 2003 heatwave (Johnson et al., 2005). During the same period, 15,000 excess deaths were reported in France with roughly 50% resulting from the domestic sector as the mortality rate increased by 74% in homes and 91% in care homes (Fouillet et al., 2006). Importantly, a harvesting effect (where the mortality rate decreases following an event) that would compensate for the increased mortality rate over summer was not observed in most countries (Robine et al., 2008). Therefore, it is possible that preventable deaths

occurred that might have been avoided if the right actions were taken. Since the 2003 heatwave, several other major episodes resulted in significant excess deaths in England, including the heatwaves in 2006 (2323 deaths), 2018 (863 deaths), 2019 (892 deaths) and 2020 (2,556 deaths, taking out the effects of Covid-19) (Kovats and Brisley, 2021). As may be noted, more excess deaths were recorded during the 2020 heatwaves than in previous years, with the cumulative excess all-cause mortality being the highest recorded since the introduction of the Heatwave Plan for England in 2004 (Kovats and Brisley, 2021). Many factors might have contributed to the comparatively high number of excess deaths, and it is unclear whether the increased time spent at home due to the Covid-19 pandemic played a part. It is worth highlighting that excess heat-related deaths are thought to occur even in periods not classified as heatwaves,[1] and their cumulative mortality burden might in some cases be greater than that of the heatwave period (Hajat et al., 2006). Furthermore, in the absence of adaptation and with the projected increase in global temperature, the problem of heat-related mortality is expected to grow in magnitude. Hajat et al. (2014) estimated a 257 % increase in heat-related deaths by 2050 compared to a baseline of around 2000 deaths in the 2000s has been estimated.

When levels of heat stress are not life-threatening, the thermal discomfort resulting from indoor overheating can still cause several adverse impacts to individuals and society. Evidence of reduced cognitive performance due to thermal discomfort suggests that domestic indoor overheating could impair productivity when working from home (Lan et al., 2011). The thermal environment is also thought to be a key determinant of nocturnal sleep quality, with deviations from the thermoneutral temperature range impacting the duration and onset of Rapid-Eye Movement (REM) and Slow-Wave Sleep (SWS) (Joshi et al., 2016).[2] As the SWS and REM stages are essential to physical recovery, memory consolidation and learning (Lan et al., 2017),

[1] A *heatwave* is defined by the Met Office as "an extended period of hot weather relative to the expected conditions of the area at that time of year, which may be accompanied by high humidity" (Met Office, 2022b). In London, the threshold is 28 °C, but the relative risk of heat-mortality is thought to increase above a two-day mean max temperature of 24.7 °C (Armstrong et al., 2010).

[2] The thermoneutral temperature range was defined by Joshi et al. (2016) as the range of ambient temperatures over which the body is not required to make any effort (through the process of thermoregulation) to maintain thermal homeostasis.

poor sleep quality is believed to reduce productivity in the workplace regardless of its thermal environment (AECOM, 2019). Poor sleep quality may also lead to increased risk of accidents, poor mental and physical health – including cardiovascular disease and reduced ability to maintain a healthy immune system – and overall poor quality of life (ZCH, 2015).

Beyond the impact on individuals and their families, indoor overheating through its increase in mortality, morbidity, risk of accidents and reduced productivity has an economic cost to the society (AECOM, 2019). Policies that aim to mitigate overheating risk could, thus, not only protect individuals but limit such costs. In addition, the implementation of such policies will likely lessen the growth in cooling demand that poses a threat to the UK's net zero ambitions (BEIS, 2021b). An important step in the process of devising and refining policies that may bring such benefits is the evaluation of indoor overheating risk prevalence, at present and under different future scenarios.

2.1.3 Quantifying Indoor Overheating

Within built environment research and practice, it is more common to determine whether a building's indoor environment is likely to result in thermal discomfort, than if it is likely to increase the risk of mortality. Since the methods and theory associated with each task are different, they are discussed separately in the following sections.

2.1.3.1 Assessing Thermal Discomfort

A direct way to identify whether a building's occupants are thermally comfortable is to ask them (CIBSE, 2013). *Post Occupancy Evaluation* (POE) is an approach that aims to evaluate the building's performance through feedback from the occupants, and may be used to investigate their experience of the thermal environment and other relevant qualities (CIBSE, 2013; CIBSE, 2020). While the POE method can offer invaluable insights, over the last few decades researchers have worked on the development of thermal comfort models that may be used to quantify thermal

discomfort using environmental parameters that may be monitored or modelled.[3]

For decades following its invention in 1970, the dominant model of thermal comfort was Fanger's Predicted Mean Vote (PMV) and Predicted Percent Dissatisfied (PPD) (de Dear et al., 2013). The PMV/PPD model, developed using measurements from climate chamber experiments and relying on a thermal balance equation, aims to predict the mean comfort vote that would be cast by a group of subjects given a set of clothing level and environmental conditions, and the proportion of people that would be dissatisfied (CIBSE, 2013). While still in use today for mechanically conditioned buildings, de Dear et al. (2013) noted that a "paradigm shift" began around the mid-1990s away from the physically-based PMV/PPD model and towards adaptive comfort models, at least for naturally conditioned buildings. Such approaches utilise field surveys of thermal comfort and concurrent monitoring of indoor and outdoor environmental parameters to develop statistical relationships between the acceptable indoor temperature and a variable derived from the outdoor temperature (CIBSE, 2013). An adaptive comfort model forms the basis for the assessment of thermal comfort in naturally conditioned buildings both in ANSI/ASHRAE Standard 55, and in BS EN 15251 (CIBSE, 2013).

As discussed by Lomas et al. (2021), the assessment of building overheating in the 1990s and for the following two decades focused on the use of static criteria which were also incorporated in CIBSE Guide A. With *static criteria*, the threshold temperatures above which overheating is deemed to occur do not change with ambient temperature (Lomas and Porritt, 2017). In 2013, CIBSE published Technical Memorandum 52 (TM52) where it was argued that "the assumption that there is a single indoor temperature limit irrespective of outdoor conditions is no longer considered sufficient", and proposed three criteria for assessing indoor overheating in naturally conditioned buildings based on BS EN 15251 (CIBSE, 2013). The more recent Technical Memorandum 59 (TM59) provides guidance on the assessment of indoor overheating risk in new homes (CIBSE, 2017), and a method that is largely

[3]Contrary to the POE approach of identifying indoor overheating risk, thermal comfort models enable the prediction of thermal discomfort using building physics models. With the advancement and more widespread application of machine learning techniques, it might be possible to predict indoor overheating risk using statistical models trained on POE data.

based on TM59 is one of the two ways that compliance may be demonstrated with Part O of the Building Regulations, according to Approved Document O (HMG, 2021a). For naturally ventilated homes, TM59 defines two criteria, and failure to satisfy either of them indicates a high risk of indoor overheating (CIBSE, 2017).

The first criterion is based on the European standard BS EN 15251 (BSI, 2007), which has now been superseded by 16798-1:2019 (BSI, 2019), and stems from the adaptive principle (Nicol and Humphreys, 2002) that people react in ways which tend to restore their comfort. The field data underpinning this equation were collected in 26 European offices, under the EU Project Smart controls and Thermal Comfort (SCATs) project (McCartney and Nicol, 2002; Nicol and Humphreys, 2010). The applicability of a thermal comfort model derived from office based data in the domestic context may be questioned, and it has been suggested that factors other than temperature could also influence the level of comfort and tolerance to the thermal environment (Brotas and Nicol, 2017). The second criterion is static in nature and was derived from research on sleep quality, published in 1979 and discussed in CIBSE in Guide A (CIBSE, 2015). A static criterion was likely preferred due to the assumption that there is limited capacity to adapt overnight. Nicol and Humphreys (2018) argued that adaptation can take place over a wide range of indoor temperature through a change in duvet and sleepwear, and the threshold of 26 °C used by TM59 is potentially 3–5 °C lower than the maximum comfort temperature for nude sleeping people. Therefore, a threshold of 29–32 °C may be more appropriate, although further research is required to support this (Nicol and Humphreys, 2018).

2.1.3.2 Quantifying Excess Heat-Related Mortality

Several epidemiological studies have sought to examine and quantify the association of high ambient temperatures with excess mortality (Basu and Samet, 2002), how this relationship may vary within countries (Armstrong et al., 2010), between countries (Gasparrini et al., 2015), and how it might be influenced by climate change (Vicedo-Cabrera et al., 2021).

In one such study, Armstrong et al. (2010) used ecological time-series regression between daily counts of all-cause mortality and predictors based on daily ambient

temperature between the summers of 1993 to 2006 in 10 government regions in England and Wales. The analysis quantified a threshold, the 93rd centile of the two-day mean max ambient temperature, over which there is an increase in mortality with increase in temperatures in all regions except the North-East. The thresholds ranged from 21.6 °C in Wales to 24.7 °C in London, while the mean increase in relative risks ranged from 1.3 %/°C in the North West to 3.8 %/°C in London. While this large scale analysis of over 2.286 million deaths provides strong evidence for the association of mortality with elevated summer temperatures, such clear and adequate evidence of the association between indoor temperatures and health effects does not exist (Anderson et al., 2013; WHO, 2018).

To link heat-related mortality with building characteristics in the absence of empirical data, Taylor et al. (2015) assumed that on days which exceeded the external temperature threshold, the temperature-mortality relationship was the same for indoor temperatures as for outdoor temperatures. For example, if in London the threshold was exceeded by 1 °C resulting in a heat-related increase in relative risk of 3.8 %, under the same outdoor conditions, a change in the building characteristics that would result in a 1 °C increase in the average maximum indoor temperature of London homes would be associated with a further 3.8 % increase in relative risk. Thus, the overall increase in relative risk of 7.6 %. Taylor et al. (2015) did not have any empirical evidence for the use of indoor daily maximum temperature, but in the absence of evidence this seems like a reasonable assumption. In other epidemiological studies (e.g. Hajat et al., 2014), mortality was associated with daily mean outdoor temperatures, therefore a case for the use of daily indoor mean temperatures can equally be made.

2.2 Building Stock Modelling

Modelling, the process of creating an abstraction of a natural system, enables the estimation of quantities that might have been hard, or impossible to otherwise determine, and it can also be used as a way to understand the natural system better (Saltelli et al., 2008). Since 2017, designers of new homes have been encouraged and

guided by CIBSE TM59 to carry out indoor overheating risk modelling to identify and avoid building designs likely to result in thermal discomfort (CIBSE, 2017). The value of modelling in limiting indoor overheating risk is also recognised by the English government's guidance for compliance with the building regulations; according to Approved Document O, modelling in the form of CIBSE TM59 is one of the two ways that compliance with Part O of the 2021 amendment to the building regulations may be demonstrated (HMG, 2021a). While Part O is limited to new homes, the prevalence of indoor overheating is not; approximately 19 % of the existing housing stock was found to overheat in 2018 (Lomas et al., 2021). Thus, further policies that mitigate overheating risk in existing homes are required (CCC, 2022). The formulation of such policies can be supported by building stock models whose value in guiding energy-related policies has been recognised (Kavgic et al., 2010; Oraiopoulos and Howard, 2022).

2.2.1 Approaches to Building Stock Modelling

Building stock modelling often falls within one of two classes: (i) *top-down* and (ii) *bottom-up* (Kavgic et al., 2010). A top-down approach will typically employ statistical techniques on aggregate building stock data to try to infer relationships between the quantity of interest (e.g. energy consumption or CO_2 emissions) and other variables. Bottom-up methods work at the disaggregated level, with data combined from a hierarchy of individual components. Bottom-up approaches can further be separated into *statistical* and *engineering*, depending on the modelling methods used (Lim and Zhai, 2017b). Whereas a statistical approach would rely purely on historical data to gain an understanding of the building being studied and forecast future trends, engineering methods rely on established building physics theory. An engineering method will require detailed input information that would often necessitate several assumptions. On the contrary, statistical methods may be employed faster, provided historical data are available, and will likely require less assumptions. A key weakness of statistical methods is their inability to make accurate predictions outside their training domain and to effectively assess the impact of new technologies, building interventions and weather forecasts for which historical data

does not exist. This limits their usefulness as policy design tools, with bottom-up engineering models often preferred for such applications (Booth et al., 2012; Cerezo et al., 2017; Lim and Zhai, 2017b).

Dynamic Thermal Simulation (DTS – also referred to as Building Performance Simulation or Building Energy Models) software are engineering (or building physics) tools that have long been used to predict the energy demand of buildings, assess their indoor environment and compare different retrofit options (Lim and Zhai, 2017b). Typically, such tools require a large set of model inputs and can have significant computational time. Applying the typical DTS approach to every single dwelling within the housing stock would therefore require an almost prohibitive amount of data and simulations.

A commonly used approach to overcome this difficulty is the use of *archetypes* (Lim and Zhai, 2017b). A building archetype may be defined as a building definition that represents a group of buildings with similar properties (Reinhart and Cerezo Davila, 2016). Such a process can allow for significant reduction in the amount of models and data required. The specification of archetype models requires two key steps (Reinhart and Cerezo Davila, 2016):

1. **Classification (or Segmentation or Clustering):** The building stock under investigation is divided into groups according to classifiers such as building shape, age, use, occupancy, climate and systems.

2. **Characterisation:** A complete set of building geometry, thermal properties, occupancy patterns and building systems have to be defined for the archetype buildings representing the previously defined groups.

The following sections discuss in detail some of the most important approaches to classification and characterisation.

2.2.1.1 Classification

Classification seeks to cluster a *heterogeneous* building stock into *homogeneous* groups of buildings with similar characteristics, that may each be represented by an archetype (Booth et al., 2012). While this process enables the treatment of some

differences between archetypes, it does not eliminate heterogeneity, since some differences within each group will remain. Thus, it might be more useful to treat heterogeneity and homogeneity as the two ends of a spectrum, where increasing the number of archetypes may result in increased modelling detail and reduced levels of heterogeneity at the expense of a greater computational burden and data requirement (Lim and Zhai, 2017b).[4]

In practice, the number of archetypes following classification can vary from less than ten to several thousands (Reinhart and Cerezo Davila, 2016). In the absence of data, this process has mostly been carried out on an *ad hoc* basis, with building typology and age being common classifiers (Reinhart and Cerezo Davila, 2016). Several examples of this approach exist within the literature (Filogamo et al., 2014; Mata et al., 2014). When data are available, then a statistical approach may be employed instead (Cerezo et al., 2017). Such an approach might involve, but is not limited to, the use of multivariate regression analysis to determine influential variables that could be used as classifiers (Famuyibo et al., 2012; Sokol et al., 2017), or the use of clustering algorithms (Tardioli et al., 2018)

Following an *ad hoc* approach will not necessarily result in a set of inappropriate classifiers, nor would a statistical approach ensure a better set of archetype models. However, where adequate empirical data is available, a rigorous data-driven approach may provide useful insights that could otherwise be missed, resulting in a potentially more detailed classification process. In some cases, as demonstrated by Sokol et al. (2017), a data-driven approach can result in archetype models with better predictive performance.

2.2.1.2 Characterisation

Several methods of building characterisation have been used, from site audits to detailed, high resolution data collection and intrusive testing methods (Coakley et al., 2014). In the case of building stock models, modellers will often rely on national building surveys, building codes and standards (Cerezo et al., 2017). National

[4]Depending on the quantity of interest, the spectrum is likely to be different. For example, the heating system might induce heterogeneity for a modeller interested in heating demand or carbon footprint but may be ignored if the focus is on illuminance levels.

building surveys, an example of which is the English Housing Survey (EHS), may often provide useful information relating to the dwelling's physical characteristics and occupancy for a nationally representative sample of the housing stock (DLUHC, 2021). Several model inputs could be read directly off the survey datasets, for example floor-to-ceiling height or floor area. Despite the plethora of information available in these datasets, they will not necessarily contain information relating to all the required model inputs. In such cases, a modeller might either assume some default values or try to infer values through the use of *proxy variables*, defined in this thesis as: variables that can be measured, or otherwise quantified, more easily than a variable of interest and are used to infer a value for the variable of interest. For example, since the EHS does not contain any information on the building's thermal transmittance characteristics, dwelling age and wall construction may be used as proxy variables, together with reference tables (e.g. Appendix S of the Standard Assessment Procedure – RdSAP) to infer the desired model input (BRE, 2019).[5]

2.2.2 UK Housing Stock Model

In response to increasing concerns surrounding indoor overheating, researchers of the *Air pollution and WEather-related health impacts: methodological study based on Spatio-temporally disaggregated multi-pollutant models for present-day and future* (AWESOME) and the *Health Protection Research Unit* (HPRU) projects developed an archetype-based modelling framework for assessing the summer indoor temperatures of the UK housing stock (UCL IEDE, 2016; UCL IEDE, 2017). The following paragraphs outline some key applications of the UK Housing Stock Model (UK-HSM), while modelling details are reserved for Section 3.5.

As illustrated in Figure 2.2, the first publications on UK-HSM introduced the modelling framework and assessed the impact of retrofit under current and future weather on summer indoor temperatures (Oikonomou et al., 2012; Mavrogianni et al., 2012). The effect of occupancy patterns, window and shading operation – found to be substantial – was later assessed (Mavrogianni et al., 2014), and the risk of indoor

[5] In the example presented in text, the U-value may be inferred in this way. This may then need to be transformed into thickness and thermal conductivity of each layer of the construction element to be used in common dynamic thermal models.

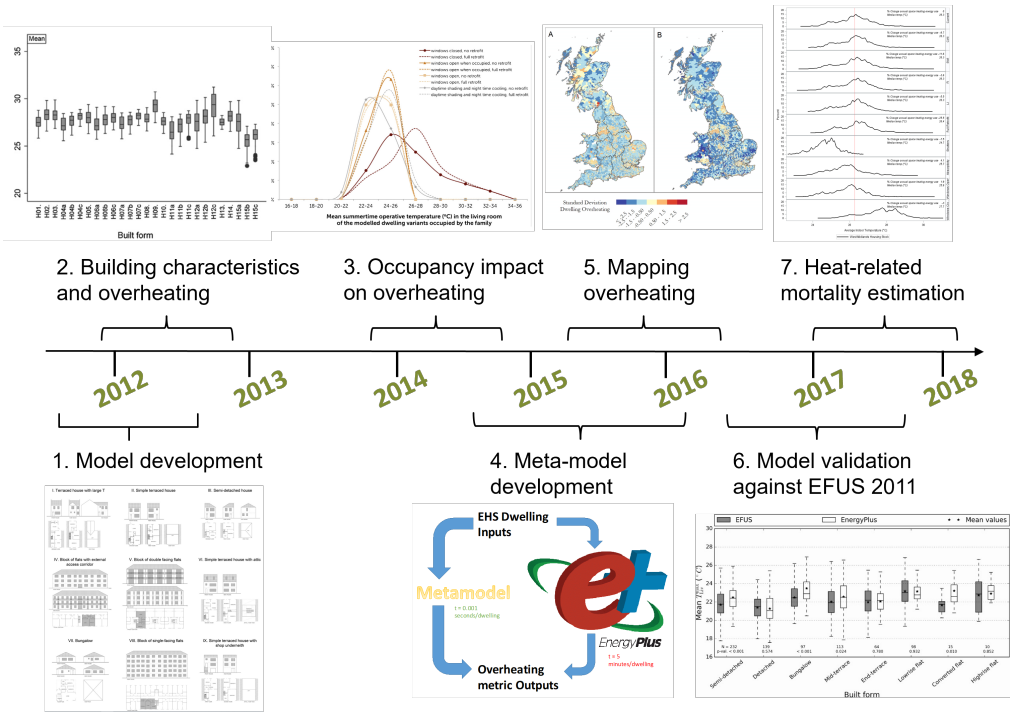


Figure 2.2: Timeline of key publications associated with the UK Housing Stock Model. Relevant papers: 1. Oikonomou et al. (2012), 2. Mavrogianni et al. (2012), 3. Mavrogianni et al. (2014), 4. Symonds et al. (2016), 5. Taylor et al. (2016), 6. Symonds et al. (2017), 7. Taylor et al. (2018b).

overheating was mapped across Great Britain (Taylor et al., 2016). The modelling framework has also been used to provide estimates of heat-related mortality in London (Taylor et al., 2015) and the West Midlands (Taylor et al., 2018b). To reduce the substantial computational cost of UK-HSM, Symonds et al. (2016) developed a set of Artificial Neural Network (ANN) models using the archetype-based modelling framework, enabling the rapid assessment of indoor temperatures for a large number of dwellings.

Through the different stages of model development, the number of classifiers and archetypes varied. The choice of classifiers was largely informed by literature available at the time, with UK-HSM also being used as an investigation tool to identify and rank parameters of influence (Mavrogianni et al., 2012; Taylor et al., 2014). For the London-centric version of the model, Mavrogianni et al. (2012) used construction age and built form type to form 15 archetypes. By varying the floor level of purpose-built flats (ground/mid/top), and running a parametric analysis

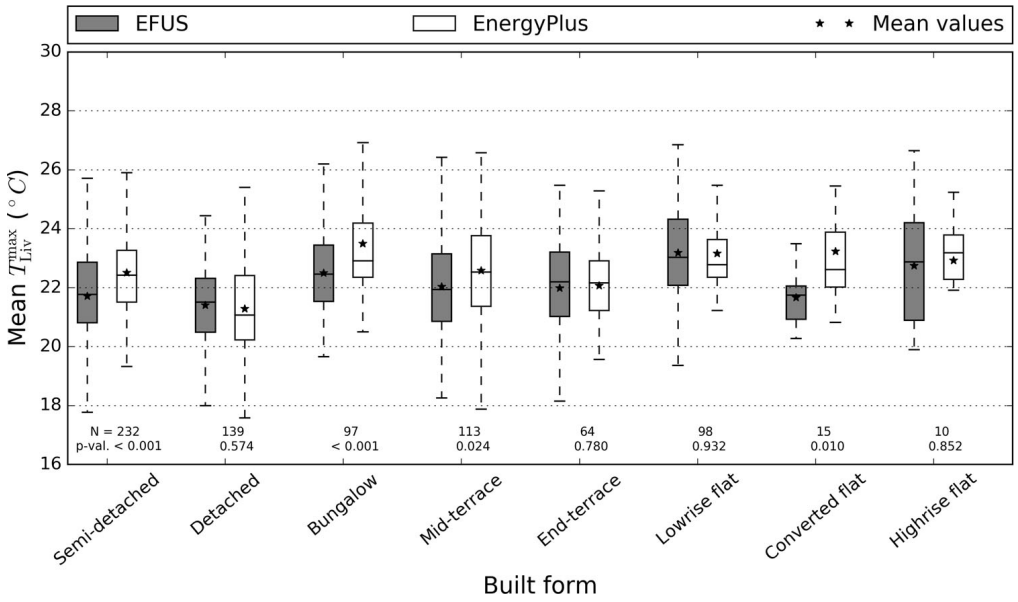


Figure 2.3: Comparison of empirical and modelled mean of the daily max living room temperatures (Symonds et al., 2017). The empirical measurements were collected during the 2011 Energy Follow-Up Survey.

for the insulation level (as-built/retrofitted), orientation (four cardinal directions) and external environment morphology (stand-alone/part of larger building structure), 3456 simulation were run. In later applications, the archetypes became more detailed, allowing for parameters such as the wall type (solid/cavity wall), terrain (city/urban/rural), floor area and floor-to-ceiling height to be specified (Taylor et al., 2015; Taylor et al., 2016).

The floor area and ceiling height were often informed directly from survey data (Taylor et al., 2016) and in some cases from GIS maps (Mavrogianni et al., 2012). Model inputs of building thermal characteristics and air permeability were inferred based on the dwellings' construction age through the use of RdSAP in combination with the EHS (Taylor et al., 2015; Taylor et al., 2018a) or EPC dataset (Taylor et al., 2019). Empirical data to characterise the indoor temperature at which windows open, an important model input for indoor overheating risk (Mavrogianni et al., 2014), were and still are limited. Mavrogianni et al. (2012) used a threshold of 25 °C for the living rooms and 23 °C for the bedroom based on the indoor temperatures recommended by CIBSE Guide A 2006 as thermally comfortable in non-air conditioned dwellings during the summer. The same rationale was used by Taylor et al. (2019), although

based on CIBSE Guide A 2015, with daytime temperature of 23 °C for the living room and night time temperature of 21 °C for the bedroom. Fixed values were used for electrical gains based on previously published work (Symonds et al., 2016).

To evaluate the predictive performance of UK-HSM, Symonds et al. (2017) undertook a detailed empirical validation study against the indoor temperatures monitored in 823 dwellings during the 2010/2011 EFUS. The model inputs relating to the dwellings' thermal characteristics were inferred from their construction age using RdSAP. For each dwelling, three different values were used for the internal gains from electrical appliances and window opening threshold, corresponding to a base case, an upward and downward variation. The comparison between modelled and monitored daily maximum temperatures for the living room and bedroom showed a better agreement at aggregate level (for each typology in each region) than at the individual building level (see Figure 2.3 for a comparison of aggregated living room temperatures). For the semi-detached typology, the most frequently occurring building typology in England, the archetype-level Root Mean Square Error ranged between 0.94 to 1.73 °C depending on location while for individual dwellings it ranged from 2.34 to 2.89 °C.

2.3 Treatment of Modelling Uncertainties

Symonds et al. (2017) hypothesised that the differences observed could have been due to several factors, including the lack of knowledge about certain model inputs, simplifying modelling assumptions and biases in the monitored data. Following from the results, crucial next steps in the model's development are the quantification of its uncertainties and the improvement of its predictive performance (Symonds et al., 2017). To achieve this, there is value in reflecting on the relationship between modelling and uncertainties, and identifying common sources of uncertainties before discussing ways to characterise and minimise them.

2.3.1 The Nature of Modelling and its Uncertainties

Uncertainties are an integral part of the modelling practice and the scientific method (Saltelli et al., 2008). To understand why, one has to first look at the nature of the

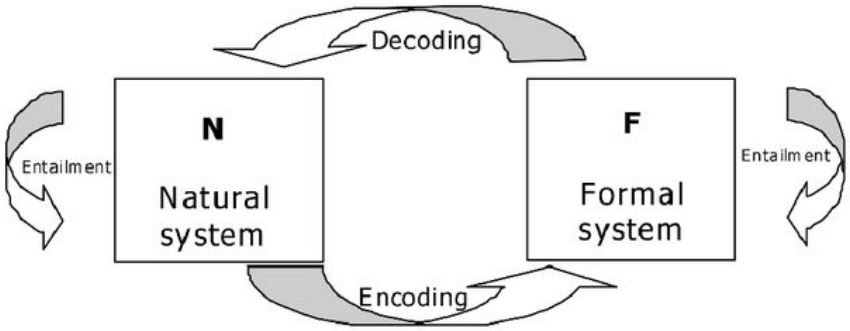


Figure 2.4: Rosen's modelling diagram, published in 1991 and reproduced from Saltelli et al. (2008).

modelling process. A useful way of visualising the modelling process is provided by Rosen's diagram (fig. 2.4), published in 1991 and discussed by Saltelli et al. (2008). For a Natural System (**N**) that we might observe, we can model it by hypothesising its structure and encoding it into a Formal system (**F**). Through decoding, we can then make inferences from **F** and compare them with our observations of **N**. While both systems are entailed, following their own set of rules, there is no rule on how to create **F** as to accurately reproduce the behaviour of **N** (the importance of this insight by Rosen is highlighted by Saltelli et al. (2008)). In essence, **F** depends on our observations of an arbitrary portion of **N**, our understanding of the causality rules that **N** follows and our ability to capture this in a mathematical (or any other) manner. Therefore, a model is derived from aspects of the real system and aspects from the modeller's perception of the system (Mulligan and Wainwright, 2013). Hence, the term *model* may be defined as "an abstraction of a real system; it is a simplification in which only those components that are seen to be significant to the problem at hand are represented in the model" (Mulligan and Wainwright, 2013). Different interpretations of the same **N** (i.e. different models), depending on the modeller, their skills, knowledge and purpose may arise. And in the process of observing the natural system, understanding its causality rules, encoding them into a formal system and using it for prediction, uncertainties may arise due to our limited capacity to perform all the aforementioned steps.

2.3.2 Classes and Sources of Uncertainty

Uncertainties are often classified as (Kiureghian and Ditlevsen, 2009):

1. **Aleatory uncertainties:** The result of inherent randomness of the system studied that may be described using a distribution. This definition derives from the Latin word for dice player, *aleator* (McClarren, 2018).

2. **Epistemic uncertainties:** These uncertainties arise from the lack of knowledge about the system studied. The term derives from the Greek word for knowledge (Kiureghian and Ditlevsen, 2009).

The key distinction between aleatory and epistemic uncertainties is that the latter could be reduced or eliminated if enough information becomes available. It has been argued that this distinction is an artificial one; based on the modeller's assumptions, a source of uncertainty might be classified as *epistemic* if they wish to reduce it or *aleatory* if not and between different models the same source of uncertainty might be classified as reducible or not (Kiureghian and Ditlevsen, 2009).

The six sources of uncertainty expected to be common amongst most computer models are (Kennedy and O'Hagan, 2001):

1. **Parameter uncertainty:** The result of lack of knowledge about the true values of some model inputs. They may be context-specific or global parameters that could be reduced if enough information was available, thus falling within the class of *epistemic* uncertainties.

2. **Residual variability:** The process may not always take the same value for the same known conditions, possibly due to the process being inherently unpredictable and stochastic. This variation, even when the conditions are fully specified, is referred to as *residual variability*.

3. **Model inadequacy:** If it is assumed that the true, real-world, values of the model inputs are accurately known, differences between the mean of the real process (i.e. averaging out residual variability) and the model output might still exist since no model is perfect. This uncertainty might exist because there is

lack of knowledge on how to model the true process (referred to as *ignorance* by Booth et al. (2012)) or because simplifying assumptions were made during the model construction.

4. **Parametric Variability:** Some model inputs might be allowed to vary because the modeller does not want (or is unable) to fully specify them. By allowing model inputs to vary, the output acquires this extra uncertainty.

5. **Observation error:** Any observation of the true system used as a model input or for calibration purposes can have an associated uncertainty (or measurement error) that might depend on the measuring process and the resolution of the measuring instrument.

6. **Code uncertainty:** Given a set of inputs for a complex model, the outputs are not known until a simulation is performed and thus may be considered unknown. If the code is computationally expensive, it might be impractical to run the code for all desired inputs, resulting in an associated uncertainty.

An important point to highlight from the discussion around model inadequacy is that, in some cases, the model input values which result in the best agreement between model output and real-world observations may not equal the real-world (true) values of the physical parameters they represent (Kennedy and O'Hagan, 2001).

2.3.3 Uncertainties in Building Stock Modelling

Uncertainties are also integral to building modelling. Several papers have discussed the topic of uncertainty within the built environment field, with a comprehensive review on the topic provided by Tian et al. (2018). With regard to the uncertainties associated with archetype-based modelling of summer indoor temperature, it is worth revisiting the steps of classification and characterisation.

Since eliminating heterogeneity is – in practical terms – not possible, uncertainties arise during the process of classification. Heterogeneity, as a source of uncertainty in building stock modelling, was discussed by Booth et al. (2012) who drew a parallel with research in the medical field and the heterogeneity between individuals. While a thorough classification process can reduce its effects, any remaining

heterogeneity will result in an uncertainty that may fall under parameter uncertainty, model inadequacy and residual variability (depending on the modelling viewpoint).

Uncertainties also arise during the characterisation process, especially when model inputs have to be inferred using proxy variables. This may be demonstrated by considering the fieldwork carried out by the Building Research Establishment (BRE) that aimed to compare the thermal performance of walls in 300 English dwellings against values assumed in RdSAP (Hulme and Doran, 2014). In-situ measurements were taken using heat flux plates (Hukseflux HFP01) and surface temperature measurements for a period of two weeks. In a subset of 10 cavity wall dwellings, inspected in more detail since their measured U-value was in large disagreement with their assumed value, nine were found to have been miss-classified as uninsulated within the EHS. A comparison of the mean measured and corresponding RdSAP U-value for each construction type revealed differences of up to 34 % (BRE, 2016). Finally, for each construction type, a large spread of U-values was observed, with the standard deviation ranging per category from 0.23 to 0.32 $\mathrm{W/m^2K}$ (Hulme and Doran, 2014). This spread is partly due to natural variability of wall constructions, with evidence of varying quality and density of insulation when present. In combination, errors in survey data (proxy variables) along with inaccurate reference tables that do not capture the possible spread in values, may all contribute to the poor selection of model inputs and large parameter uncertainties. Uncertainties surrounding model inputs also arise when the variable of interested is measured directly, since a difference between the measured quantity and the true value (*observational error*) will exist.

2.3.4 Uncertainty Analysis

Uncertainty analysis aims at *quantifying* the uncertainty of a Quantity of Interest (Saltelli, 2004), and generally falls under two broad categories (Tian et al., 2018):

1. **Forward:** Also known as *uncertainty propagation*, this process aims at quantifying the uncertainty in the model outputs due to uncertain input variables.

2. **Inverse:** Also referred to as *calibration*, this process tries to determine unknown input variables through mathematical models from measurement data.

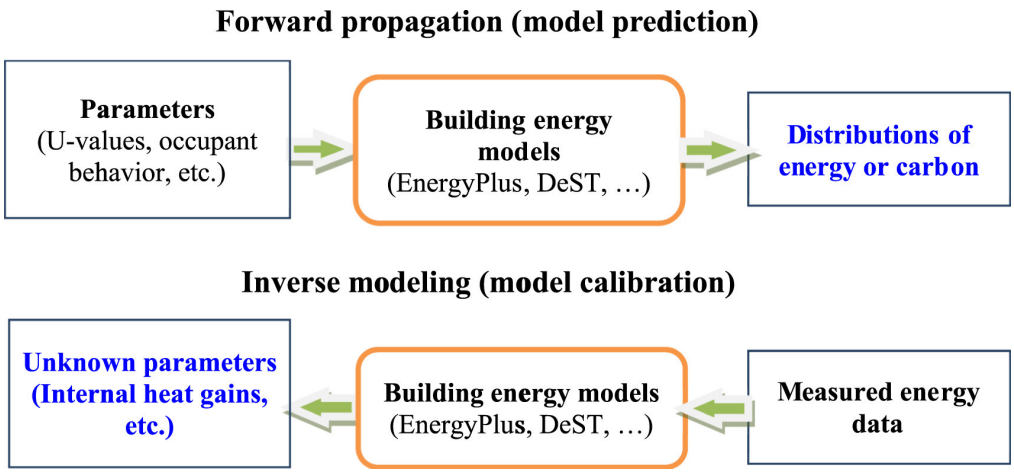


Figure 2.5: Illustration of the forward and inverse uncertainty quantification in building energy analysis, as presented by Tian et al. (2018).

An example of how each process might be used in building energy analysis is provided by Tian et al. (2018) and captured by Figure 2.5. Forward uncertainty quantification can quantify the variation in energy use or carbon emissions predicted by a building energy model given some input variations. Inverse uncertainty quantification can quantify the variation of unknown input variations through building energy models with the use of monitored energy data. Forward and inverse modelling are often linked, and several methods exist for each category (Smith, 2013; Tian et al., 2018). The following subsection will provide an overview of calibration methods.

2.3.5 Approaches to Calibration

Kennedy and O'Hagan (2001) defined *calibration* as the process of learning the values of unknown model inputs using field observations of the model output. This term is often used interchangeably with *inverse uncertainty analysis* (Tian et al., 2018) or *model tuning* (Coakley et al., 2014), although their equivalence depends on the calibration method chosen and its intended use. The following section will discuss some of the calibration approaches implemented within the field of building modelling and place emphasis on statistical techniques.

Coakley et al. (2014) conducted a thorough review of methods relating to the development and calibration of building simulation models and categorised the various approaches within two broad groups:

1. **Manual:** Approaches that predominantly rely on iterative pragmatic intervention by the modeller and do not make use of automation or mathematical/statistical methods. The modeller's interventions might vary from ad-hoc parameter-tuning approaches based solely on their expertise to more systematic efforts that might utilise advanced graphical or procedural methods.

2. **Automated:** Approaches that have some form of not-user-driven input and may include mathematical/statistical methods such as penalty or objective function optimisation and Bayesian inference.

A concern when calibrating complex models, such as BPS models, is that multiple solutions might exist; this has been referred to as *equifinality* or *model indeterminacy* (Coakley et al., 2014). An automated, statistical, approach can generally allow for a more efficient and rigorous exploration of the possible solutions when compared to manual methods. This led Coakley et al. (2014) to conclude that the papers reviewed were at best using a method "based on an optimisation process used to identify multiple solutions within a parameter space identified from a knowledge-base of templates of influential parameters" and at worst "based on an ad-hoc approach in which the analyst manually tunes the myriad of parameters until a solution is obtained".

Despite the potential strengths of automated techniques, Coakley et al. (2014) discovered that manual methods were more popular, at least at the time of publication, with 74 % of the papers reviewed having employed manual techniques. This could be the result of the "Manual" category encompassing more approaches, approaches that are more widely known, or that built environment researchers did not have the expertise required to adopt more advanced, automated approaches. In addition, the computational expense and in some cases the lack of data might have prevented the use of some automated techniques. Nowadays, with the increase in computing power and data availability, statistical (machine learning) techniques stemming from the fields of computer science and applied statistics are on the rise and may be used to solve various calibration problems.

To explain the framework that underpins statistical techniques of inverse model-

ling, the general problem of computer model calibration from a statistical perspective will be derived and the two general approaches to solving it will be discussed.

2.3.6 Theory of Statistical Calibration

For a physical (or natural) system, denoted as $\zeta(\cdot)$, a set of n observations $\mathbf{y}$ ($\mathbf{y} = y_{i=1}, \ldots, y_n$) are made at conditions $\mathbf{W}$ ($\mathbf{W} = \mathbf{w}_{i=1}, \ldots, \mathbf{w}_n$) (Kennedy and O'Hagan, 2001; Higdon et al., 2004):

$$y_i = y(\mathbf{w}_i) = \zeta(\mathbf{w}_i) + \varepsilon_i, \quad i = 1, ..., n \tag{2.1}$$

where ε_i denotes the observation (or measurement) error. Within the built environment context, and for simplicity of this derivation, the physical system could be a single-zone test house with the observations (captured by $\mathbf{y}$) being the monitored monthly mean indoor temperatures given some weather conditions (captured by $\mathbf{W}$).[6] All conditions that could influence the indoor temperature, other than the weather, are assumed to be constant over the duration of monitoring. The observation error captures the discrepancy between "true" indoor temperature and the value measured by the monitoring equipment.[7]

To simulate the physical system, a computer model may be used that would likely require several inputs that may fall into one of three groups:

1. Inputs thought to be known accurately and which vary between observations; in this example, this refers to the weather variables.

2. Inputs considered to be accurately known and which do not vary (these can effectively be ignored for the rest of the derivation).

3. Inputs that do not vary between observations, but whose value is unknown. These variables are represented by $\mathbf{t}$.[8]

[6] The weather conditions are represented by a matrix, instead of a vector, since each month is associated with a few weather variables.

[7] A "true" value is one that would be obtained from a *perfect measurement* and is, by nature, indeterminate (BIPM et al., 2008). In this example, a true value might refer to the mean indoor temperature of the zone, if it could be measured without any measurement error. The use of the term "true" in this derivation stems from the work of Kennedy and O'Hagan (2001).

[8] This simple derivation does not address inputs whose values are unknown and vary between

Given a set of model inputs $(\mathbf{W}, \mathbf{t})$, the computer model can provide a set of outputs $\eta(\mathbf{W}, \mathbf{t})$. When the right selection of calibration parameters is made ($\mathbf{t} = \boldsymbol{\theta}$), the simulator ($\eta(\mathbf{W}, \boldsymbol{\theta})$) can effectively simulate the physical system ($\zeta(\mathbf{W})$) and the following statistical relationship between the observations of the physical process and the model outputs may be established:

$$y(\mathbf{w}_i) = \eta(\mathbf{w}_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.2}$$

Equation 2.2 captures the basic form of the calibration problem, where the aim is "to determine $\boldsymbol{\theta}$ given the noisy observations $\mathbf{y}$" (Smith, 2013). As discussed in Section 2.3.2, model form (or structural) uncertainty may arise from the computer model's simplifications and approximations. In some cases this might be negligible, but often it has to be accounted for. This can be done by adding a model discrepancy term to Equation 2.2, yielding the complete formulation of the calibration problem (Kennedy and O'Hagan, 2001):

$$y(\mathbf{w}_i) = \eta(\mathbf{w}_i, \boldsymbol{\theta}) + \delta(\mathbf{w}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.3}$$

where $\delta(\mathbf{w}_i)$ captures the discrepancy between the computer model output $\eta(\mathbf{w}_i, \boldsymbol{\theta})$ and true process $\zeta(\mathbf{w}_i)$ and prevents over-estimation of the calibration values (Higdon et al., 2004; Li et al., 2016).

The calibration problem may be solved using a *frequentist* or *Bayesian* approach (Smith, 2013; Tian et al., 2018). A thorough comparison of these two paradigms can be found in Bolstad and Curran (2017) and Smith (2013). Briefly, a frequentist approach treats parameters as unknown but fixed, and relies on observations alone to estimate their value. On the other hand, a Bayesian approach considers parameters to be associated with probability density functions and their inference is informed both by data and by expert knowledge. In the Bayesian paradigm, the probability density function associated with each unknown parameter following calibration embodies the uncertainty surrounding its value. This uncertainty, within

observations.

the Bayesian world, can be "naturally" propagated to the model output to quantify the epistemic uncertainty that remains (Smith, 2013). Recognising that uncertainties cannot simply be eliminated, due to the complexity of the natural and formal system, and limitations in data availability, a calibration framework for housing stock models that captures uncertainty in addition to improving predictive performance will likely lead to better-informed decision-making (Booth et al., 2012). In practice, Bayesian model calibration has been shown to perform well, with studies that employed such techniques generally reporting lower errors than other methods, according to the review of urban building energy modelling by Oraiopoulos and Howard (2022). For these reasons, the following sections will focus on the theory and application of Bayesian calibration. A critical evaluation of published work on the Bayesian calibration of archetype-based housing stock models is offered in Section 2.4.

2.3.7 Theory of Bayesian Inference

At the core of Bayesian inference is Bayes' theorem (McElreath, 2020):

$$\text{Posterior} = \frac{\text{Probability of the data} \times \text{Prior}}{\text{Average probability of the data}} \tag{2.4}$$

The components of Bayes' theorem are:

- *Prior* ($p(\boldsymbol{\theta})$): This captures how plausible each value of a parameter (e.g. θ_1) is, according to the modeller's subjective opinion, before observing the data (Bolstad and Curran, 2017).
- *Probability of the data (Likelihood)* ($p(\mathbf{y}|\boldsymbol{\theta})$): Represents the relative weights of belief for the observed data ($\mathbf{y}$), given a value for the unknown parameter (Bolstad and Curran, 2017).
- *Average probability of the data* ($p(\mathbf{y})$): It is the probability of the data averaged over the prior, and it acts to standardise the posterior so that it sums (or integrates) to one (McElreath, 2020).
- *Posterior* ($p(\boldsymbol{\theta}|\mathbf{y})$): Represents the relative weights of belief for each parameter value after analysing the data (Bolstad and Curran, 2017).

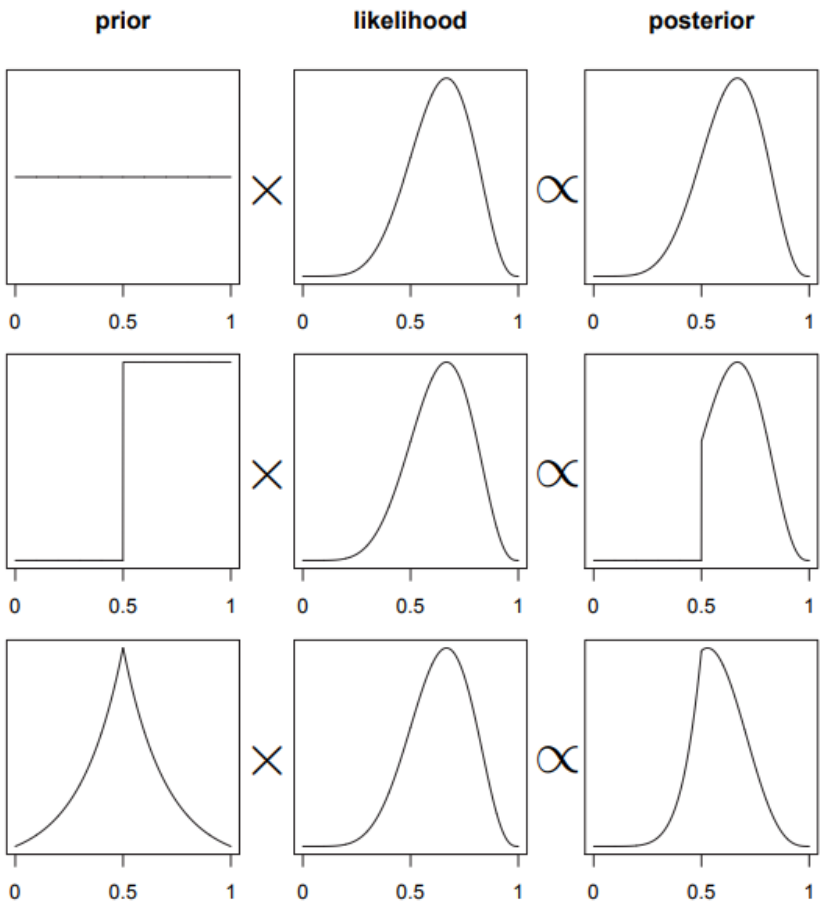


Figure 2.6: Estimating the posteriors for the same likelihood assuming different priors. Reproduced from McElreath (2020).

For a fixed set of observations, $p(\mathbf{y})$ can be considered constant and the unnormalised posterior density is defined as (Gelman, 2014):

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}) \tag{2.5}$$

The interaction of the likelihood and prior can be visualised in Figure 2.6, which highlights the importance that the choice of prior can play in Bayesian inference.

2.3.8 Steps to Bayesian Calibration

The Bayesian calibration of DTS models is a multistep process. As Chong and Menberg (2018) identified, guidance on carrying out this procedure within the built environment field has been limited. In response, Chong and Menberg (2018)

published guidelines on its implementation. Six key steps were identified, depicted in Figure 2.7 and discussed below.

1. Collect field data. The collected field must be cleaned and potentially aggregated according to the desired temporal resolution. An investigation of the effect of temporal resolution on the calibration process revealed a systematic trend of increasing predictive accuracy with increased (district heating) training data resolution (Kristensen et al., 2017b), although differences between models and data may exist.

2. Parameter screening. A key step in the Bayesian calibration process is the selection of variables to calibrate. Selecting only the most influential parameters to calibrate reduces the computational time – that can often be prohibitive – and can also reduce the risk of *non-identifiability*: when a unique combination of calibration parameters does not exist or cannot be determined, resulting in either weak (uninformative) posterior distributions or ones that mirror the priors (Kristensen et al., 2018; Menberg et al., 2019). Chong and Menberg (2018) caution that concerns regarding parameter identifiability must be balanced against that of *over-fitting*, as reducing the number of parameters can result in unreasonably tight posterior distributions with poor out-of-sample performance. When parameter screening is applied carefully, calibrating only a subset of parameters can yield reliable posterior distributions and result in models with good predictive capabilities, as demonstrated by Heo et al. (2015). A frequently used approach to selecting the calibration parameters is the Morris method, as revealed by the review of Hou et al. (2021).

3. Create computer data. It is common practice in Bayesian calibration applications to use a surrogate model (also known as metamodel or emulator) to reduce computational time. Briefly, a *surrogate model* is a statistical model trained to reproduce the predictions of a computationally expensive model at a fraction of a time (for a detailed example, please refer to Symonds et al., 2016). If this approach is chosen, computer simulations are performed to generate the data needed to train a surrogate model. The explanatory model inputs and calibration parameters must be varied within the range of possible values. For the calibration parameters, this is often done using a sampling approach such as the Latin Hypercube Sampling (LHS) method,

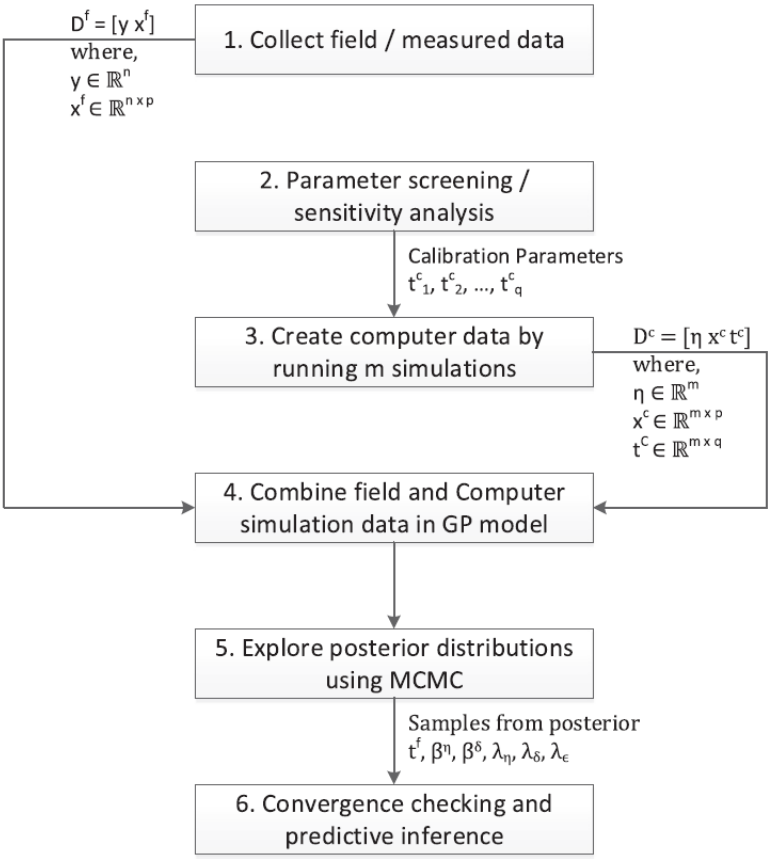


Figure 2.7: Bayesian calibration procedure, reproduced from Chong and Menberg (2018). GP stands for Gaussian process, MCMC for Markov Chain Monte Carlo.

attempts to maximise the coverage of the input space (Chong and Menberg, 2018).

4. Combine field and computer simulation data in GP model (Surrogate Model). In the commonly used approach introduced by Higdon et al. (2004) and preferred by Chong and Menberg (2018), a Gaussian Process (GP) is used as the surrogate model which is trained on both the field and computer data at the same time as the model parameters are calibrated. GP is a powerful and flexible surrogate model that has often been used in the Bayesian calibration of building models (Lim and Zhai, 2017a). A more detailed, mathematical, description of GP is offered in Section A of the appendices. Training a GP model concurrently with parameter calibration, enables the uncertainty of the GP's *hyperparameters* (parameters that define the GP model) to be captured. The GP's main limitation is its scalability with data; training a GP model has a runtime complexity of $\mathcal{O}(N^3)$ where N is the combined number of

field data and computer data upon which the surrogate model is trained on (Chong et al., 2017). Other surrogate models may be used, and their training can take place ahead of the calibration (Lim and Zhai, 2017a; Tardioli et al., 2020).

5. Explore posterior distributions using MCMC. The most frequently used method for estimating posteriors relies on the use of Markov Chain Monte Carlo (MCMC) (Hou et al., 2021). This is a sampling-based approach, where Bayes' theorem (Equation: 2.5) is iteratively computed at different values of the calibrated parameters until convergence (this is discussed in more detail in Section D.3 of the appendices). In a study that compared the performance of MCMC algorithms in the Bayesian calibration of building energy models, Chong et al. (2017) determined that the No-U-Turn-Sampler (NUTS), an extension to the Hamiltonian Monte Carlo (HMC) algorithm, performed better than other popular algorithms such as the Random Walk Metropolis (RWM) or Gibbs sampling algorithm.[9] Although not captured within the flowchart of Chong and Menberg (2018), before the posterior exploration it is necessary to decide on the priors and the likelihood function:

- **Prior:** Parameter priors are chosen based on best available knowledge surrounding the possible parameter values. If justified, the use of informative priors could help improve posterior identifiability (Smith, 2013). If a Bayesian approach to training the surrogate model is followed, the priors for the surrogate model's hyperparameters must also be defined (Chong and Menberg, 2018). If a term for the model bias is also considered, priors associated with that component must also be defined - this is also often treated as GP.

- **Likelihood:** It is common practice to assume a normal likelihood function by treating the error term in eq. 2.3 as being independently and identically distributed, following a normal distribution (Higdon et al., 2004). Other likelihood functions can also be used (Cerezo et al., 2017).

[9]This is because HMC NUTS avoids random walk behaviour and is insensitive to correlated parameters, thus often resulting in faster convergence when dealing highly multi-dimensional problems (Chong et al., 2017).

6. Convergence checking and predictive inference. MCMC convergence is assessed separately for each parameter by using trace plots of multiple Markov chains, and by evaluating the *potential scale reduction*, also known as Gelman-Rubin statistic (Chong and Menberg, 2018; Gelman, 2014). If a satisfactory level of convergence has been achieved, the posterior distributions may be used for predictive inference. This may be done directly from the surrogate model.

2.4 Bayesian Calibration of Housing Stock Models: A Critical Evaluation

The first implementation of Bayesian calibration on archetype-based building stock models was proposed by Booth et al. (2012). Since then, several other applications of Bayesian inference for archetype model calibration have been published (see reviews by Oraiopoulos and Howard (2022) and Hou et al. (2021)). In this section, the studies that focused on archetype-based housing stock models are critically reviewed, compared and discussed. A summary of each study's key characteristics is provided in Tables 2.1–2.4.

2.4.1 Location, Model, Data Resolution & Housing Stock

The housing stock's location, the dwelling types and sample size used for the calibration varied between application (Table 2.1). It was common for papers to focus on a single dwelling type, or a small number of dwelling types. Dwelling sample sizes varied from 35 flats located in West Salford (UK), in the work of Booth et al. (2012), to 2972 Swiss multi-family buildings in the study by Tardioli et al. (2020). EnergyPlus was the most commonly used Dynamic Thermal Simulation tool and was used in three instances, while ISO-based models were used in four papers. Wang et al. (2020) used CitySim, a dynamic urban energy model which treats each building as single thermal zone and can aggregate heating demand to postcode level (Robinson et al., 2009).

In all papers reviewed, the observations (field data) used for the calibration related to energy use; no example of domestic archetype model calibration that relied

on indoor environmental parameters could be found. The observations were in some cases normalised against the floor area (Booth et al., 2012; Cerezo et al., 2017; Sokol et al., 2017) and building (postcode) volume (Wang et al., 2020). The temporal resolution of the observations ranged from hourly (Hedegaard et al., 2019) to annual (Cerezo et al., 2017; Sokol et al., 2017; Tardioli et al., 2020; Wang et al., 2020).

2.4.2 Classification

The approach to, and level of discussion on, the classification of dwellings into archetypes varied between papers (Table 2.2). Half of the authors used a single archetype in their analysis, while at the upper end of the spectrum Wang et al. (2020) used 18. A single archetype was assumed by Booth et al. (2012) since the 35 'physically similar' flats in the study were thought to form a homogeneous cluster. Factors relating to occupant characteristics were not considered by Booth et al. (2012), although they were mentioned in the paper's discussion on ways to extend their work. Kristensen et al. (2017a) also used a single archetype (detached single-family dwellings constructed between 1979-1998) which was derived from the TABULA project. The TABULA and EPISCOPE projects sought to develop residential building typologies in 21 European countries that could be used in housing stock models of energy consumption and guide policies (Loga et al., 2016; TABULA Project Team, 2012; TABULA Project Team, 2017). Occupancy is not considered in the classification process, with the main (and often only) classifiers being the dwelling age and typology. Hedegaard et al. (2019) and Kristensen et al. (2018) also used a single archetype in their analysis. Hedegaard et al. (2019) does not discuss classification while Kristensen et al. (2018) mentions the classifiers used to select their archetype, but does not explain the rationale behind this choice. The 18 archetypes used by Wang et al. (2020) were also informed by TABULA/ EPISCOPE, together with the Dutch national reference home standard. Four archetypes were used by Cerezo et al. (2017) with the only classifier being the construction or renovation period, and with limited discussion on why this choice of classifiers was preferred.

Only two papers used data driven methods to inform their classification process. Sokol et al. (2017) employed multivariate linear regression analysis to identify

several variables that had a statistically significant association with energy use intensity. Two such variables, effective year built and heating system efficiency, were used as classifiers, while some of the remaining variables were modelled explicitly. Tardioli et al. (2020) used 14 archetypes, derived through a process that included the classification of dwellings using three categorical variables (building typology, construction period and final use), the normalisation of continuous building characteristics, and the further subdivision of the stock using a set of clustering algorithms; the process is described in depth in Tardioli et al. (2018). The approach followed by Tardioli et al. (2020) purposefully placed equal weight on continuous building characteristics being considered, including the variable that corresponded to the observations (e.g. energy use) used for the calibration.

2.4.3 Choice of Calibration Parameters

In all cases, only a subset of the model inputs were calibrated (Table 2.2). Sokol et al. (2017) and Hedegaard et al. (2019) provided limited discussion regarding their approach to selecting calibration parameters. Cerezo et al. (2017) cited modeller's expertise and a simplified sensitivity analysis to explain their choice of four calibration parameters, while Tardioli et al. (2020) selected six parameters based on literature. Sobol's method was used by Kristensen et al. (2017a) to select seven calibration parameters; a resource intensive approach that required 15,000 simulations. The Morris method was used by Booth et al. (2012), Kristensen et al. (2018) and Wang et al. (2020). Booth et al. (2012) were the only authors to have repeated the calibration for a different number of variables (2, 4 and 6) and compared the associated posterior distributions (but not the influence of this choice on predictive performance). This parametric exercise demonstrated the "lumping" of uncertainties from uncalibrated model parameters onto the calibrated ones, which prevents a modeller from inferring a parameter's real-world value from the posterior distribution.

2.4.4 Surrogate Modelling

In three out of the eight papers, surrogate modelling was used to speed up the calibration process (Table 2.3). Kristensen et al. (2017a) used a Gaussian process, Sokol

et al. (2017) used polynomial regression, while Tardioli et al. (2020) compared the use of four techniques: Generalised linear models, artificial neural networks (ANN), support vector machines and random forests. In addition, Tardioli et al. (2020) also assessed the effects of the "surrogate strategy" by comparing the predictive performance of using one emulator for each building (OEFEB), one emulator for each cluster (OEFEC) and one emulator for a representative building (OEFRB). OEFEB resulted in the best performing emulators, followed by OEFEC and OEFRB. However, implementing the OEFEB requires rich data for each building in the cluster, and was shown to not generalise as well as OEFEC and OEFRB when considering out-of-sample buildings (Tardioli et al., 2020). Across the three surrogate strategies and the 14 clusters, ANN was the emulator that performed best in most cases.

2.4.5 Bayesian framework

2.4.5.1 Choice of Priors for the Calibrated Parameter

Cerezo et al. (2017), Sokol et al. (2017) and Wang et al. (2020) utilised uniform distributions for the priors of their calibration parameters (Table 2.3). All other authors used non-uniform priors or a mixture of both. A uniform prior assumes that any value within the feasible range is equally probable which rarely reflects the modeller's state of knowledge about a parameter. The rationale behind the use of uniform distributions, sometimes, is to purposefully be non-informative and "to let the data speak for themselves" (Gelman, 2014). However, through the choice of a feasible range the modeller assumes two bounds where an infinitesimal change in either direction changes the probability from 1 to 0. In some cases the choice of upper or lower bound might relate to real constraints in the parameter values (e.g. a system's coefficient of performance), however in other cases the choice might not be straightforward (e.g. the heating setpoint).

Regardless of the choice of priors, a common thread amongst most authors was the limited discussion on the rationale behind their chosen priors. Exceptions to the rule were Booth et al. (2012) and Hedegaard et al. (2019) who discussed their choice of priors in more detail, and in both cases the choice was partly informed by empirical data and partly by expert judgment. Yet, when considering the choice of

non-uniform priors, clear examples of how a modeller might determine what prior distribution to use (e.g. gamma or Normal) and with what distributional parameters is missing.

2.4.5.2 Likelihood

A normal likelihood was the most common choice in the papers reviewed (Table 2.3). The only exceptions were by Cerezo et al. (2017) and Sokol et al. (2017), who used a binary likelihood. For each simulation run, the likelihood would equal 1 if the difference between the model's output and the field data was smaller than a predefined threshold (accepting the vector of parameter values used for that simulation run as feasible solutions) and 0 otherwise (rejecting the simulation run's parameter values). Since the prior distributions were assumed to be uniform, the posterior distributions for each archetype were the result of combining all accepted vectors of calibrated parameter values into a single multivariate joint probability distribution. The choice of an arbitrary threshold, whose validity is not based on data or theory, can be questioned. Kristensen et al. (2018) argued that a binary likelihood approach might prove too simple to fully exploit the information available in high resolution data.

2.4.5.3 Data Aggregation

An important consideration when Bayesian calibration is applied to building stock models is how observations from different buildings are treated (Table 2.3). Booth et al. (2012) averaged the daily observations within the cluster before the calibration, eliminating any inter-dwelling variability and enforcing the assumption of homogeneity. On the other hand, Kristensen et al. (2017a) combined the observations of annual energy use for each dwelling into a single vector, and evaluated the likelihood of this vector at the building level.[10] Through this approach, the error term included (and was likely dominated by) residual variability other than measurement error, such as stochastic occupant behaviour and violations of the cluster homogeneity assumption. Both approaches may be thought to be prone to bias from extreme values due to the averaging that takes place before (Booth et al., 2012) or during (Kristensen et al., 2017a) the calibration. Tardioli et al. (2020) also treated each observation as

[10] by assuming that each observation is independently and identically distributed.

coming from the same archetype, similarly to Kristensen et al. (2017a), but the exact implementation varied between the OEFEB surrogate strategy and the OEFEC and OEFRB strategy.

In some cases, the authors chose to calibrate each building separately, resulting in a separate set of posteriors for each dwelling (Hedegaard et al., 2019; Wang et al., 2020; Cerezo et al., 2017; Sokol et al., 2017).[11] Following calibration, Cerezo et al. (2017) and Sokol et al. (2017) combined the posteriors for all dwellings classified under the same archetype to form a single multivariate joint distribution.

A different approach was proposed by Kristensen et al. (2018), who employed a hierarchical structure to infer archetype-parameters. According to the authors, a calibration procedure based on "partially pooled" data would result in parameter estimates that are less prone to outliers, compared to the other types of data aggregation (Kristensen et al., 2018). This does seem to be the case in some applications within the statistics literature (Gelman and Hill, 2007), but it has yet to be shown whether this approach results in better-performing archetype-based models following calibration when compared to other approaches.

2.4.5.4 Model Bias

The Bayesian calibration framework proposed by Kennedy and O'Hagan (2001) included a model bias (or discrepancy) term used to capture the inadequacy of a model to represent the true process, even if the unknown (calibration) parameters were accurately known (Table 2.3). This component was included in three of the eight papers. In Booth et al. (2012), the model bias was determined to be negative and increase in magnitude as the external temperatures decreased, suggesting an inadequacy of the CEN-ISO model in predicting energy use at colder temperatures. The model bias was found to be negligible in the case of Kristensen et al. (2017a), while its magnitude was not explicitly discussed by Tardioli et al. (2020) despite being used in the calibration. In all cases, a Gaussian process was used to represent the model bias.

[11] To be exact, Wang et al. (2020) applied a separate calibration approach for each postcode, corresponding to a group of buildings.

Table 2.1: Comparison of the location, typology, sample size, simulation tool, observations and temporal resolution (Res.) in the studies reviewed.

Authors	Location	Typology	Sample Size	Simulation Tool	Observations	Res.
Booth et al. (2012)	West Salford, UK	F	35	CEN-ISO	Energy use [kWh/m^2/day]	D
Cerezo et al. (2017)	AlQadisyah district, Kuwait City, Kuwait	V	336	EnergyPlus	Energy use intensity [kWh/m^2/year]	A
Hedegaard et al. (2019)	Aarhus, Denmark	DSF	159	Modified hourly ISO 13790:2008	District heating energy use [kWh/hour]	H
Kristensen et al. (2017a)	Aarhus, Denmark	DSF	600	Modified hourly ISO 13790:2008	Energy use [kWh/year]	A
Kristensen et al. (2018)	Aarhus, Denmark	DSF	150	Modified hourly ISO 13790:2008	Energy use [kW/3hour]	3-H
Sokol et al. (2017)	Cambridge, MA, USA	LRB	2662	EnergyPlus	Gas and electricity use intensity [kWh/m^2/year or kWh/m^2/month]	A; M
Tardioli et al. (2020)	Meyrin District, Geneva, Switzerland	MFB	2972	EnergyPlus	Heating demand and hot water use [kWh/year]	A
Wang et al. (2020)	Amsterdam, Netherlands	SF; T; MF	84*	CitySim	Energy use intensity [kWh/m^3/year]	A

Disambiguation. Typology: F = Flats, V = Villas, DSF = Detached single-family, LRB = Low-rise buildings with 1-4 dwelling units, MFB = Multi-family buildings, SF = Single-family, T = Terrace, MF = Multi-family; Resolution (Res.): D = Daily, A = Annual, H = Hourly, M = Monthly.

Table 2.2: Comparison of the classification process, sensitivity analysis (SA) and number of calibration parameters (P) in the studies reviewed.

Authors	Classification Archetypes	Classifiers	Rationale	SA	P
Booth et al. (2012)	1	N/A	Flats from the same block	Morris Method	2, 4, 6
Cerezo et al. (2017)	4	construction (or renovation) period	Limited discussion	Modeller Expertise and simplified SA	4
Hedegaard et al. (2019)	1	Not discussed	Not discussed	Limited discussion	5
Kristensen et al. (2017a)	1	Not discussed	TABULA	Sobol's Method	7
Kristensen et al. (2018)	1	Usage/Type; Construction Period; Location; Stories; Basement; Attic utilised for living; Heating source; Suppl. heating	Limited discussion	Morris Method	5
Sokol et al. (2017)	8	Effective Year Built; Heating COP	Multivariate linear regression (for COP)	Limited discussion	6
Tardioli et al. (2020)	14	Pre-clustering: Building typology; Construction Period; Final use	Automated clustering process (Tardioli et al., 2018)	Literature	6
Wang et al. (2020)	18	Dwelling type; Construction year	Dutch Standard, EPISCOPE and TABULA	Morris Method	2

Table 2.3: Comparison of the surrogate approach and Bayesian calibration framework in the studies reviewed.

Authors	**Surrogate**		**Bayesian Calibration Framework**			
	Model	**Strategy**	**Priors**	**Likelihood**	**Bias**	**Posterior Estimation**
Booth et al. (2012)	Not used	N/A	N; B; F	Normal (Assumed)	Yes - GP	Not discussed
Cerezo et al. (2017)	Not used	N/A	U	Binary	No	Parametric
Hedegaard et al. (2019)	Not used	N/A	B; G; HC	Normal	No	MCMC (Metropolis)
Kristensen et al. (2017a)	GP	OEFEB	G; N; B; U	Normal	Yes - GP	MCMC (Metropolis-Hastings)
Kristensen et al. (2018)	Not used	N/A	MVN; NIW; U; HC	Normal	No	MCMC (Metropolis-Hastings)
Sokol et al. (2017)	PR	OEFEB	U	Binary	No	Parametric
Tardioli et al. (2020)	GLM; ANN; SVM;RF	OEFEB; OEFEC; OEFRB	Not-U	Normal	Yes - GP	MCMC (Metropolis-Hastings)
Wang et al. (2020)	Not used	N/A	U	Normal	No	Parametric

Disambiguation. Priors: N = Normal, B = Beta, F = Fréchet, U = Uniform, HC = Half-Cauchy, G = Gamma, MVN = Multivariate Normal, NIW = Normal-Inverse-Wishart, Not-U = Not Uniform; Surrogate Strategy: OEFEC = One Emulator For Each Cluster, OEFEB = One Emulator For Each Building, OEFRB = One Emulator For Representative Building; Surrogate Model: GP = Gaussian Process, PR = Polynomial Regression, GLM = Generalised Linear Models, ANN = Artificial Neural Networks, SVM = Support Vector Machines, RF = Random Forests.

Table 2.4: Comparison of the validation procedure, pre- and post-calibration performance for the studies reviewed.

Authors	**Validation Procedure**	**Performance - Training**		**Performance - Validation**	
		Pre-calibration	**Post-calibration**	**Pre-calibration**	**Post-calibration**
Booth et al. (2012)	T & V on the same data	PE: 17.6 %	PE: 0.5 %.	-	-
Cerezo et al. (2017)	For the same period, • T: 164 homes (51 %) • V: 159 homes (49 %)	PE: 4 %	PE: $<$ 1 %	-	PE: 3 %
Hedegaard et al. (2019)	For the same homes, • T: 1 month (Jan), • V: 1 month (Feb).	-	CVRMSE: 4.66 % NMBE: 0.08 %	-	CVRMSE: 5.58 % NMBE: -1.39 %
Kristensen et al. (2017a)	For the same period, • T: 450 homes (75 %) • V: 150 homes (25 %)	-	NMBE: -0.3 % MAPE: 20.0 % CVRMSE: 24.1 %	-	NMBE: 2.3 % MAPE: 21.9 % CVRMSE: 26.5 %
Kristensen et al. (2018)	• T: 50 homes (25 %) for 1 month (Jan) • V: 150 homes (75 %) for 1 month (Feb)	-	NMBE: -3.0 % CVRMSE: 7.2 %	-	NMBE: 2.9 % CVRMSE: 7.8 %
Sokol et al. (2017)	For the same period, • T: 399 homes (15 %) • V: 2263 homes (85 %)	PE: 54.8 %	Annual PE: 25.5 % Monthly PE: 19.4 %	PE: 69 %	Anual PE: 47 % Month PE: 44 %
Tardioli et al. (2020)	For the same period, • T: 326 homes (11 %) • V: 2646 homes (89 %)	-	OEFEB PE: 0.75 % OEFEC PE: 2.5 % OERFB PE: 1.34 %	-	OEFEB PE: 2.3 % OEFEC PE: 8.2 % OEFRB PE: 2 %
Wang et al. (2020)	For the same homes, • T: six years (75 %), • V: two years (25 %).	-	-	2016 PE: 25.0 % 2017 PE: 19.9 %	2016 PE: 8.3 % 2017 PE: 7.7 %

Disambiguation: T = Training, V = Validation, PE = Percentage Error, CVRMSE = Coefficient of Variation of Root Mean Square Error, NMBE = Normalised Mean Bias Error, MAPE = Mean Absolute Percentage Error, OEFEB = One Emulator For Each Building, OEFEC = One Emulator For Each Cluster, OEFRB = One Emulator For Representative Building.

2.4.5.5 Posterior Estimation

In four out of the eight studies, a Metropolis or Metropolis-Hastings MCMC algorithm was used to compute the posteriors (Table 2.3). A parametric approach was used in three papers, where the posteriors were estimated at fixed parameter values. Booth et al. (2012) did not explain how the posteriors were estimated in their analysis.

2.4.6 Predictive Performance

With the exception of Booth et al. (2012), who used the same data for model calibration and validation, all other authors used some of the data for training and some for validation (although they often evaluated the performance for the training data as well). Table 2.4 summarises the validation approach and results of the studies under review. Cerezo et al. (2017), Kristensen et al. (2017a), Sokol et al. (2017) and Tardioli et al. (2020) performed the calibration and validation over the same period but on different dwellings. On the other hand, Hedegaard et al. (2019) and Wang et al. (2020) carried out the calibration and validation on the same buildings but for two different periods. The split of training/validation data ranged from 75 % / 25 % to 11 % / 89 %. Kristensen et al. (2018) were the only authors who used two different groups of dwellings and periods for calibration and validation.

Where a comparison pre- and post-calibration was made, a clear improvement in predictive performance is observed, although most authors did not carry out such comparison. For many of the studies, the post-calibration error metrics were below commonly used thresholds,[12] deeming the models as "calibrated" (Ruiz and Bandera, 2017). While the errors reported by the authors vary significantly between studies, it is not possible to contrast the efficacy of each study's calibration method due to the differences in data and models used.

[12]The ASHRAE Guideline 14 thresholds for hourly values are CV(RMSE) = 30 % and NMBE = $\pm$10 %, and for monthly values are CV(RMSE) = 15 % and NMBE = $\pm$5 % (ASHRAE, 2002).

2.5 Summary & Research Gaps

As described in Section 2.1, indoor overheating is a complex phenomenon driven by warm weather, being intensified by climate change and substantially modified by building and urban characteristics, along with occupant behaviour (Kougionis, 2018; Lomas and Porritt, 2017; Mavrogianni et al., 2014). Its implications for individuals and society can be substantial (Section 2.1.2); these include reduced enjoyment of the indoor space, sleep quality, productivity, and increased risk of mortality (AECOM, 2019; Joshi et al., 2016; Lan et al., 2011). Mitigating the risk posed to human health, wellbeing and productivity from increased exposure to heat in homes and other buildings is one of the highest priorities for adaptation (CCC, 2021b). While positive steps have been taken to tackle such risks in new homes, there is a policy gap in adapting the existing housing stock to the increasing risk of indoor overheating (CCC, 2022).

Building stock models have long been used to guide energy policy (Kavgic et al., 2010; Oraiopoulos and Howard, 2022). Similarly, models adapted to predict summer indoor temperature, such as the bottom-up UK-HSM, can support the formulation of policies to mitigate indoor overheating risk. Uncertainties are integral to the modelling process (Saltelli et al., 2008), and their quantification and reduction can result in better-informed decision-making. The importance of uncertainties was recognised during the empirical validation of UK-HSM, with their quantification and reduction considered essential future work (Symonds et al., 2017). The Bayesian approach to model calibration has been shown to perform well and has gained traction within the field of built environment (Hou et al., 2021; Oraiopoulos and Howard, 2022). Differences exist between the various implementations of this approach, and in section 2.4, published examples of the Bayesian calibration of archetype-based housing stock models were critically evaluated.

Through the review of literature on indoor overheating, building stock modelling and the calibration of archetype-based housing stock models using Bayesian inference, several research gaps were identified. Pertinent to this thesis are the following:

1. All published examples of Bayesian calibration on archetype-based housing stock models focused on energy. At the time of writing, no such application on models of indoor temperature or overheating could be found. Given the growing importance of indoor overheating, and the differences in the dynamics of indoor temperature compared to energy use when used for calibration purposes, this is a key research gap.

2. Classification, an important step in the Bayesian calibration of archetype-based models, has often been done using an *ad hoc* basis. A data-driven approach has the potential to result in a better-informed classification process. In the limited examples where such an approach was implemented, household characteristics were not considered, and it was unclear how this approach linked with the subsequent calibration process.

3. The process of identifying probability distributions that represent available evidence on the possible values of influential model inputs is important for uncertainty quantification and Bayesian calibration. In the work reviewed, there was very limited discussion of this process, with uniform distributions often used despite the fact that they are unlikely to capture all available evidence.

The growing concerns surrounding indoor overheating adaptation and the aforementioned research gaps have motivated this doctoral study. The aim of thesis is to *quantify and reduce uncertainties of archetype-based housing stock models of summer indoor temperature*, with a discussion surrounding the work's scope and research objectives offered in Section 1.3. In the following chapter, the methods used to address the research aim and objectives are described. In addition, the datasets that are instrumental to this work are discussed and modelling details regarding UK-HSM are provided.

Chapter 3

Methods & Methodology

Chapter 2 presented a review of the theory and literature relevant to this work, focussing on indoor overheating, building stock modelling and model calibration. As highlighted in Chapter 1 and Chapter 2, there is an urgent need to adapt the UK housing stock to indoor overheating. A useful tool in supporting such efforts is archetype-based building stock modelling, with one such example being the UK Housing Stock Model (Section 2.2.2). Yet, as with any modelling endeavour, uncertainties are present and their quantification and minimisation through a calibration process is a necessary step in ensuring good predictive performance without an unrealistic level of confidence. For this reason, a Bayesian approach was selected, due to its ability to incorporate uncertainties within the calibration process. Several frameworks of Bayesian calibration of housing stock models were reviewed (Section 2.4), and several research gaps were identified (Section 2.5). Most importantly, no published example that focused on the Bayesian calibration of housing stock models of summer indoor temperature could be identified.

The purpose of this chapter is to describe the overall process through which the aim and research objectives of this thesis, set out in Section 1.3, will be achieved (Figure 3.1). In Section 3.1, the research objectives are revisited and an outline of the steps required to address them is provided. Subsequently, Section 3.2 introduces the Bayesian calibration framework developed in response to the first research objective of this work. The quantity of interest for this study, and arguments for this choice, are presented in Section 3.3. In Section 3.4, the key datasets used within this thesis are

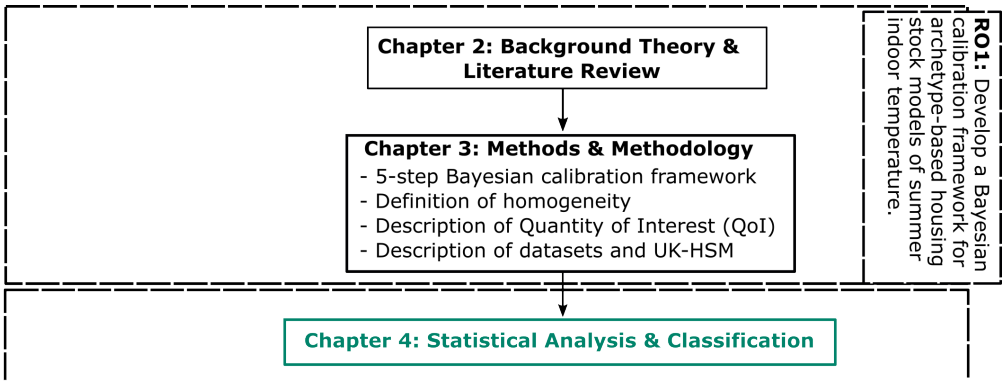


Figure 3.1: Chapter 3 flowchart. This is an abridged version of Figure 1.3, focusing on Chapter 3. RO1 is a shortened version of Research Objective 1. UK-HSM stands for UK Housing Stock Model.

described, while the modelling details of the UK-Housing Stock Model are provided in Section 3.5. The chapter concludes with a summary in Section 3.6

3.1 Addressing the Research Objectives

In Section 1.3, three research objectives were identified that collectively address the aim *to quantify and reduce uncertainties of archetype-based housing stock models of summer indoor temperature*. The rest of this section outlines the steps required to fulfil each research objective.

Research Objective 1: *To develop a Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature*. The following steps are required:

1. Review theory and literature on the Bayesian calibration of archetype-based housing stock models. This was presented in Sections 2.3.6–2.4.

2. Define what constitutes a *homogeneous* group of dwellings for this work.

3. Based on the chosen definition of homogeneity, propose a framework to:

    - carry out a data-driven classification of dwellings into homogeneous groups, and
    - select and calibrate uncertain, influential model inputs

Research Objective 2: *To quantify the uncertainty of the UK Housing Stock Model (UK-HSM) inputs with the greatest influence on summer indoor temperature for a single homogeneous group of dwellings.* Addressing this research objective requires implementing the framework developed in response to the first object. In particular, the steps that focus on classification and influential model input selection:

1. Select the model output of interest (also referred to as Quantity of Interest - QoI).
2. Select a homogeneous group of dwellings following an analysis of empirical observations of the QoI and linked dwelling and household characteristics.
3. For the UK-HSM archetype model that corresponds to the selected homogeneous group of dwellings, identify a probability distribution for each continuous model input. This should be done using empirical data where possible, or based on theoretical assumptions in their absence. These distributions represent the uncertainty for each input.
4. Identify the model inputs, where given their uncertainty, have the greatest influence on the QoI.

Research Objective 3: *To quantify the level of improvement in the predictive ability of the UK Housing Stock Model following application of the Bayesian calibration framework and reduce model input uncertainty for a single homogeneous group of dwellings.* The following steps are required:

1. Choose a method to aggregate the empirical observations to be used in the Bayesian calibration process.
2. Select a surrogate model and a method to computationally implement the Bayesian calibration.
3. Use a subset of the monitored data (train set) of the chosen homogeneous cluster to calibrate the equivalent archetype UK-HSM model. The calibration variables and their priors will be informed by the work carried out in addressing the second research objective.

4. The predictive performance of the calibrated model will be quantified using the unseen part of the field data (test set).

3.2 The Bayesian Calibration Framework

In response to the first research objective, a Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature was developed. The framework is modular and flexible; a modeller can choose what methods to use for the different steps depending on the data available, their model and preference. A high-level description of the framework is offered within this section. Details regarding the framework's implementation in this thesis are provided in the methods sections of Chapters 4–7.

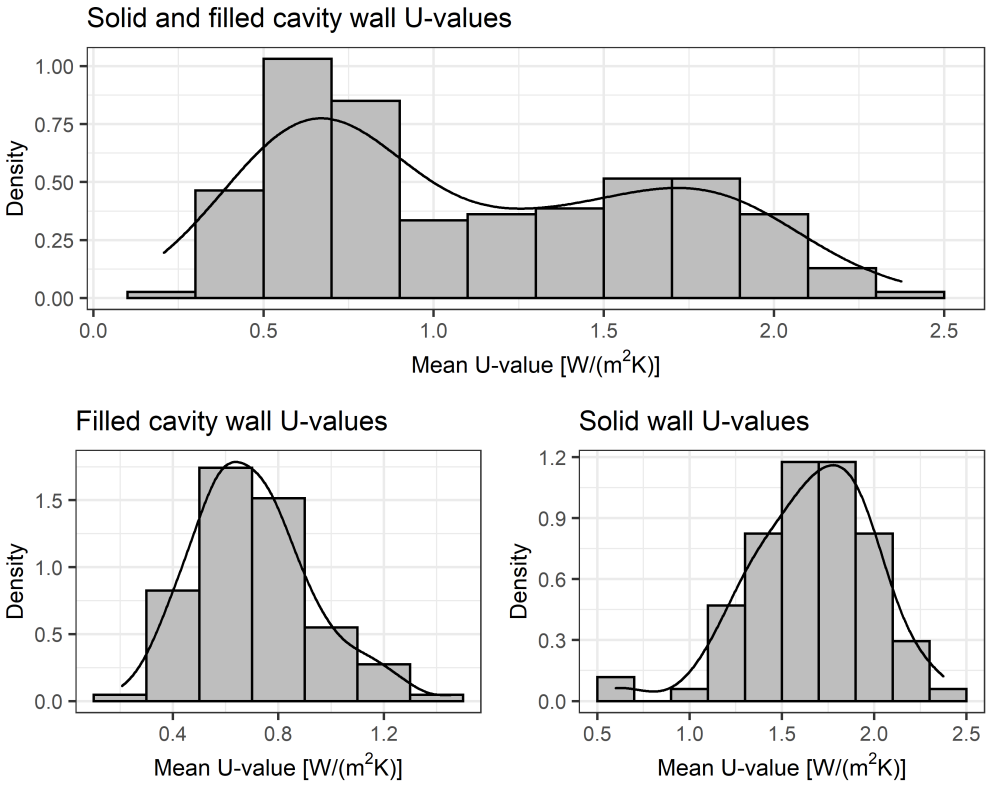


Figure 3.2: Distributions of wall U-value, before (top) and after (bottom) segmenting the measurements based on the wall type. Data from Hulme and Doran (2014).

As with the studies reviewed in Section 2.4.2, a prerequisite to the Bayesian calibration of archetype-based housing stock models is the classification (also referred to as segmentation or clustering) of dwellings into homogeneous groups. This

step is necessary to reduce the effect of uncontrollable parameters on the calibration process and the misattribution of uncertainty. Contrary to most papers reviewed, a data driven approach to classification is preferred, which relies on a clear and practical definition of homogeneity.

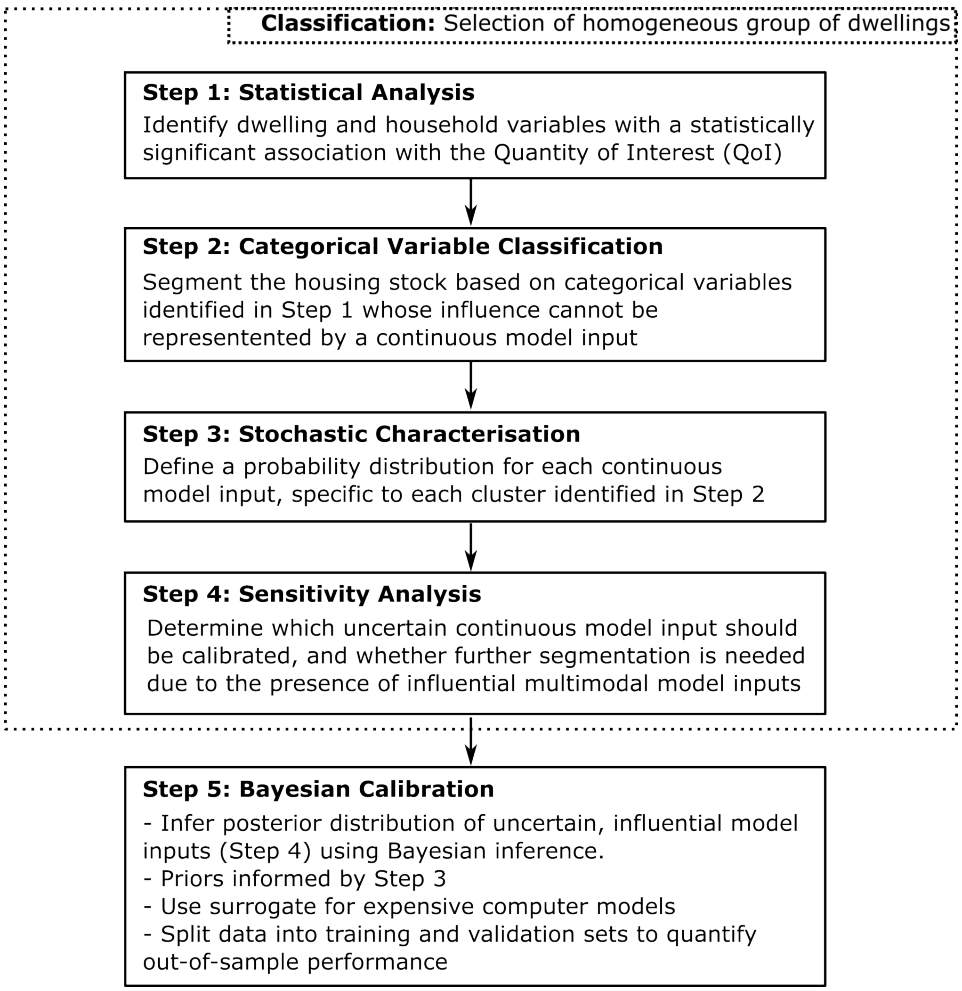


Figure 3.3: Workflow diagram for Bayesian calibration framework.

In this work, a group of dwellings is considered homogeneous *if the variability of influential building parameters could be described by unimodal distributions*. Only the most influential parameters are considered, to avoid an excessive segmentation of the building stock based on parameters that would not largely influence the QoI. In addition, it is advisable to calibrate only the most influential parameters to reduce computational cost (since a greater number of parameters would require a greater number of training points) but also to reduce the risk of parameter identifiability,

as discussed in Section 2.3.8. If a building parameter within a group of dwellings is described by a multimodal distribution, it is likely that one or more factors exist that are responsible for the distinct modes. For example, the wall construction type is the factor largely responsible for the multimodal distribution of wall U-values in Figure 3.2. Segmenting the group of dwellings based on the identified factor and carrying out the calibration separately for each group offers a few benefits:

1. It facilitates a comparison between groups of dwellings whose distinct characteristics, such as those associated with different wall type construction, might be the result of a change in building regulations or common industry practice. A comparison of the posterior distributions of building parameters could reveal to what extent a change in construction practice has had the desired effect on the building parameter, for example, a reduction in wall U-value. This could be particularly useful for building parameters that are hard to measure, although caution should be used when interpreting such results from complex models and with limited data (Booth et al., 2012; Kennedy and O'Hagan, 2001). In addition, a comparison of QoI between groups could determine the effect of a change in construction practice or building regulations on the QoI.
2. Furthermore, as argued by Booth et al. (2012), segmenting the housing stock based on key factors also make the outcomes of a Bayesian calibration analysis more informative to policymakers who might be interested in applying a measure to a group of similar dwellings.
3. Finally, a practical consideration is that popular Markov Chain Monte Carlo (MCMC) algorithms used in Bayesian model calibration, including Hamiltonian Monte Carlo discussed in Section 2.3.8, do not perform as well when faced with multimodal posteriors (Pompe et al., 2020; Yun et al., 2020).[1]

Following from the definition of a homogeneous group, a Bayesian calibration framework was developed and is visualised in Figure 3.3. Each step of this process is described in more detail below:

[1]Although it is worth noting that MCMC algorithms are being developed to tackle multimodal posteriors (Pompe et al., 2020; Yun et al., 2020).

Step 1: Statistical analysis. The aim of this step is to determine which variables have a statistically significant association with the QoI. This process relies on the use of empirical observations of the QoI and matched metadata, such as dwelling and household characteristics. The QoI might need to be estimated from raw observations. For example, the daily maximum indoor temperature could be derived from hourly observations of indoor temperature. The metadata, which in the statistical analysis are treated as explanatory variables, could be continuous or categorical. Bivariate and multivariate methods of analysis may be used to establish whether any such associations exist.

Step 2: Categorical variable classification. Following the statistical analysis, the housing stock is segmented based on statistically significant categorical variables whose effect cannot be captured by a continuous model input. This step must take into consideration the model being calibrated, and the empirical data available for calibration. For example, assume *dwelling type* and *floor area* were both shown to be associated with the QoI at a statistically significant level. In this example, floor area can be specified in the archetype model as a continuous model input. Thus, even if it was a categorical variable in Step 1, it does not need to be used as a classifier and the housing stock data will not be segmented based on this model input. On the other hand, the housing stock would need to be clustered based on the dwelling type, since this is not modelled as a continuous model input. If the model has yet to be developed, or further development is desired, Step 1 could inform this process.

Step 3: Stochastic characterisation. For each cluster, a probability distribution is defined for each continuous model input. Where possible, the probability distribution functions should be informed by empirical data; this process is not constrained to the dataset used in Steps 1 and 2. Methods for identifying appropriate distributions depending on the data available are described in Section 5.1.

Step 4: Sensitivity analysis. This step has two main aims: 1) To determine which of the uncertain continuous model inputs 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, as discussed in Section 2.3.8. Several techniques exist that could accomplish this task, with the Morris method most frequently used (Hou et al., 2021). The second aim is specific to this framework and is based on the proposed definition of homogeneity. To implement this step, the chosen method of sensitivity analysis should first be used to rank model inputs when they are allowed to vary within their feasible range (e.g. their 95 % or 99 % percentile interval); this will be informed by the work carried out in Step 3. Subsequently, if any influential variables are described by multimodal distributions, the housing stock is further segmented for each mode. A probability distribution will need to be identified for each mode and the sensitivity analysis will be repeated for the newly formed clusters. Influential model inputs characterised by a unimodal distribution can be calibrated. Fixed values may be used to describe non-influential model inputs.

Step 5: Bayesian calibration. The aim of this step is to use Bayes' theorem to combine empirical observations of the QoI with any prior knowledge about the distributional form of the uncertain calibration variables, and infer the posterior distributions of calibration variables. The implementation will depend on a few factors, such as, whether the model is computationally expensive, or what method is used to compute the posteriors:

- For a computationally expensive model, a surrogate model would need to be trained based on model simulations, which would subsequently (or concurrently) be calibrated (Higdon et al., 2004). A fast computer model could instead be calibrated directly.
- An MCMC is the most frequently used method for computing the posterior distributions (Hou et al., 2021), and the Hamiltonian Monte Carlo algorithm has been shown to be promising (Chong et al., 2017).
- Another important consideration in the calibration of archetype-based models is the choice of aggregation of empirical observations of the QoI from different dwellings (see Section 2.4.5.3). Related to this is also the choice of likelihood model used (see Section 2.4.5.2.
- The choice of priors is also crucial; this could be informed by Step 3.

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 the out-of-sample predictive ability of the model following calibration, part of the empirical observations should be reserved for validation. While the predictive performance post-calibration is expected to improve, the posterior distributions of model inputs (e.g. the wall U-value) should not be treated as direct estimates of the real-world physical quantities. As Kennedy and O'Hagan (2001) argues, this is "inevitable in calibration when we do not believe that the model can ever be a perfect fit", which is the case when modelling a complex system – such as an occupied building or groups of buildings – where several simplifying assumptions are necessary. In addition, selecting only the most influential model inputs for calibration can result in the lumping of uncertainty from uncalibrated parameters into the posteriors, further diluting the physical meaning of the calibrated parameters (Booth et al., 2012). The extent to which the posterior distributions are representative of the physical quantities cannot be verified in most applications, since a ground truth[2] about the physical quantities does not exist, unless a study is concerned with synthetic or test cell data.

Ideally, the dataset which informs the statistical analysis, classification and calibration is large and representative of the building stock. The dataset would be considered representative if the proportions and correlations of dwelling and household parameters within the dataset (sample) are similar to those expected of the housing stock (population). A sample would be sufficient in size if there is enough data for each cluster to be calibrated. No guidance could be identified on what number this should be, with Booth et al. (2012) having used 35 similar dwellings in a single cluster while Kristensen et al. (2017a) used 450. Often, neither of the aforementioned requirements will be met, which will require pragmatic decision-making and caution when generalising from the results. For instance, in the first stage of the classification, the influence of some categorical variables may not be

[2]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* 2023)

taken into account to ensure the clusters are large enough to conduct archetype-based calibration. It should be noted that through the proposed classification procedure, correlations between dwelling and household categorical variables and the QoI are captured.

3.3 Quantity of Interest

To implement the Bayesian calibration framework, a quantity of interest must be specified based on the calibrated model's intended use. In this thesis, two such quantities were selected for Step 1 of the calibration framework, but only one was used for Steps 2-5. The choice of these quantities 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 (Taylor et al., 2015; Taylor et al., 2018a; Taylor et al., 2018b). The two quantities selected were:

- **The Mean of the Daytime Living Room Temperature (MDLRT)**: The mean of the hourly living room temperature was estimated per day between 08:00-22:00.
- **The Mean of the Nighttime Bedroom Temperature (MNBT)**: The mean of the hourly bedroom temperature was estimated per day between 22:00-08:00.

The use of these variables for heat-mortality modelling stems from the work of Taylor et al. (2015). While several studies have derived epidemiological relationships between excess heat-related mortality and ambient air temperature (Armstrong et al., 2010; Hajat et al., 2014; Vicedo-Cabrera et al., 2021), a similar set of relationships for indoor temperature do not currently exist. To account for the modifying effect of building characteristics on excess heat-mortality, and as discussed in Section 2.1.3.2, Taylor et al. (2015) assumed the relationship between daily maximum temperature and excess mortality above the temperature-mortality threshold to be the same for the living room as that suggested by Armstrong et al. (2010) for the ambient temperature. Later heat-mortality work by Taylor et al. relied on the daytime (08:00-22:00) living room temperature based on the UK-HSM metamodels developed Symonds et al. (Symonds et al., 2016; Taylor et al., 2018b; Taylor et al., 2021).

During Step 1, statistical analysis was carried out to identify dwelling and household characteristics that have a statistically significant association with summer indoor temperature. The analysis was conducted for both the living room (MDLRT) and bedroom (MNBT) metrics, since outputs of this analysis would contribute to current knowledge of indoor overheating risk and could potentially have implications to industry or policy, in addition to informing the rest of the calibration procedure.

For Steps 2-5, only MDLRT was used for the calibration. The focus on living room temperature is in line with previous heat-related mortality work based on UK-HSM (Taylor et al., 2015; Taylor et al., 2018a; Taylor et al., 2018b). However, contrary to these studies, the daily mean of the living room temperature was used instead of the daily maximum. Even after data cleaning, it is still possible for the daily maximum of hourly (or sub-hourly) indoor temperature observations to be biased by the data logger's exposure to brief yet substantial levels of heat (e.g. direct sunlight). The daily mean is less sensitive to such exposure. Moreover, there is no strong evidence to support the use of maximum instead of mean daily indoor temperatures, and epidemiological relationships between ambient temperature and excess mortality have been established for daily mean and maximum temperatures (Armstrong et al., 2010; Hajat et al., 2014; Vicedo-Cabrera et al., 2021).

3.4 Datasets

To implement the proposed Bayesian calibration framework in this thesis, two datasets of monitored indoor temperatures were used: (i) The 2011 Energy Follow-Up Survey (EFUS) for Step 1 (statistical analysis) and (ii) The 4M dataset for Step 5 (Bayesian calibration). A summary of the datasets' key characteristics is provided in Table 3.1, with further information in Section 3.4.1 for EFUS and Section 3.4.3 for 4M. Section 3.4.2 explains why two datasets were used.

3.4.1 EFUS & EHS

The English Housing Survey (EHS) is a national survey, commissioned by the Department for Levelling Up, Housing and Communities (previously Ministry of Housing Communities and Local Government, and Department for Communities and

Table 3.1: Key characteristics of the two datasets of monitored indoor temperature used in this study, the Energy Follow-Up Survey (EFUS) and the 4M. For EFUS, information is based on Hulme et al. (2013a), for 4M on Lomas and Kane (2013) and Tempcon Instrumentation Ltd (2022).

	EFUS	4M
Number of homes*	823	193
Start date	Dec 2010 - April 2011	July 2009
End date	April-May 2012	March 2010
Measuring frequency	20-min	60-min
Rooms monitored	Living room; hallway; main bedrooom	Living room; main bedroom
Data Logger	TinyTag Transit 2	Onset HOBO UA-001-08
Data Logger's accuracy	$\pm 0.2\,^{\circ}$C	$\pm\ 0.4\,^{\circ}$C**
Data Logger's resolution	$0.01\,^{\circ}$C	$0.14\,^{\circ}$C at $25\,^{\circ}$C
Data Logger's range	$-70\,^{\circ}$C to 40°C	$-20\,^{\circ}$C to $70\,^{\circ}$C

* This is the number of homes for which data was available to use in this study
** The value $\pm\ 0.4\,^{\circ}$C, quoted by Lomas and Kane (2013) is different to that suggested by the manufacturer (Tempcon Instrumentation Ltd, 2022). Since the accuracy is temperature-dependent, the value of $\pm\ 0.4\,^{\circ}$C is possibly over a narrower temperature range than the value offered by Tempcon Instrumentation Ltd (2022) ($\pm\ 0.53\,^{\circ}$C from $0\,^{\circ}$C to $50\,^{\circ}$C).

Local Government), that takes place every two years and consists of household interviews and physical surveys (DLUHC, 2021). The interviews typically cover topics such as demography, employment and income while the surveys gather information regarding the dwelling conditions and energy efficiency measures.

As a follow-up to the 2010-11 EHS, the Energy Follow-Up Survey conducted further interviews and surveys in 2,616 dwellings with the purpose of updating modelling assumptions regarding how energy is used at home (Hulme et al., 2013a). For 943 dwellings, the indoor air temperature was monitored using TinyTag Transit 2 loggers at 20-minute intervals in the living room, bedroom and hallway from the time of installation (December 2010 to April 2011) until they were returned (April/May 2012), with the data stored internally. As summarised in Table 3.1, the loggers had an accuracy of $\pm 0.2\,^{\circ}$C, a resolution of $0.01\,^{\circ}$C and a temperature range of $-70\,^{\circ}$C to 40°C (Hulme et al., 2013a). Although the loggers were new and calibrated by the manufacturer, BRE conducted additional tests on a sample of them, with their performance being as expected. Occupants were provided with instructions on the

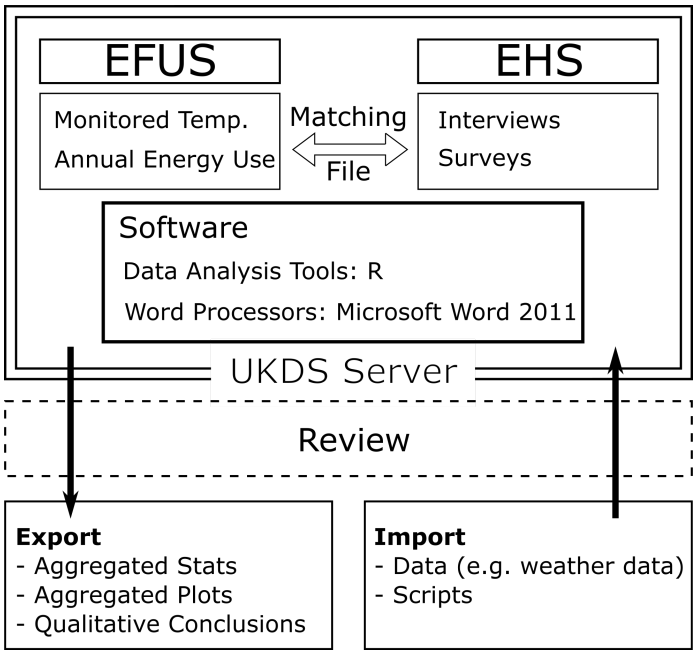


Figure 3.4: Flowchart of the process used to upload and download information off the UK Data Service server.

correct placement of the data loggers during the interview, on internal walls away from direct sources of heat or sunlight and at a height that can be reached by adults but not by children. Adequate data for at least one room were returned by 823 dwellings.

The monitored temperatures, interviews and survey data can be linked to the data within the EHS through access to the UK Data Service (UKDS) (Building Research Establishment and Department of Energy and Climate Change, 2016a; Building Research Establishment and Department of Energy and Climate Change, 2016b; Department for Communities and Local Government, 2017).

3.4.1.1 Data Access

A key consideration is the process of accessing and using the matched EHS and EFUS datasets (Figure 3.4). The two datasets, along with a matching file that allows the two datasets to be linked, are stored by the UKDS. Secure access to the data for the purpose of this research was obtained following successful training to become an Economic and Social Research Council (ESRC) Accredited Researcher. The data is stored on a remote server, which could only be accessed through a specific computer

and desk within UCL that were approved by UKDS during the application process. One could only use the software installed on the UKDS server for analysis, with the software used shown in Figure 3.4. Exporting or importing has to be reviewed and approved by UKDS staff. Exporting was particularly restrictive as only specific formats are allowed (e.g. manuscripts) with the results being either qualitative or aggregated quantitative statistics to satisfy a set of Statistical Disclosure Control (SDC) checks in order to reduce the risk of identifying individuals within the study.

3.4.1.2 Data Cleaning

At the pre-processing stage, the monitored indoor temperatures of each dwelling were analysed with the purpose of identifying extreme values that could be the result of faulty or misplaced data loggers (e.g. positioned near heat sources). Given the relatively cool conditions during the summer of 2011, individual recordings that exceeded 40 °C were removed and the temperatures measured at 20-minute intervals were averaged to give hourly values. In the case that multiple recordings exceeded 40 °C, that logger was removed from the dataset. Subsequently, for each region, the temperature profiles of statistical outliers were qualitatively assessed to determine whether further elimination was required (e.g. in case of year-long flat temperature profiles). In the case of missing data from bedroom or living room loggers during the period May-September (inclusive), the rooms of these dwellings were not included in the overheating assessment. Following the pre-processing stage, the temperature monitored in 795 living rooms and 799 bedrooms were considered adequate for analysis, out of an initial sample of 823 dwellings.

3.4.2 The Impact of Covid-19 on Data Access

The intention at the start of this study was to use the 2011 EFUS dataset for both classification and calibration. While the summer of 2011 was not particularly warm, the level of detail provided by this dataset together with its representativeness of the English housing stock were considered ideal. As explained in Section 3.4.1.1, access to this dataset was obtained through UKDS, with several restrictions in place, including the physical location of the computer being used to access the data on the

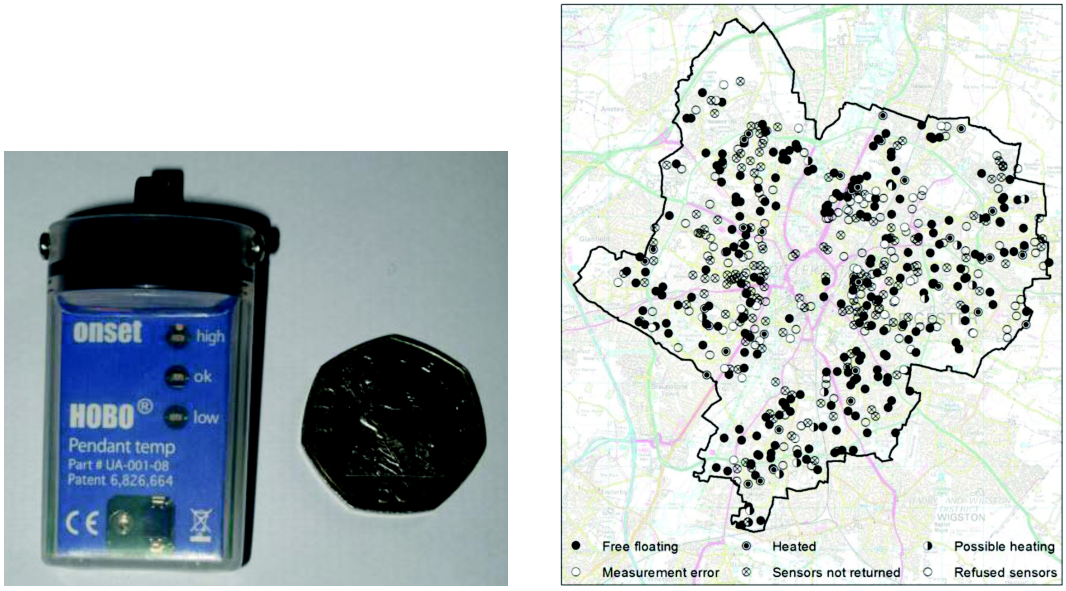
remote server. On the 17th of March 2020, due to the increasing number of people with SARS-CoV-2 within the UK, UCL requested its academic and research staff to work from home; something that became official guidance from the UK government a few days later. The building that houses the Bartlett School of Environment, Energy and Resources remained shut for several months, with restricted access being reinstated in September 2020.

Working from home meant that accessing the 2011 EFUS dataset or any analysis stored on the UKDS server was no longer possible. Given the uncertain nature of the pandemic, the author decided that the best course of action was to use a different dataset for the calibration, the 4M study, for which the steps of data exploration and cleaning had to be repeated. As detailed in Section 3.4.3, 4M was an empirical study which focused on the city of Leicester and contained less information on the monitored dwellings' characteristics and occupancy than EFUS. Yet, the use of 4M allowed for an effective application of the calibration framework developed and for the research objectives outlined in Section 1.3 to be met.

3.4.3 The 4M dataset

The core aim of the 4M project was to determine the carbon footprint of Leicester (Lomas and Kane, 2013). At the time, Leicester was the UK's 15th largest city, with households that covered a wide range of socio-economic categories. Face-to-face questionnaires were completed by 575 homes, which were randomly selected following stratification based on the percentage of detached homes and the percentage of homes with no dependent children. The questionnaire was compiled by the 4M team and was carried out on its behalf by the National Centre for Social Research (NATCEN). The survey included questions regarding the dwelling type, number of occupants, age of the oldest occupant and whether loft or wall insulation were present.

As part of the survey, occupants of all 575 homes were asked to place temperature sensors in the living room and main bedroom, with guidance provided on their correct installation; away from heat sources and not in direct sunlight. Out of the 575 households, 94 did not agree to the use of data loggers and only 321 returned

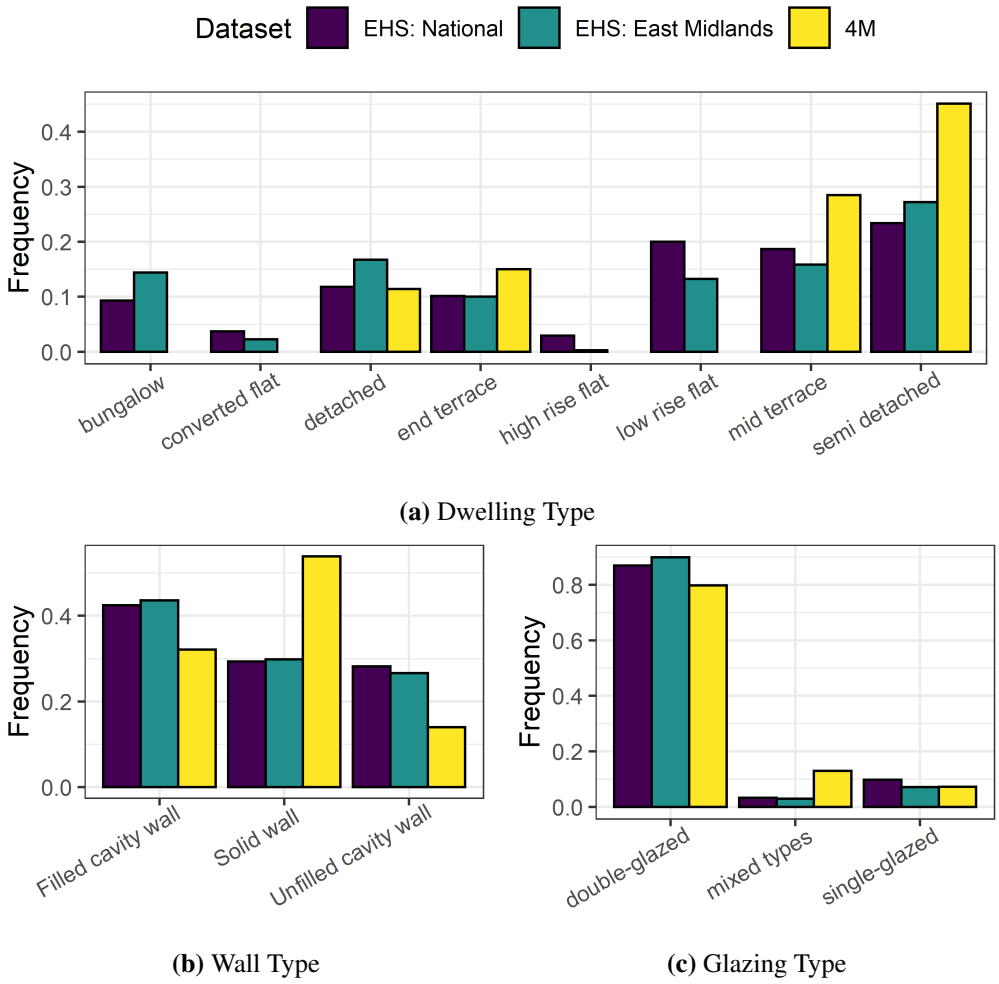


(a) **(b)**

Figure 3.5: Reproduced from Lomas and Kane (2013), (a) is an image of the Hobo data logger used to monitor indoor temperatures and (b) is a map of Leicester showing the spatial distribution of homes that took part in 4M.

them at the end of the monitoring period (Figure 3.5(b)). A subset of the 4M dataset (193 homes) with adequate metadata required for the calibration part of this doctoral study were kindly provided in an anonymised format by Prof. Kevin Lomas and Dr. David Allinson.

Table 3.1 offers a summary of the temperature monitoring. Onset HOBO UA-001-08 pendant-type temperature sensors (Figure 3.5(a)) were used to monitor internal temperatures over an eight-month period, beginning on the 1st of July 2009 (Lomas and Kane, 2013; Tempcon Instrumentation Ltd, 2022). Each hour, the sensors took spot measurements of air temperature, but it is likely that they also recorded part of the radiant component since they were not shielded (Lomas and Kane, 2013). While it cannot be claimed that such observations are of air temperature alone, it has been argued by Lomas and Porritt (2017) that temperature measurements which are a function of air and radiant temperature may correspond better to the temperature experienced by the occupants. The loggers were calibrated by the manufacturer and were found to be accurate to $\pm$ 0.4 °C (Lomas and Kane, 2013).



(a) Dwelling Type

(b) Wall Type

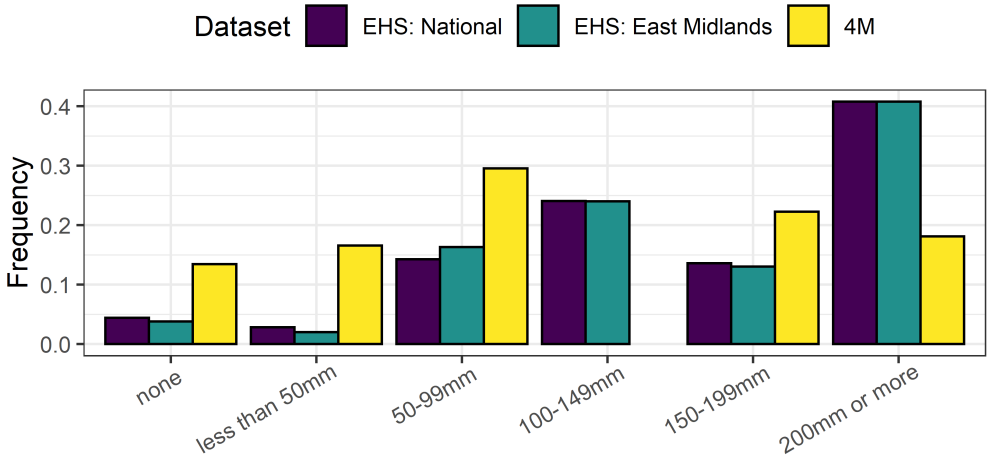
(c) Glazing Type

Figure 3.6: Bar charts comparing the prevalence of different dwelling, wall and glazing types within the entire 2012 English Housing Survey (EHS), the subset of EHS dwellings located in East Midlands and the 4M dataset. Mixed glazing types indicate the presence of single and double glazing.

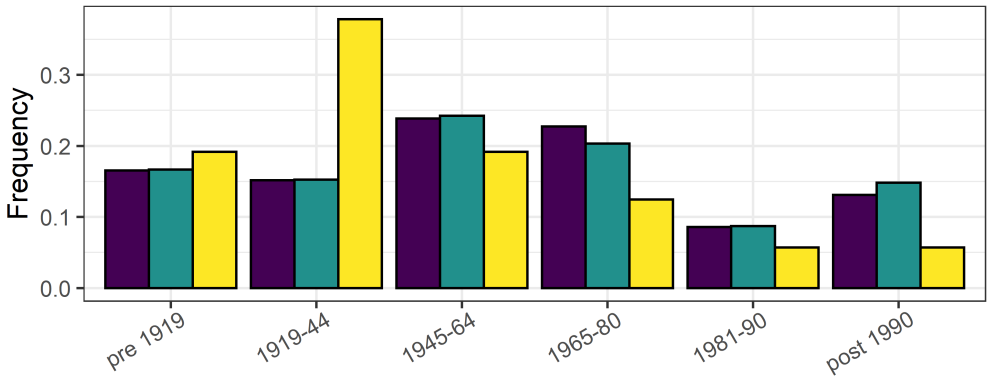
3.4.3.1 Exploratory Data Analysis

Figures 3.6–3.8 compare the prevalence of different dwelling and household characteristics within the 4M dataset, the entire 2012 English Housing Survey (EHS) and the East Midlands subset of EHS (EHS-EM).

Figure 3.6(a) indicates a higher prevalence of mid-terrace and semi-detached homes in 4M compared to EHS or EHS-EM, while there is a smaller prevalence of bungalows and flats. Semi-detached is the most frequently occurring dwelling type within the 4M dataset, with monitored temperatures and sufficient metadata



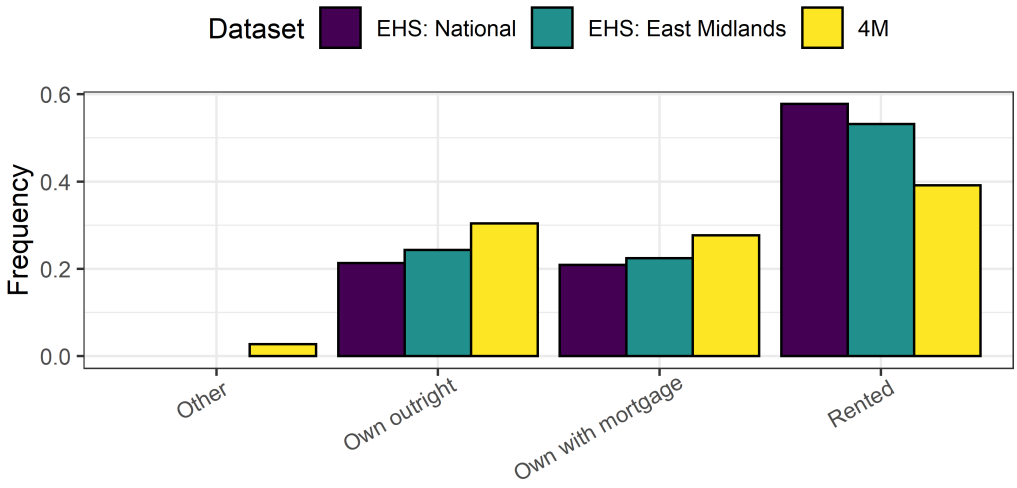
(a) Loft Insulation



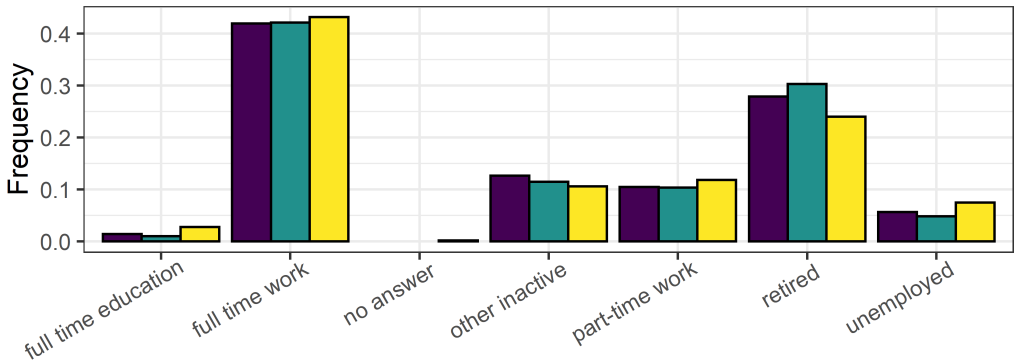
(b) Construction Age

Figure 3.7: Bar charts comparing the prevalence of different loft insulation levels and construction age within the entire 2012 English Housing Survey (EHS), the subset of EHS dwellings located in East Midlands and the 4M dataset.

available for 86 homes. The percentage of solid wall dwellings monitored by the 4M project is higher than the equivalent percentage in the EHS and EHS-EM datasets (Figure 3.6(b)). At the same time, cavity wall dwellings are less prevalent in the 4M dataset than in the EHS and EHS-EM datasets. The high prevalence of semi-detached and mid-terrace dwellings, along with the high prevalence of solid wall dwellings, are likely linked to the high percentage of pre-1944 dwellings within the 4M dataset (Figure 3.7(b)). While Figure 3.6(c) demonstrates that double glazing is the dominant type of glazing for all three datasets, Figure 3.7(a) suggests large discrepancies in the levels of loft insulation between the datasets. It is more common



(a) Tenure



(b) Employment

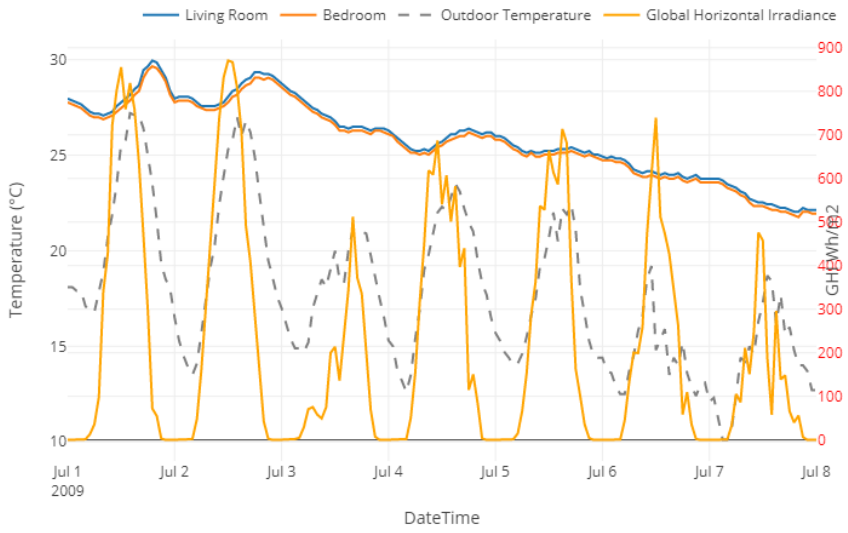
Figure 3.8: Bar charts comparing the prevalence of different tenure and employment groups within the entire 2012 English Housing Survey (EHS), the subset of EHS dwellings located in East Midlands and the 4M dataset.

within the 4M dataset to have insulation thickness less than 100 mm but less common to have insulation thickness of 200 mm or more than either of the EHS and EHS-EM datasets. The percentage of occupants who own their household is higher within the 4M dataset than in EHS and EHS-EM (Figure 3.8(a)). Employment status is similar amongst all three datasets (Figure 3.8(b)).

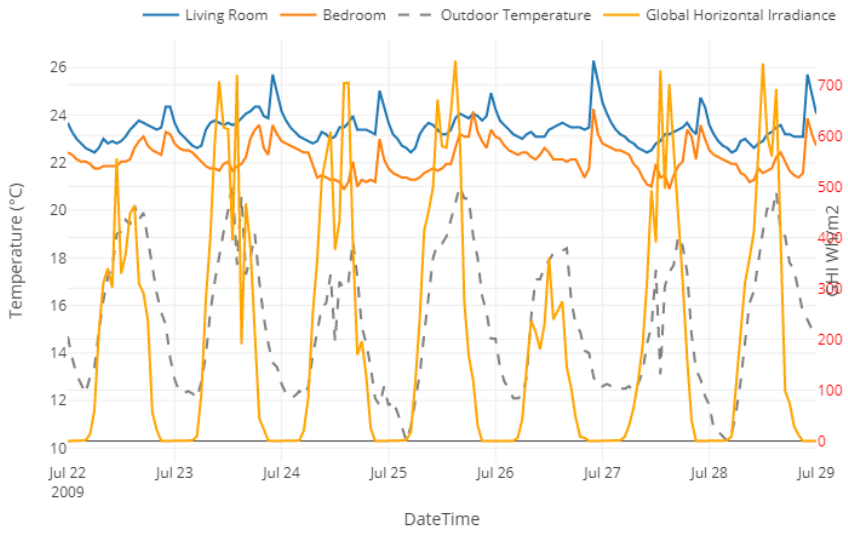
Following this comparison, it may be concluded that in many regards, the 4M dataset is not representative of the national or the East Midlands housing stock. Lomas and Kane (2013) considered the 4M dataset as largely representative of the city of Leicester. However, it is assumed that the findings from the statistical analysis of EFUS, described in Section 4.1, may still inform the classification process, even

if 4M is used for the calibration.

3.4.3.2 Data Cleaning



(a) Temperature sensors placed in the same room.



(b) Heating is likely on at 21:00.

Figure 3.9: Examples of timeseries plots of temperature profiles used for data cleaning.

Since the raw 4M data were provided for this study, data cleaning had to be carried out before being used. The data cleaning process was based on the approach of Lomas and Kane (2013). For each semi-detached home, a set of interactive, weekly timeseries plots of hourly indoor temperature for both rooms, outdoor ambient temperature and Global Horizontal Irradiance (GHI) were generated

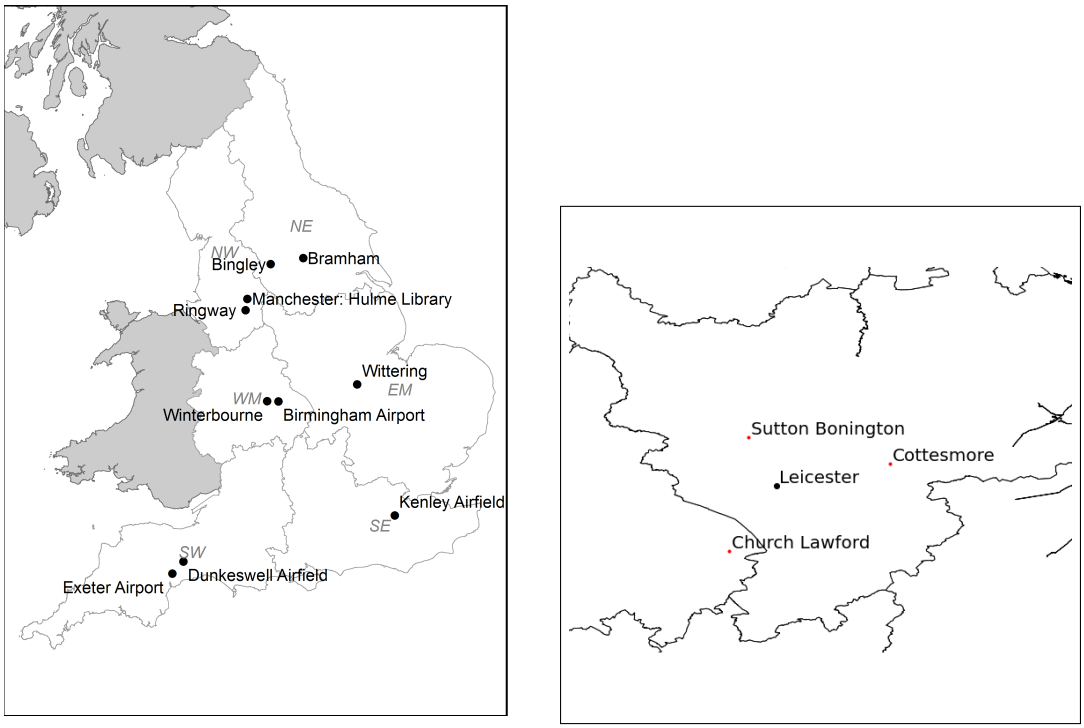
using the R library **plotly** (Sievert, 2020). An example is shown in Figure 3.9. By visually inspecting the plots of each home, data were rejected if any of the following conditions were true:

- **Sensors placed together**: identical temperature profiles for the living room and bedroom (Figure 3.9(a)).
- **Sensors moved location**: step changes in temperature profiles.
- **Sensors placed outdoors**: indoor temperature profiles closely matching the outdoor ambient temperature.
- **Sensors with direct solar radiation**: indicated by extreme increases in indoor temperature, correlated with measurements of GHI.
- **Sensors placed in a container, cupboard or drawer**: unresponsive (flat) temperature profiles.
- **Periodically unoccupied homes**: indicated by periods of different indoor temperature behaviour compared to the rest of the time. For example, this condition might be indicated by a period of unresponsive and comparatively low indoor temperatures, whereas a more dynamic indoor temperature behaviour was observed during the rest of the time.
- **Heated homes**: indicated by periodic increase of indoor temperature that diverged from the ambient outdoor temperature. An example is shown in Figure 3.9(b), where indoor temperature increased in both rooms, every day, at about 21:00.

Heated homes were excluded from the dataset used for calibration, since UK-HSM assumes no heating during the summer period. While this is true for most homes, a small percentage of occupants utilise heating during the summer and future version of the housing stock model should account for that (Lomas and Kane, 2013). Only the period July–August was used for calibration, since the data cleaning process revealed that a significant fraction of homes used heating in September. Data cleaning was only carried out for the final three clusters of dwellings considered for calibration, with its outcomes summarised in Figure 4.4 of Chapter 4.

3.4.4 Weather Data

Weather data already available from a previous study that compared the UK-HSM prediction to the EFUS dataset were used (Symonds et al., 2017). As described in more detail by Symonds et al., the weather data were obtained from the Met Office Integrated Data Archive System (MIDAS) database for the weather stations in the six regions identified in Figure 3.10(a) (Met Office, 2018b). Since not all weather stations had a complete dataset, a second weather station was sometimes used within the same region to fill the gaps. A summary of the daily-mean temperature of each region is provided in Figure 3.11. Depending on its Government Office Region (GOR), each dwelling was associated with one of the six regions.



(a) **(b)**

Figure 3.10: Part (a), reproduced from Symonds et al. (2017), maps the location of the weather stations used in the analysis of the EFUS dataset. Part (b) shows the location of the weather stations used for the 4M analysis.

For the analysis of the 4M dataset, MIDAS data from three weather stations, shown in Figure 3.10(b), were used to construct the weather file needed for the UK-HSM simulations (Met Office, 2018b). The stations were selected based on data availability and their proximity to the centre of Leicester (coordinates = 52.634444,

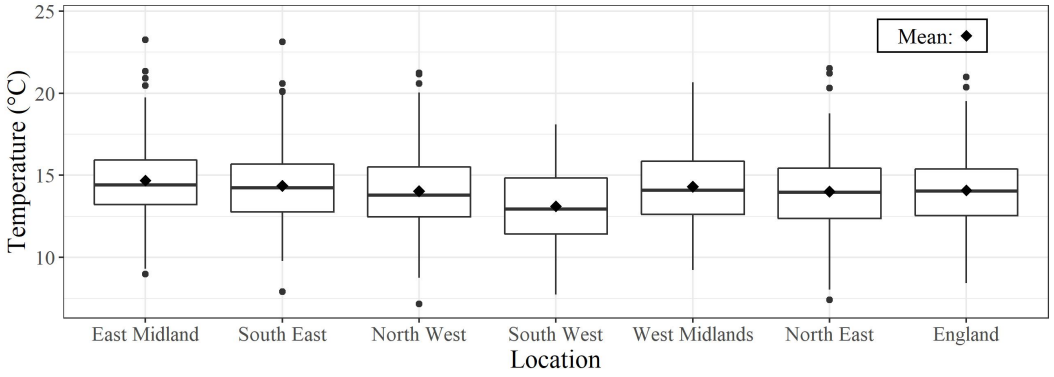


Figure 3.11: Box plots of the daily mean temperature between May and September 2011 for six English regions. The England box-plot represents the average daily mean temperature across the six regions. Data provided by Symonds et al. (2017).

-1.131944). 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.

The period of interest is the 1st of July to 31st of August 2009 when indoor temperature measurements were available, and heating was considered 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 and August of 16.2 °C and 16.6 °C, below the 10-year average. From the 28th of June until the 2nd of July, a heatwave resulted in the average daily temperature exceeding 19 °C and peaking on the 1st of July at 24.1 °C. There was one more occasion when the daily mean ambient temperature exceeded 19 °C on the 19th of August.

3.5 UK Housing Stock Model

In Section 2.2.2, UK-HSM was introduced, and the timeline of its development and applications was described. This section focuses the main modelling characteristics

of UK-HSM, as used in this work. While this study sought to calibrate UK-HSM's continuous model inputs, it did not attempt to alter the model's structure and hard-coded assumptions. The main component of UK-HSM is a UCL IEDE in-house parametric tool, EPGenerator 3 (EPG3), written in Python. While a thorough description of EPG3 and UK-HSM is provided by Mavrogianni et al. (2014), Symonds et al. (2016), Taylor et al. (2016) and Taylor et al. (2019), the rest of this section will provide a brief summary.

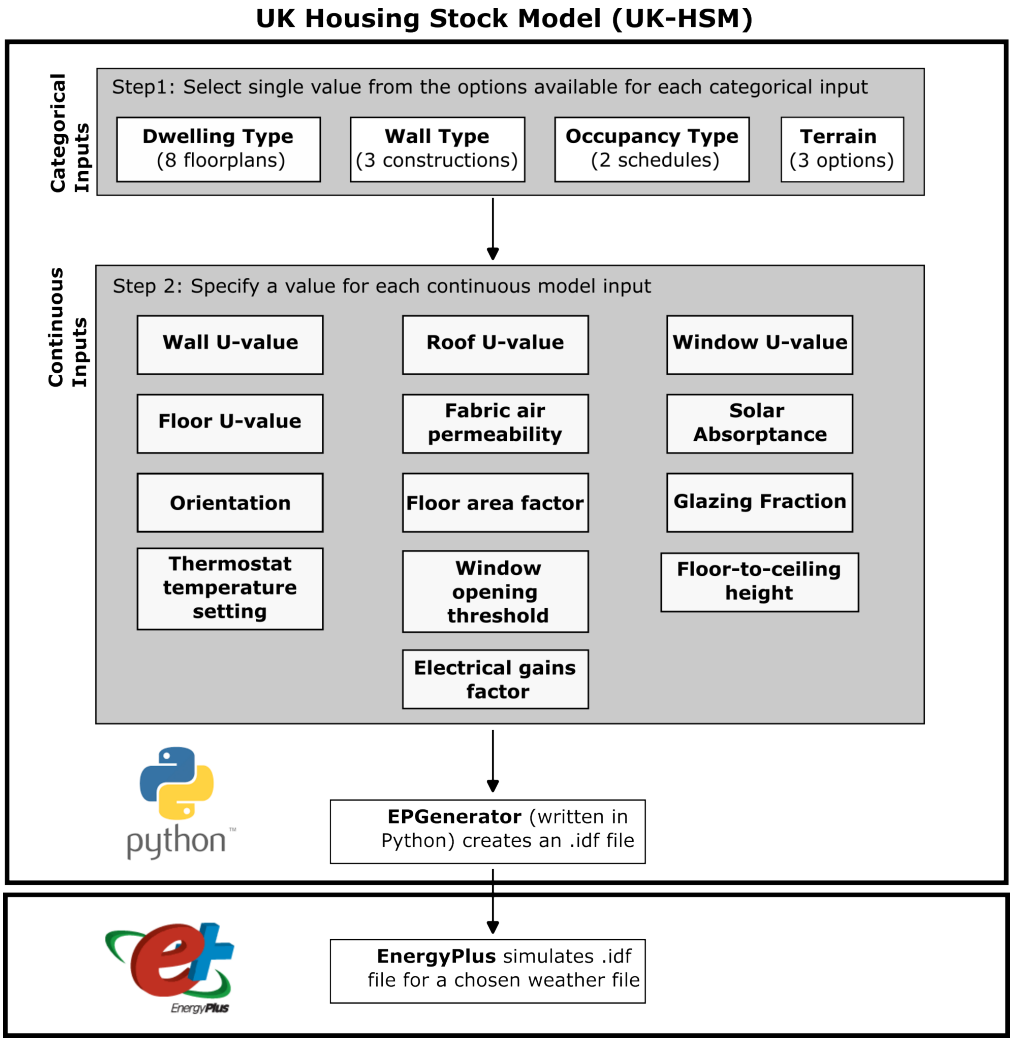


Figure 3.12: UK Housing Stock Model (UK-HSM) workflow diagram.

As illustrated in Figure 3.12, UK-HSM requires the specification of four categorical and twelve continuous model inputs. The possible values for the categorical variables are summarised in Table 3.2. Following input specification, as a comma

separated file (.csv) or directly in Python, EPG3 generates an EnergyPlus Input Data File (.idf) for each unique combination of model inputs which are subsequently simulated in EnergyPlus. EnergyPlus is a free, open-source and cross-platform whole building energy simulation tool developed by the United States Department of Energy (DOE), in collaboration with the National Energy Research Laboratory, academic institutions and private firms (DOE, 2022). It has descended from two legacy building simulation software, DOE-2 and BLAST, with its beta release in December 1999 (Crawley et al., 2001). Since then, it has received regular updates, currently at the rate of two releases per year, substantially expanding its capabilities (DOE, 2022); this has been facilitated by its *modular structure* which enables the easy integration of new features (Crawley et al., 2001). Another key characteristic of EnergyPlus is its *integrated solution manager*, which through the simultaneous evaluation of energy and moisture balance in the building, system and plant, enables a more accurate prediction of space temperatures than its predecessors (Crawley et al., 2001; DOE, 2016). For the present analysis, EnergyPlus version 8.8.0. was used, with six timesteps per hour.

Table 3.2: Categorical model inputs of the UK Housing Stock Model.

Parameter	Values
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

3.5.1 Building Characteristics

For the purposes of modelling the English housing stock, eight typologies (see Table 3.2) were derived from EHS and are defined in EPG3 (Taylor et al., 2016). These were the result of a statistical analysis of key dwelling characteristics within the EHS, with the aim to identify the most representative typologies within the housing stock (Oikonomou et al., 2018). Internal layouts were derived using typical floor plans for the typologies of the corresponding age and form, and these are included in Figures B.1–B.2 of the appendices. A summary of the number of rooms

(inc. hallways), bedrooms, ground floor area and total volume (excl. roof) per typology is provided in Table B.1, also in the appendices. By varying the *floor area factor*, the model floor area varies in proportion to the mean floor area for each typology. The model's *floor-to-ceiling height* and *orientation* can also be varied. To account for shading and wind exposure sheltering from nearby dwellings, replicates of each model are created depending on the built form. For example, a mirrored replicate of a semi-detached dwelling is created with a party wall separating them. The *terrain* variable modifies the local wind speed near the modelled building by varying the wind speed profile exponent and boundary layer thickness (DOE, 2016).

3.5.2 Building Fabric Characteristics

For each built form, the wall type is specified as one of the three (masonry) construction types: solid, cavity and filled cavity (Symonds et al., 2016). The choice of construction in UK-HSM accounts for differences in thermal mass, but related characteristics such as thermal insulation, air tightness and albedo are controlled independently. EPG3 uses pre-defined material and construction libraries, and adjusts their characteristics (e.g. thickness of insulation) depending on the *Wall U-value* (only for external walls), *Roof U-value*, *Window U-value* and *Floor U-value*. For the solid wall construction, internal wall insulation is assumed since this is considered to perform worse with regard to overheating compared to external wall insulation, resulting in the evaluation of the worst case scenario (Peacock et al., 2010; Symonds et al., 2016). When the Window U-value is ≤ 2.0 W/m^{2}K, the windows are assumed to be post-2002 and are modelled with trickle vents installed (Symonds et al., 2016). The *fabric air permeability* parameter determines the unintended airflow through the fabric – this is modelled as crack air flow within EnergyPlus, where two (a high and a low) cracks are modelled for each façade with their *air mass flow coefficient* adjusted depending on the specified air permeability (see DOE (2016) for how EnergyPlus models crack air flow). *Solar absorptance* is an input used to assess the impact of changing the albedo level of external roofs and walls of the model, while the *Glazing Fraction* controls the ratio of glazed area to external wall area.

3.5.3 Occupant Characteristics

To account for diversity in occupant's actions and presence, two types of occupancy are defined within EPG3:

1. *Pensioners*: Assumes two pensioners spending all day at home.

2. *Family*: Assumes a family of five which is out during the day on weekdays between 8 am and 6 pm.

For both occupancy types, windows can only be opened between May and September (inclusive), during the hours specified in Table 3.3 and if the indoor temperature exceeds a threshold temperature and is greater than the outdoor temperature. This is a simplified way of modelling window operation. Fabi et al. (2012) identified a number of factors that influence window opening behaviour, categorised into five groups: physical environmental, contextual, psychological, physiological and social. Indoor temperature was classified as an influencing physical environmental group, but so were other factors, such as, perceived illumination, smoking behaviour, outdoor temperature and time of the day (Fabi et al., 2012). In addition, Meinke et al. (2017) demonstrated through experiment that people perceive indoor temperatures and act to change their thermal comfort differently, even under controlled conditions. Furthermore, operating windows only when the internal temperature is greater than the external temperature is most effective with regard to reducing indoor overheating risk, but real-life occupancy is unlikely to be as consistent. Given the likely substantial influence that window operation will have on the two quantities of interest (Section 3.3), this simplified modelling approach, and the scarce data to inform the *window opening threshold*, are key sources of uncertainty; both of model inadequacy and parameter (2.3.2). A Bayesian approach to model calibration could identify the most likely value for the model input and quantify parameter uncertainties and model bias.

Heating is also assumed to be seasonal (January–April and October–December), differ between occupancy types and controlled by the *Thermostat temperature setting*. Since this study focuses on summer indoor temperatures, assumptions

Table 3.3: Hours during which windows were modelled as open between May and September if the indoor temperature exceeded the threshold and outdoor temperature.

	Bedrooms	Other Rooms
Pensioners	22:00–08:00	08:00–22:00
Family	22:00–08:00	08:00–09:00; 18:00–22:00

relating to heating will not be considered. However, it is acknowledged that a minority of occupants heat their homes during summer, potentially exacerbating indoor overheating problems (Lomas and Kane, 2013). In some cases, this could be due to the use of communal heating systems that do not always allow for occupant control.

The use of electrical equipment (including for cooking) may also contribute to indoor overheating through heat generated as a waste-product of their intended use. The assumptions regarding the schedules of use and power for each appliance are based on datasets available at the time of model development (Symonds et al., 2016). A single model input, *electrical gains factor*, is used to linearly vary the power level of electrical equipment, thus controlling the levels of internal gains. Given the stochastic nature of energy and appliance use, and the changes in appliance energy efficiency and practices over time, this component of the UK-HSM is also considered to be simplified and associated with significant uncertainties. Metabolic gains are also modelled and depend on whether the *pensioners* or the *family* occupancy profile is selected.

3.5.4 Modelling Details for Chosen Archetype

As will be detailed in Section 4.2.6.2, following Step 2 of the Bayesian calibration framework, the archetype model selected for calibration in this thesis was that of a semi-detached typology, with filled cavity walls, double glazing and occupied by pensioners. The *pensioners* occupancy profile was used due to the higher prevalence of this occupancy type, compared to the *family* profile, in the cluster of 4M dwellings identified following the classification process. For the brevity of this chapter, the modelling details regarding this archetype model have been included in the appendices: Table B.2 lists the key algorithms assumed, Tables B.3 and B.4 provides a

summary of the construction and building characteristics, and Table B.5 details the occupancy schedules.

3.6 Summary

This chapter set out a step-by-step process for addressing the aim and research objectives of this work, subsequent chapters will present the implementation of these steps. A modular and flexible Bayesian calibration framework for archetype-based models of summer indoor temperature was proposed. It consists of five steps and follows from the definition of homogeneity provided within the same section. The framework is not prescriptive with the methods to be used in each step, instead, it offers a flexible structure that can be adapted depending on the model and data available; in-depth description of the methods used for the framework's application in this thesis is provided in the "Methods" sections of Chapters 4–7. The quantity of interest for this work, the mean of the daytime living room temperature, was introduced and the arguments for its selection were discussed. The datasets of monitored summer indoor temperature, their associated metadata and weather data were also detailed within this chapter, along with the UK-HSM modelling structure and assumptions. The following chapter presents the implementation of Steps 1 and 2 of the proposed Bayesian calibration framework. The implementation of the remaining steps is then discussed in Chapters 5–7.

Chapter 4

Statistical Analysis & Categorical Variable Classification

The previous chapter introduced the Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature in Section 3.2. Step 1 of the framework requires the statistical analysis of an empirical dataset to identify variables that are significantly associated with the quantity of interest. The outputs of this process inform Step 2 of the framework, the categorical variable classification, where the housing stock is clustered into groups of dwellings suspected to be homogeneous. This chapter presents and discusses the methods of Step 1 and 2 in Section 4.1, and the results in Section 4.2. The outcomes of this chapter will inform Chapters 5 and 6, as illustrated in Figure 4.1.

To explore the potential associations between dwelling and household characteristics with the summer indoor temperatures monitored during the 2011 Energy Follow-Up Survey (Step 1), an approach similar to that used by Hamilton et al. (2017) is followed. The summer indoor temperatures are standardised to account for the inter-regional variation in local weather, a method similar to that introduced by Oreszczyn et al. (2006) is used (Section 4.1.1). Subsequently, a set of statistical tests are conducted to identify variables that are significantly associated with the standardised indoor temperature, as detailed in Section 4.1.2. The prevalence of indoor overheating risk within the housing stock based on the CIBSE TM59 metrics is quantified, with the approach described in Section 4.1.3, while this is not a re-

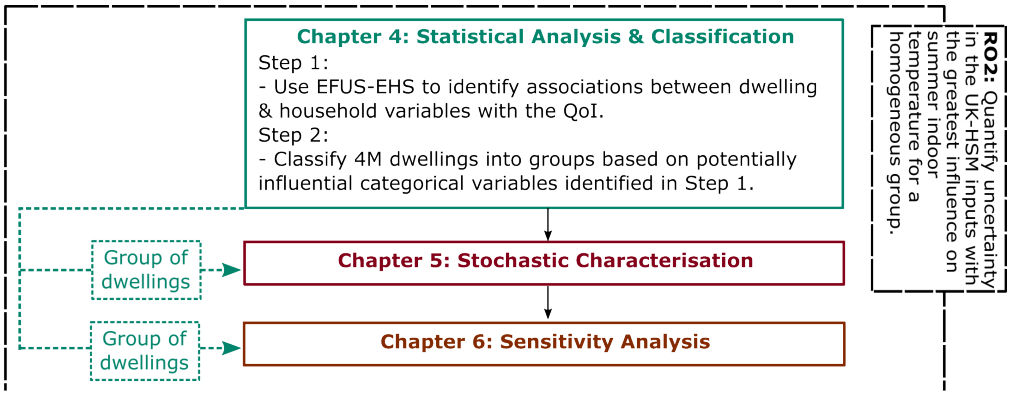


Figure 4.1: Chapter 4 flowchart. This is an abridged version of Figure 1.3, focusing on Chapter 4 and its outputs. RO2 is a shortened version of Research Objective 2. UK-HSM stands for UK Housing Stock Model, QoI for Quantity of Interest, EFUS for Energy Follow-Up Survey and EHS for English Housing Survey.

quirement for the Bayesian calibration framework it is included for completeness of this analysis. The results of the overheating and statistical analysis are summarised and discussed in Sections 4.2.1, 4.2.2 and 4.2.3. This part of the thesis has been published in a journal article (Petrou et al., 2019b).

Having identified variables that are significantly associated with summer indoor temperatures, the process of categorical variable classification (Step 2) is implemented, with the procedure outlined in Section 4.1.4. The results from this process are presented and discussed in Section 4.2.6. The limitations of the work presented in this chapter are summarised in Section 4.3, and the chapter concludes with a summary in Section 4.4.

4.1 Methods

4.1.1 Standardisation of Indoor Temperature

Both in the winter, and summer, the outdoor weather is a key determinant of indoor temperature. To compare internal winter temperatures against dwelling characteristics for homes monitored across the country, Hamilton et al. (2017) considered it "necessary to create a common baseline". They employed a method introduced by Wilkinson et al. (2001) and expanded on by Oreszczyn et al. (2006) to standardise the monitored indoor temperatures against external weather conditions and compare

the standardised indoor temperature (SIT) for dwellings with different characteristics. For each monitored home, a regression-based model was constructed using observations of indoor and outdoor temperature. The regression models were subsequently used to predict a single SIT for each home for the same outdoor temperature. This approach enables for outdoor temperature to be controlled, diminishing its effect as a confounder when examining the association between indoor temperature and dwelling characteristics. With the same rationale, the summer indoor temperature may also be standardised against weather conditions to carry out a similar analysis of association between SIT, dwelling and household characteristics. However, the choice of temporal resolution, data points included in the analysis and standardised weather conditions will depend on the purpose of the analysis.

Given the focus on periods when occupants would most likely be at home, during cold conditions and when the heating system would most likely be used, Hamilton et al. (2017) derived a regression function for each living room using hourly observations between 07:00-09:59 and 19:00-21:59, and for each bedroom between 20:00-07:59. In addition to estimating the SIT at an ambient temperature of 5 °C, as was previously done by Wilkinson et al. (2001) and Oreszczyn et al. (2006), Hamilton et al. (2017) also estimated the SIT at an ambient temperature of 0 °C and 10 °C to investigate whether their findings change at colder and warmer outdoor conditions. This analysis revealed that the trends in the differences in SIT for different building characteristics were broadly similar, but that the magnitudes tended to be greater as outdoor conditions became colder.

In this thesis, twelve regression models were assessed in the standardisation of the Mean Daytime Living Room Temperature (MDLRT) and the Mean Nighttime Bedroom Temperature (MNBT), the quantities of interest for this thesis (Section 3.3). The models were trained only for the period of May to September (inclusive), with the explanatory variables being the outdoor temperature and global horizontal irradiance (GHI); the choice of these weather variables was informed by the literature (Taylor et al., 2014). The models varied in their combination of weather variables used, and the temporal resolution of the predictors (see Section C of the appendices for

more details). To evaluate each model's performance, the adjusted coefficient-of-determination (R^2) was calculated for each home, and each model's distribution of adjusted R^2 was visually assessed. R^2 quantifies the proportion of total variation of the dependent variable (in this case, MDLRT and MNBT) that can be explained by the explanatory variables (Rencher and Christensen, 2012), with a value of 1 indicating a perfect fit. Adjusted R^2 has a similar interpretation, but penalises the number of explanatory variables to avoid overfitting (Reimann et al., 2008). While weather variables are expected to have a major influence on indoor temperature, other influential variables will also exist, thus a value of R^2 equal to 1 was not expected.

Amongst the numerous models evaluated, a balance was struck between model efficacy and simplicity for a model that was based on linear terms of daily-mean outdoor temperature ($T_{out,mean}$) and Global Horizontal Irradiance (GHI_{mean}) as described by equation:

$$\mathrm{SIT}_{room} = \beta_0 + \beta_1 T_{out,mean} + \beta_2 GHI_{mean} \tag{4.1}$$

SIT_{room} is the mean day-time (08:00-22:00) indoor temperature estimated for the living room or the mean nighttime (22:00-08-00) temperature for the bedroom. β_{0-2} are the linear regression coefficients. Different regression coefficients were estimated for each room of each dwelling using the **stats** package of the programming language R (R Core Team, 2018). The median of the adjusted R^2 for the MDLRT and MNBT was 0.48 and 0.64, respectively (see Section C of the appendices for the full set of results).

Given the focus on warm summer conditions and excess heat-related mortality, a relatively high outdoor temperature had to be chosen as the standardisation temperature. Hajat et al. (2014) identified different thresholds over which heat-mortality takes place for each region of England based on the daily mean temperature. These thresholds ranged from 16.6 °C in the North East to 19.6 °C in London. To ensure that standardised temperature is considered warm across the entire of England, a daily-mean temperature of 20 °C was used, since it exceeded all regional heat-mortality thresholds. This value was considered an upper end of what might be

used for the standardisation, given the relatively mild summer conditions of 2011 (see Figure 3.11) for which the models could be trained on. The GHI_{mean} value of 210 Wh/m^2 was used, since it was the average daily-mean GHI across the days that the daily-mean outdoor temperature exceeded 19 °C.

It should be noted that if the SIT was used to compare the severity of overheating in different regions, with a standardised temperature at the upper extreme of summer conditions in the north, then it would indicate overheating to be more severe than what occupants experience. Similarly, using a standardised temperature at the lower extreme of regions in the south would erroneously suggest overheating to be less severe than what is experienced by occupants. However, SIT was not used in this thesis to compare the extent of overheating in different regions, but only to study the association of dwelling and household characteristics with summer indoor temperature.

4.1.2 Statistical Analysis

The approach to statistical analysis is based on precedence, since it has been previously used in the analysis of (winter) indoor temperatures (Hamilton et al., 2017) and allows for a thorough assessment of each explanatory variable.

To explore the association of household and dwelling characteristics with standardised indoor temperatures, 20 categorical variables were selected, which are summarised in Tables 4.1-4.2. The selection of variables was based on a combination of characteristics which have been associated with indoor overheating by published modelling and empirical studies (Beizaee et al., 2013; Firth and Wright, 2008; Lomas and Kane, 2013; Mavrogianni et al., 2012; Pathan et al., 2017), and variables suspected to be influential but which have never been previously investigated. One variable investigated was SAP 09. 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 associated with space heating, water heating, ventilation and lighting, minus cost savings from energy generation technologies (BRE, 2011). It ranges from 1 (highest energy costs) to 100 (lowest energy costs).

A number of statistical techniques were used to investigate the differences in

Table 4.1: Summary of the household variables analysed. HRP is the household reference person, LA stands for Local Authority and RSL for Registered Social Landlord.

Variable names	Value
Household composition	couple, no dept. child(ren) < 60; couple, no dept. child(ren) ≥ 60; couple with dept. child(ren); lone parent with dept. child(ren); other multi-person households; one person < 60; one person aged 60 or over
No of people in the household	1; 2; 3; 4; 5; 6 or higher
Age band of youngest person	0-4; 5-10; 11-15; 16-24; 25-59; 60-74; 75-84; 85 or more
Age band of oldest person	16-34; 35-49; 50-59; 60-74; 75-84; 85 or more
Employment status of HRP and partner	1 or more work full time; 1 or more work part-time; none working, one or more retired; none working, and none retired
Tenure	own with mortgage; own outright; privately rent; rent from LA; rent from RSL
Anyone illness or disability	Yes; No
All households - income in 5 bands	lowest 20 %; quintile 2; quintile 3; quintile 4; highest 20 %
Occupant on means tested or certain disability related benefits	Yes; No

SIT associated with the selected dwelling and household characteristics, summarised in Table 4.3. To assess whether statistically significant differences exist for the SIT of each variable, the Kruskal-Wallis test was used. This test has been previously used for a similar analysis by Hamilton et al. (2017), it does not assume normality, and it is able to deal with extreme data. Whether a significant difference exists was indicated by the p-value: a measure of how likely the observed data are under the null hypothesis, between 0 for impossible and 1 for certain (Greenland et al., 2016). The maximum acceptable p-value (significance level) was chosen in advance to be 0.05, a common choice in statistical analysis (Reimann et al., 2008). If the p-value ≤ 0.05 for any variable (e.g. dwelling type), there is enough evidence to support a statistically significant difference between the median SIT of the variable's levels (e.g. bungalow, detached etc.). To compare the median SIT of each variable's

Table 4.2: Summary of the dwelling variables analysed. Total useable floor area represents the entire area within the dwelling's footprint, excluding the area occupied by staircases, internal and external walls.

Variable names	Values
Dwelling Type	bungalow; converted flat; detached; end-terrace; mid-terrace; purpose built flat; semi-detached
Dwelling Age	pre-1850; 1850-1899; 1900-1918; 1919-1944; 1945-1964; 1965-1974; 1975-1980; 1981-1990; post-1990
Total Useable Floor Area	less than 50 sqm; 50 to 69 sqm; 70 to 89 sqm; 90 to 109 sqm; 110 sqm or more
No. of Storeys	1; 2; 3; 4; 5 or more
Construction	solid masonry; cavity masonry; timber frame; steel frame; concrete frame; concrete boxwall
Double Glazing	no double glazing; less than half; more than half; entire house
Nature of Area	city centre; other urban centre; suburban residential; rural residential; village centre
Traffic Problems	Yes; No
Main Heating System	boiler system with radiators; storage radiators; room heater; communal
Loft Insulation	none; less than 100mm; 100 up to 150mm; 150mm or more
SAP 09	less than 30; 30 to 50; 51 to 70; more than 70

levels, their SIT variance and distribution must be similar (McDonald, 2014). If this assumption is not satisfied, *stochastic dominance* could still be demonstrated; that is, assessing whether it is likely that an observation from one level is greater than an observation in the other (Mangiafico, 2016). In this situation, a p-value less or equal to 0.05 would be interpreted as the distribution of SIT still being significantly different, but it is not necessarily true that their median values are different.

As additional statistical measures to support this analysis, 95 % Confidence Intervals (CI) were estimated and Pairwise Mann-Whitney U-tests were conducted (Table 4.3). A CI provides a sense of how accurate the sample median is relative to the population median (Mangiafico, 2016). In addition and for each level, the Pairwise Mann-Whitney U-tests for multiple comparisons with the False Discovery Rate (FDR) p-adjustment method were performed. A p-value smaller or equal to 0.05 indicates a statistically significant difference between that level's SIT and that

of the first level of each variable.

Table 4.3: Statistical tests and techniques used to compare the standardised indoor temperature between dwellings.

Test	Null Hypothesis	Explanation
Kruskal-Wallis (Mangiafico, 2016; McDonald, 2014)	Variance and distribution similar: *The median SIT across the different levels (sub-groups) of each explanatory variable is the same at a significance level of* 5 %	If the p-value $\leq$ 0.05 for any variable (e.g. dwelling type), there is enough evidence to support a statistically significant difference between the median SIT of the variable's levels (e.g. bungalow, detached etc.)
	Otherwise: *The probability of a randomly selected SIT in one sub-group being greater than a randomly selected SIT from the other sub-group is 50 % at a significance level of* 5 %.	A p-value smaller or equal to 0.05 indicates a statistically significant difference between an observation in one group and that in the other.
95 % Confidence Interval (CI) (McDonald, 2014)		By estimating a 95 % CI, it is assumed that if repeated random samples were taken from the population and the median and confidence intervals were estimated, the confidence interval for 95 % of the samples would include the population median.
Pairwise Mann-Whitney U-tests for multiple comparisons with the False Discovery Rate (FDR) p-adjustment method (Divine et al., 2018)	*The probability of a randomly selected SIT in one sub-group being greater than a randomly selected SIT from the other sub-group is 50 % at a significance level of* 5 %.	A p-value smaller or equal to 0.05 indicates a statistically significant difference between that level's SIT and that of the first level of each variable.
Fisher's exact test (McDonald, 2014)	*There is no statistical association between categorical explanatory variables at a significance level of* 5 %.	If the p-value for any combination of variables (e.g. household composition and dwelling type) was less or equal to 0.05, a statistically significant association was assumed.

To determine whether the dwelling characteristics are correlated to the house-

hold characteristics, Fisher's exact test was used to explore whether the proportions of a dwelling variable are different depending on the values of the household variable (McDonald, 2014). For the above analysis, cases where the occupants did not provide an answer or stated that the survey question is not applicable to them were excluded.

4.1.3 Indoor Overheating Assessment

To translate indoor temperatures into overheating risk, the two criteria defined in TM59 for naturally ventilated dwellings were used (CIBSE, 2017). The validation of TM59, which is a design stage guidance tool, is beyond the scope of this thesis. A form of these criteria has been used in the past to assess overheating risk in previous in-use studies (Lomas and Kane, 2013; Mavrogianni et al., 2016; Vellei et al., 2017). According to TM59, there is a high risk of overheating if either of the following thresholds is exceeded (CIBSE, 2017):

1. The percentage of occupied hours where the operative temperature T_{op} exceeds the maximum allowable temperature T_{max} by 1 °C or more during the period May to September, inclusive, exceeds 3%.

2. Bedroom operative temperature exceeds 26 °C for more than 1% of the assumed sleeping hours (22:00-07:00) annually (equivalent to 32 hours – 33 hours or more above 26 °C are recorded as overheating hours).

Operative temperature (T_{op}) is the weighted average of the room's air (T_{air}) and mean radiant (T_{rad}) temperature, defined as:

$$T_{op} = AT_{air} + (1-A)T_{rad}, \tag{4.2}$$

where A is the ratio $h_c/(h_c+h_r)$, which depends on the surface heat transfer coefficient of the clothed body by convection (h_c) and radiation (h_r) (Nicol and Humphreys, 2010). According to CIBSE Guide A, the *air temperature* is the "temperature registered by a dry thermometer, shielded from radiation, suspended in the air", while the *mean radiant temperature* is defined as "uniform surface temperature of a radiantly black enclosure in which an occupant would exchange the same amount of

radiant heat as in the actual non-uniform space" (CIBSE, 2015). Under the assumption of low air flow speeds, it is reasonable to assume that (Nicol and Humphreys, 2010):

$$T_{op} = \frac{1}{2}(T_{air} + T_{rad}), \tag{4.3}$$

T_{max} is the estimated maximum acceptable temperature according to the adaptive thermal comfort model (CIBSE, 2013; CIBSE, 2015):

$$T_{max} = 0.33 \times T_{rm} + 21.8^{o}\mathrm{C}, \tag{4.4}$$

where T_{rm} is the exponentially weighted running mean of outdoor ambient temperature. For a series of days, T_{rm} can be approximated using (CIBSE, 2013):

$$T_{rm} = (1-\alpha)T_{od-1} + \alpha T_{rm-1}, \tag{4.5}$$

where α is a constant commonly taken as 0.8 while T_{od-1} and T_{rm-1} are the daily mean and running mean temperatures, respectively, of the day previous to the day of interest (T_{od}).

Local weather data were used to estimate a T_{max} for each region using the equations in CIBSE TM52 (CIBSE, 2013). The dwellings were assumed to be predominantly naturally ventilated, with the living room being occupied between 09:00 and 22:00 and the bedroom being always occupied, as suggested in TM59 (CIBSE, 2017). The indoor temperatures monitored during the 2011 EFUS were used in this analysis; although they are likely a function of air and radiant temperature, they are not expected to correspond to the operative temperature as assumed by TM59. However, given the data available, this limitation could not be overcome, and it is common amongst empirical studies (Lomas and Porritt, 2017; Lomas et al., 2021).

4.1.4 Categorical Variable Classification

Categorical variables identified as having a statistically significant association with summer indoor temperatures were used to select a group of dwellings within the 4M

dataset suspected to be homogeneous, in accordance with the Bayesian calibration framework and definition of homogeneity provided in Section 3.2. Whether the group is homogeneous will be determined following the sensitivity analysis in Chapter 6. Since the model used for the calibration (UK-HSM) already exists, and model development was not within the scope of this thesis, the classification process took into account the model's structure. This part of the analysis had two components.

In the first part, the outcomes of the statistical analysis were considered in conjunction with the UK-HSM model structure to determine which variables would be used as classifiers. The guidelines for this process were:

- If a statistically significant categorical variable was modelled explicitly (e.g. dwelling type), it was used as a classifier and the appropriate modelling option was used for each group. If the variable's possible values are not all modelled, only the empirical data associated with modelled variable values were used for calibration.

- If a statistically significant categorical variable was modelled as a continuous model input (e.g. floor area), it was not used as a classifier.

In the second part, the identified classifiers were used to select a homogeneous group of dwellings within the 4M dataset. A pragmatic approach was followed, where the classifiers considered to be the most important were used in order to avoid excessive segmentation, given the relatively small number of homes monitored.

4.2 Results & Discussion

Following from a description of the methods used, the statistical and overheating analysis results are presented and discussed in this section. Section 4.2.1, provides a summary of the indoor overheating risk results, followed by the statistical analysis results in Sections 4.2.2–4.2.3. In Section 4.2.4, the results of Fisher's Exact test looking at the correlations of dwelling and household characteristics are summarised. Results from the categorical variable classification are presented and discussed in Section 4.2.6. For clarity, the term "significant" is only used in the following sections to mean "statistically significant".

4.2.1 Indoor Overheating Assessment

For the EFUS dataset, a total of 20 out of 795 (2.5 %) living rooms exceeded the threshold of CIBSE TM59 Criterion 1 (defined in Section 4.1.3), while 178 (22.3 %) living rooms recorded some overheating hours but less than 3 % of occupied hours. A similarly small number of dwellings failed Criterion 1 for the bedroom (19 out of 799, 2.4 %), while a greater number (284, 35.5 %) experienced some hours of overheating. The extent of indoor overheating appears to be different when Criterion 2 is used, with 204 (25.5 %) bedrooms having exceeded the static threshold. A substantially greater prevalence of indoor overheating according to Criterion 2 (69 %) relative to Criterion 1 (19 %) was also observed by Lomas et al. (2021) in their analysis of indoor temperatures collected during the warm summer of 2018.

Table 4.4: Summary of the TM59 assessment results for the bedroom (B) and living room (LR) of each dwelling.

		Number of dwellings with percentage of Overheating Hours (% OH) by range		
Criterion	Total sample size	0	$0 < \%\ \text{OH} \leq 3$	$\%\ \text{OH} > 3$
Criterion 1: LR	795	597	178	20
Criterion 1: B	799	496	284	19
		0	$0 < \%\ \text{OH} \leq 1$	$\%\ \text{OH} > 1$
Criterion 2	799	377	218	204

The estimated prevalence of indoor overheating may be compared to data on stated thermal discomfort collected during EFUS, with the occupants being asked whether they find it difficult to keep the bedroom cool. From a total number of 61 who responded positively, 29 (47.5 %) were found to exceed the Criterion 2 threshold, 21 (34.4 %) had some hours of overheating recorded while 11 (18 %) had no hours recorded. The agreement between predicted and stated indoor overheating was lower when looking at Criterion 1 (the exact number is not provided to reduce the chance of identification). On the contrary, Lomas et al. (2021) found a better agreement between Criterion 1 and stated thermal discomfort than with Criterion 2. The large (175) number of dwellings that failed Criterion 2 while their occupants did not report thermal discomfort could also be further evidence of support to the

ongoing discussion on the strictness of the 26 °C threshold (Nicol and Humphreys, 2018; Lomas et al., 2021).

4.2.2 Statistical Analysis of Household Characteristics

Tables 4.5–4.7 summarise the median and 95 % Confidence Intervals (CI) of the SIT for each household characteristic. The associated p-values indicate whether a statistically significant difference (if p-value ≤ 0.05) exists. An asterisk next to the p-value indicates that while a statistically significant difference in the medians could not be assessed, stochastic dominance was. Figure 4.2 presents a comparison of bedroom and living room SIT for the following four variables: Household composition, main heating system, income and tenure.

For household composition, there was no significant difference in the living room, with the median SIT lying within a range of 0.2 °C. On the contrary, the bedroom SIT deviated significantly, with the median SIT for a single occupant aged 60 or over being at 23 °C (CI: [22.7, 23.4] °C), 1.1 °C lower than the maximum median bedroom SIT observed for this variable. For households with a single occupant, the median bedroom SIT is 23.3 °C, 0.8–0.9 °C lower than households with three or more occupants. The number of people in the household is significant only in relation to the bedroom SIT, with the median bedroom SIT increasing when the number of occupants increase from one to three, with small to zero differences for further increases in the number of occupants. An association between the number of occupants has been previously discovered in research in overcrowded homes, defined as having five or more occupants, which indicated greater indoor temperatures in both rooms compared to non-overcrowded homes (Vellei et al., 2017). A pattern of increasing overheating risk in the bedroom as the number of occupants increased was also observed by Lomas et al. (2021).

A statistically significant association was also observed between the bedroom SIT and the age of the youngest and oldest occupants. Households whose youngest occupant was in the 60-74 age band had the lowest median bedroom SIT (23.4 °C, CI: [23.0, 23.5] °C), 0.5–0.8 °C lower than households with occupants younger than 24. Statistically significant differences were also observed based on the age band

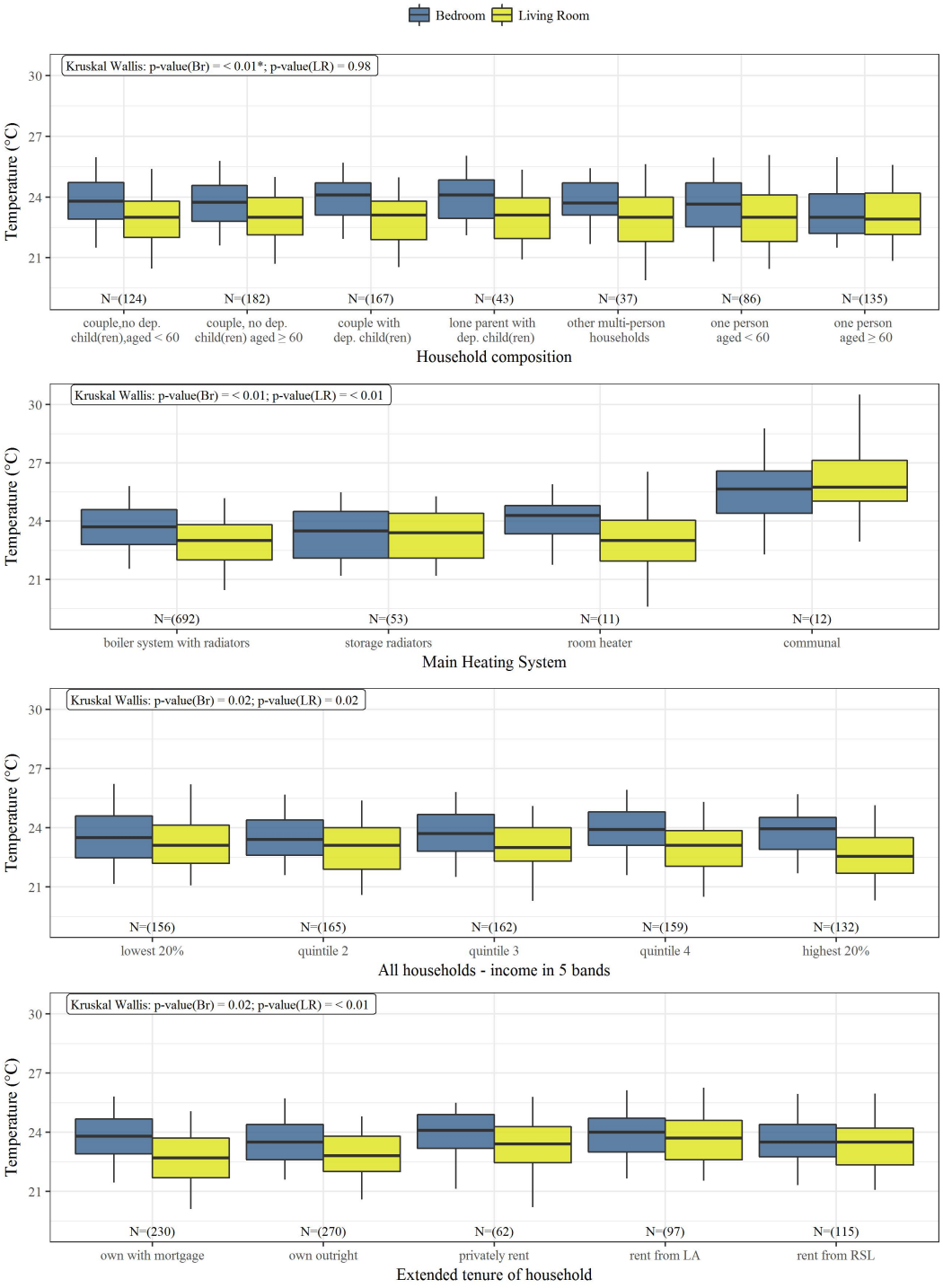


Figure 4.2: Box plots of standardised indoor bedroom and living room temperatures. The whiskers represent the 5th and 95th percentile. Outliers were masked for data privacy reasons. * on p-values indicates groups where the assumption of equal variance was not met but where the stochastic dominance could be assessed.

Table 4.5: Summary of the median standardised indoor temperatures (SIT), 95 % Confidence Interval (CI) and significance test results. The p-values (p) associated with each variable are the results the Kruskal-Wallis test and the p-values associated with each level of the variable are the result of the Pairwise Mann-Whitney U-tests. * indicates groups where the assumption of equal variance was not met but stochastic dominance could be assessed. The bold font highlights cases where the p-value ≤ 0.05.

		Bedroom SIT (°C)		Living Room SIT (°C)	
	N	Med (CI 95%)	p	Med (CI 95%)	p
Household composition	-	-	**<0.01***	-	0.98
couple, no dept. child(ren) <60	124	23.8 (23.5, 24.2)	-	23 (22.6, 23.3)	-
couple, no dept. child(ren) $\geq$ 60	182	23.8 (23.5, 24)	0.76	23 (22.7, 23.3)	0.99
couple with dept. child(ren)	167	24.1 (23.8, 24.2)	0.48	23.1 (22.8, 23.4)	0.99
lone parent with dept. child(ren)	43	24.1 (23.3, 24.5)	0.83	23.1 (22.5, 23.7)	0.99
other multi-person hholds	37	23.7 (23.2, 24.2)	0.85	23 (22.3, 23.8)	0.99
one person <60	86	23.6 (23.2, 24.1)	0.66	23 (22.6, 23.5)	0.99
one person aged 60 or over	135	23 (22.7, 23.4)	**0.01**	22.9 (22.5, 23.5)	0.99
No of persons in the household	-	-	**<0.01***	-	0.78
1	221	23.3 (23, 23.5)	-	23 (22.7, 23.4)	-
2	294	23.7 (23.5, 23.9)	**0.03**	23 (22.8, 23.3)	0.99
3	113	24.1 (23.5, 24.2)	**0.01**	22.7 (22.5, 23.2)	0.99
4	107	24.1 (23.8, 24.3)	<0.01	23.2 (22.9, 23.4)	0.99
5	24	24.1 (22.8, 24.4)	0.33	22.6 (21.9, 23.7)	0.99
6 or higher	15	24.2 (23.5, 24.7)	0.12	23.4 (21.3, 24.1)	0.99
Age band of youngest person	-	-	**<0.01**	-	0.17
0-4	71	24.1 (23.8, 24.3)	-	22.9 (22.5, 23.4)	-
5-10	68	24 (23.5, 24.3)	0.74	23.4 (22.5, 23.7)	0.86
11-15	50	23.9 (23.4, 24.3)	0.66	22.8 (22, 23.3)	0.77
16-24	73	24.2 (23.8, 24.6)	0.85	23 (22.6, 23.3)	0.95
25-59	233	23.7 (23.5, 24)	0.41	23 (22.8, 23.4)	0.86
60-74	213	23.4 (23, 23.5)	**<0.01**	22.9 (22.6, 23.1)	0.95
75-84	54	23.5 (23.1, 24.2)	0.41	23.6 (23, 23.8)	0.25
85 or more	12	23.8 (22.6, 24.7)	0.66	22.9 (22, 24.1)	0.95

Table 4.6: Summary of the median standardised indoor temperatures (SIT), 95 % Confidence Interval (CI) and significance test results. The p-values (p) associated with each variable are the results the Kruskal-Wallis test and the p-values associated with each level of the variable are the result of the Pairwise Mann-Whitney U-tests. HRP is the household reference person. * indicates groups where the assumption of equal variance was not met but stochastic dominance could be assessed. The bold font highlights cases where the p-value ≤ 0.05.

		Bedroom SIT (°C)		Living Room SIT (°C)	
	N	Med (CI 95%)	p	Med (CI 95%)	p
Age band of oldest person	-	-	**<0.01**	-	**0.04**
16-34	53	24.4 (24, 24.8)	-	23.4 (22.9, 23.8)	-
35-49	202	24 (23.7, 24.2)	0.13	23 (22.6, 23.4)	0.17
50-59	159	23.5 (23.3, 23.8)	**0.01**	22.8 (22.6, 23.2)	0.17
60-74	259	23.4 (23.2, 23.7)	**<0.01**	22.9 (22.6, 23.3)	0.21
75-84	83	23.6 (23.4, 24)	**0.02**	23.4 (23, 23.7)	0.90
85 or more	18	23.6 (22.8, 24.3)	0.11	22.9 (22, 24)	0.69
Employment status of HRP and partner	-	-	**<0.01**	-	0.12
1 or more work full time	351	24 (23.7, 24.1)	-	23 (22.7, 23.2)	-
1 or more work part time	65	23.9 (23.5, 24.3)	0.91	22.9 (22.5, 23.5)	0.65
none working, one or more retired	268	23.4 (23.2, 23.6)	**<0.01**	23 (22.8, 23.2)	0.33
none working and none retired	90	23.7 (23.3, 24.2)	0.91	23.3 (22.8, 23.7)	0.08
Extended tenure of household	-	-	**0.02**	-	**<0.01**
own with mortgage	230	23.8 (23.5, 24.1)	-	22.7 (22.4, 23)	-
own outright	270	23.5 (23.3, 23.7)	0.11	22.8 (22.6, 23)	0.31
privately rent	62	24.1 (23.8, 24.6)	0.42	23.4 (22.9, 23.8)	**0.01**
rent from LA	97	24 (23.6, 24.3)	0.70	23.7 (23.1, 23.9)	**<0.01**
rent from RSL	115	23.5 (23.3, 23.8)	0.26	23.5 (23, 23.7)	**<0.01**
Anyone illness or disability	-	-	0.40	-	**0.02**
Yes	286	23.5 (23.5, 23.8)	-	23.1 (22.9, 23.5)	-
No	482	23.8 (23.6, 24)	0.28	22.9 (22.7, 23.1)	**0.01**

Table 4.7: Summary of the median standardised indoor temperatures (SIT), 95 % Confidence Interval (CI) and significance test results. The p-values (p) associated with each variable are the results the Kruskal-Wallis test and the p-values associated with each level of the variable are the result of the Pairwise Mann-Whitney U-tests. * indicates groups where the assumption of equal variance was not met but stochastic dominance could be assessed. The bold font highlights cases where the p-value ≤ 0.05.

		Bedroom SIT (°C)		Living Room SIT (°C)	
	N	Med (CI 95%)	p	Med (CI 95%)	p
All households - income in 5 bands	-	-	**0.02**	-	**0.02**
lowest 20 %	156	23.5 (23.2, 23.8)	-	23.1 (22.9, 23.5)	-
quintile 2	165	23.4 (23.3, 23.8)	0.98	23.1 (22.8, 23.5)	0.52
quintile 3	162	23.7 (23.5, 24)	0.36	23 (22.7, 23.5)	0.53
quintile 4	159	23.9 (23.6, 24.2)	**0.05**	23.1 (22.8, 23.4)	0.48
highest 20 %	132	24 (23.6, 24.2)	0.22	22.6 (22.3, 23)	**0.01**
Occupant on means tested or certain disability related benefits	-	-	0.47	-	**<0.01**
Yes	250	23.6 (23.4, 23.8)	-	23.4 (23.1, 23.7)	-
No	524	23.8 (23.5, 23.9)	0.49	22.9 (22.7, 23)	**<0.01**

of the oldest person. Households where the oldest person was in the 50–59, 60–74 or 75–84 age band had a significantly lower bedroom SIT than homes with the oldest person being in the 16–34 age band. With a p-value of 0.04 for the Kruskal Wallis test of living room SIT and oldest person age band, a statistically significant result was again observed. However, this should be treated with caution, as the wide confidence intervals of the lowest and highest temperatures overlap. Related to the age-band variables, the presence of one or more retirees was associated with a significantly lower median bedroom SIT of 23.4 °C (CI: [23.2, 23.6] °C) compared a median bedroom SIT of 24 °C (CI: [23.7, 24.1] °C) in households where one or more are working full time. A similar result was not discovered for the living room SIT.

A statistically significant effect is also observed for the household's tenure with regard to both the bedroom and living room SIT. The bedrooms of homes rented privately (24.1 °C, CI: [23.8, 24.6] °C) or from a Local Authority (24.0 °C, CI:

[23.6, 24.3] °C) were the warmest. The median living room SIT of all three groups of rented homes were higher than homes owned with a mortgage or outright (see Figure 4.2 for a comparison), in partial agreement with Hulme et al. (2013b) and Lomas et al. (2021). Specifically, the median living room SIT for homes rented from a Local Authority was 0.9 °C and 1.0 °C higher than that of homes owned with mortgage and outright, respectively. There could be a few factors contributing to such differences, including the higher prevalence of flats being rented compared to being owner occupied (MHCLG, 2019). An important implication of this finding is that occupants exposed to higher indoor temperature would, in many cases, not be the owners and would thus largely depend on their landlond or local authority for the installation of significant structural or engineering-based overheating adaptation measures.

Households with occupants on means tested or certain disability benefits or where someone suffers from an illness or disability (but does not necessarily receive any benefits) had statistically higher median living room SIT. While for homes of occupants with illness or disability the difference is only 0.2 °C and the confidence intervals overlap, a clear difference is observed for households with occupants on benefits with a difference of 0.5 °C. A statistically significant difference in the prevalence of indoor overheating between households with occupants suffering from illness or disability and those without were also observed by Lomas et al. (2021).

A significant result was observed for bedroom and living room SIT when looking at household income, with opposite trends for each room. The bedroom median SIT was highest for the top two quintiles while the living room median SIT was lowest for the top quintile.

4.2.3 Statistical Analysis of Dwelling Characteristics

A summary of the median SIT and 95 % CI for the dwelling characteristics along with the associated p-values is provided in Tables 4.8–4.10.

Dwelling type and age, floor area, the number of storeys, construction and main heating system all appear to have a statistically significant association with the SIT in both rooms. Bungalows (22.8 °C, CI: [22.6, 23.1] °C) and converted flats (22.0 °C,

CI: [20.8, 23.2] °C) have the lowest median bedroom SIT, while mid-terraced had the highest median SIT with 24.1 °C (CI: [23.8, 24.3] °C). The median bedroom SIT for semi-detached, detached, end-terrace houses and purpose built flats are within 0.2 °C of each other. A different ranking is observed when looking at living room SIT across the different dwelling types. Purpose-built flats have the highest median living room SIT with 24.0 °C (CI: [23.6, 24.3] °C), followed by bungalows, with a 0.6 °C difference. Converted flats and detached dwellings have the lowest living room SIT with median values of 22.5 °C (CI: [21.7, 23] °C) and 22.6 °C (CI: [22.3, 22.9] °C), followed by the semi-detached 22.8 °C (CI: [22.6, 23.1] °C).[1] In general, these results are in agreement with previous work; Lomas and Kane (2013) and Beizaee et al. (2013) found the living rooms in detached houses to be significantly cooler than in other dwelling types, while for flats they were the warmest, a result also supported by modelling work in London and across Great Britain (Mavrogianni et al., 2012; Taylor et al., 2016).

For both rooms, pre-1900 dwellings are overall cooler than post-1900 homes. For the bedroom, the median SIT for pre-1850 and 1850-1899 age bands was 23.0 °C (CI: [21.6, 23.7] °C) and 23.1 °C (CI: [22.3, 23.5] °C). The age bands associated with the highest bedroom median SIT were the 1919-1944 and post-1990 with 24.0 °C (CI: [23.8, 24.3] °C) and 24.0 °C (CI: [23.6, 24.2] °C). The median bedroom SIT for the remaining age bands fluctuates between 23.5 °C to 23.8 °C. Pre-1850 had a median living room SIT of 21.2 °C (CI: [20.1, 22] °C), significantly lower than the rest of the age bands, with the exception of 1850-1899 dwellings with a median SIT of 21.8 °C (CI: [21.2, 22.5] °C). The median living room SIT of 23.7 °C (CI: [22.9, 24.0] °C) associated with 1975-1980 age band was the highest amongst all dwelling ages.

The Kruskal Wallis analysis suggested a significant difference amongst the floor area levels in both rooms. While a clear trend is seen for the living room, the same is not true for the bedroom. For the living room, the median SIT is negatively

[1] In this comparison, the small sample size of converted flats may have influenced the representativeness of the estimated median, and this uncertainty is reflected in the large confidence interval. In addition, the converted flats in this study are likely to consist of diverse typologies (before conversion they could have been semi-detached or mid-terrace).

Table 4.8: Summary of the median standardised indoor temperatures (SIT), 95 % Confidence Interval (CI) and significance test results. The p-values (p) associated with each variable are the results of the Kruskal-Wallis test, and the p-values associated with each level of the variable are the result of the Pairwise Mann-Whitney U-tests. * indicates groups where the assumption of equal variance was not met, but stochastic dominance could be assessed. The bold font highlights cases where the p-value ≤ 0.05.

		Bedroom SIT (°C)		Living Room SIT (°C)	
	N	Med (CI 95%)	p	Med (CI 95%)	p
Dwelling Type	-	-	**<0.01***	-	**<0.01***
bungalow	97	22.8 (22.6, 23.1)	-	23.4 (22.9, 23.7)	-
converted flat	15	22.0 (20.8, 23.2)	**0.04**	22.5 (21.7, 23)	**0.02**
detached	137	23.8 (23.5, 24.1)	**<0.01**	22.6 (22.3, 22.9)	**<0.01**
end-terrace	76	23.8 (23.5, 24.4)	**<0.01**	23.1 (22.5, 23.6)	0.08
mid-terrace	121	24.1 (23.8, 24.3)	**<0.01**	22.9 (22.5, 23.3)	**0.03**
purpose built flat	101	23.7 (23.3, 24.2)	**<0.01**	24 (23.6, 24.3)	**0.02**
semi-detached	227	23.9 (23.7, 24.1)	**<0.01**	22.8 (22.6, 23.1)	**<0.01**
Dwelling Age	-	-	**<0.01**	-	**<0.01**
pre-1850	24	23.0 (21.6, 23.7)	-	21.2 (20.1, 22.0)	-
1850-1899	55	23.1 (22.3, 23.5)	0.85	21.8 (21.2, 22.5)	0.06
1900-1918	40	23.7 (23.1, 24.4)	**0.05**	22.5 (21.8, 22.9)	**<0.01**
1919-1944	121	24.0 (23.8, 24.3)	**0.01**	23.3 (22.6, 23.5)	**<0.01**
1945-1964	193	23.7 (23.4, 23.9)	**0.02**	23.1 (22.8, 23.3)	**<0.01**
1965-1974	127	23.7 (23.5, 24.2)	**0.02**	23.4 (23.0, 23.6)	**<0.01**
1975-1980	64	23.8 (23.4, 24.4)	**0.02**	23.7 (22.9, 24.0)	**<0.01**
1981-1990	78	23.5 (23.3, 24.2)	**0.03**	23.0 (22.7, 23.6)	**<0.01**
post-1990	72	24.0 (23.6, 24.2)	**0.01**	23.0 (22.8, 23.6)	**<0.01**
Floor Area	-	-	**<0.01**	-	**<0.01***
less than 50 sqm	80	23.5 (23.0, 24.2)	-	23.9 (23.5, 24.4)	-
50 to 69 sqm	186	23.7 (23.5, 24.1)	0.58	23.4 (23.0, 23.6)	**<0.01**
70 to 89 sqm	202	23.9 (23.8, 24.2)	0.13	23.1 (22.8, 23.4)	**<0.01**
90 to 109 sqm	112	23.9 (23.4, 24.2)	0.26	23.1 (22.7, 23.4)	**<0.01**
110 sqm or more	194	23.5 (23.2, 23.7)	0.58	22.4 (22.1, 22.5)	**<0.01**

correlated with the floor area, with the highest value of 23.9 °C (CI: [23.5, 24.4] °C) associated with the smallest floor area level and decreasing as the floor area increases to the lowest median SIT of 22.4 °C (CI: [22.1, 22.5] °C).

Number of storeys is a significant factor for the bedroom and living room. The lowest median bedroom SIT was 22.8 °C, 1.2 °C lower than the two-storey buildings.

Table 4.9: Summary of the median standardised indoor temperatures (SIT), 95 % Confidence Interval (CI) and significance test results. The p-values (p) associated with each variable are the results of the Kruskal-Wallis test, and the p-values associated with each level of the variable are the result of the Pairwise Mann-Whitney U-tests. * indicates groups where the assumption of equal variance was not met, but stochastic dominance could be assessed. The bold font highlights cases where the p-value ≤ 0.05.

		Bedroom SIT (°C)		Living Room SIT (°C)	
	N	Med (CI 95%)	p	Med (CI 95%)	p
No. of Storeys	-	-	**<0.01**	-	**<0.01**
1	97	22.8 (22.6, 23.1)	-	23.4 (22.9, 23.7)	-
2	577	24.0 (23.8, 24.1)	**<0.01**	22.9 (22.8, 23.1)	**0.01**
3	73	23.3 (22.6, 23.8)	0.21	22.5 (21.9, 23.2)	**0.01**
4	13	23.7 (23.1, 25.8)	**0.02**	23.8 (23.4, 26.2)	**0.05**
5 or more	14	23.8 (21.5, 25.2)	0.18	24.5 (22.6, 25.5)	0.13
Traffic Problems	-	-	0.36	-	0.29
No	728	23.7 (23.5, 23.8)	-	23.0 (22.9, 23.1)	-
Yes	46	24.1 (23.3, 24.4)	0.41	23.4 (22.4, 24)	0.31
Construction	-	-	0.38	-	**<0.01***
solid masonry	156	23.8 (23.4, 24.1)	-	22.5 (22.1, 23)	-
cavity masonry	532	23.7 (23.5, 23.8)	0.87	23.0 (22.9, 23.3)	**<0.01**
timber frame	30	24.1 (23.4, 24.8)	0.51	23.7 (22.8, 24.0)	**0.02**
steel frame	13	23.1 (22.5, 24.9)	0.87	23.7 (22.6, 24.6)	**0.03**
concrete frame	13	24.6 (21.2, 25.8)	0.51	24.6 (21.9, 26.8)	**0.01**
concrete boxwall	17	23.7 (22.7, 24.4)	0.75	22.7 (21.4, 23.1)	0.89
Double Glazing	-	-	0.06	-	**<0.01**
no double glazing	47	23.5 (23.2, 24.3)	-	22.8 (22.1, 23.1)	-
less than half	43	23.1 (22.8, 23.9)	0.45	22.3 (21.8, 22.7)	0.26
more than half	81	23.8 (23.0, 24.2)	0.73	23.0 (22.3, 23.6)	0.72
entire house	603	23.8 (23.6, 23.9)	0.45	23.1 (22.9, 23.3)	0.26

The highest median living room SIT was 24.5 °C associated with 5th storey or higher buildings, although due to the small number of such cases (14), the CI is wide [22.6, 25.5] °C.

The traffic problems variable was assumed to be a possible indication of local noise or air pollution that could deter occupants from keeping their windows open. However, it is also likely that dwellings whose occupants expressed issues traffic problems are located near urban centres affected by the Urban Heat Island effect

Table 4.10: Summary of the median standardised indoor temperatures (SIT), 95 % Confidence Interval (CI) and significance test results. The p-values (p) associated with each variable are the results of the Kruskal-Wallis test, and the p-values associated with each level of the variable are the result of the Pairwise Mann-Whitney U-tests. * indicates groups where the assumption of equal variance was not met, but stochastic dominance could be assessed. The bold font highlights cases where the p-value ≤ 0.05.

		Bedroom SIT (°C)		Living Room SIT (°C)	
	N	Med (CI 95%)	p	Med (CI 95%)	p
Nature of Area	-	-	**<0.01**	-	**<0.01**
city centre	16	24.2 (23.2, 24.9)	-	23.8 (21.9, 24.6)	-
other urban centre	77	24.1 (23.7, 24.6)	0.94	23.5 (23, 23.8)	0.72
suburban residential	515	23.8 (23.6, 24.0)	0.39	23.0 (22.9, 23.2)	0.21
rural residential	102	23.4 (23.1, 23.8)	0.16	22.8 (22.5, 23.3)	0.21
village centre	42	23.1 (22.8, 23.5)	**0.01**	22.6 (21.7, 23.0)	0.12
rural	22	23.6 (21.9, 24.1)	0.16	22.5 (20.8, 23.5)	0.12
Main Heating System	-	-	**<0.01**	-	**<0.01**
boiler system with radiators	692	23.7 (23.5, 23.8)	-	23.0 (22.8, 23.1)	-
storage radiators	53	23.5 (22.8, 24.1)	0.25	23.4 (22.6, 24.0)	0.11
room heater	11	24.3 (22.5, 25.8)	0.31	23.0 (20.2, 26.3)	0.92
communal	12	25.7 (24.4, 28.0)	**<0.01**	25.8 (24.5, 27.5)	**<0.01**
Loft Insulation	-	-	0.07	-	0.65
none	20	24.2 (23.7, 24.7)	-	22.9 (21.8, 23.4)	-
less than 100mm	132	24.0 (23.7, 24.2)	0.80	23.3 (22.8, 23.6)	0.89
100 up to 150mm	209	23.9 (23.7, 24.2)	0.67	23.1 (22.6, 23.4)	0.89
150mm or more	337	23.5 (23.4, 23.8)	0.44	22.9 (22.8, 23.0)	0.89
SAP 09	-	-	0.50	-	**<0.01***
less than 30	19	23.8 (22.5, 24.3)	-	22.5 (20.2, 23.3)	-
30 to 50	176	23.8 (23.5, 24.1)	0.60	23.3 (22.8, 23.5)	0.05
51 to 70	516	23.7 (23.5, 23.8)	0.60	22.9 (22.7, 23.0)	0.08
more than 70	63	24.1 (23.3, 24.4)	0.60	23.8 (23.4, 24.5)	**<0.01**

that is not captured by this analysis. Although a statistically significant result was not observed, occupants that were influenced by traffic problems had a median temperature 0.4 °C greater than the ones that did not experience traffic problems.

Although differences in the median bedroom SIT of different construction types

were observed, they did not reach a statistically significant level. Regarding the living room SIT, significant differences were observed with the lowest median SIT of 22.5 °C (CI: [22.1, 23] °C) representing solid masonry construction. The median SIT for cavity masonry is only 0.5 °C higher, although the difference is statistically significant according to the Pairwise Mann-Whitney U-test, with a p-value < 0.01. This difference may be partly explained by the greater level of exposed thermal mass in solid wall dwellings, that has been shown to offer some protection against higher daytime indoor temperatures in modelling (Peacock et al., 2010; Petrou et al., 2019a) and monitoring (Lomas and Kane, 2013) studies, and the lower levels of airtightness (since solid wall dwellings are likely older and more leaky).

The presence of double glazing, based on how much of the glazed area of the house is double-glazed, was associated with a statistically significant difference in SIT in the living room only. With an SIT of 23.1 °C and a narrow confidence interval [22.9, 23.3] °C, double glazing across the entire house was the warmest group. The group of dwellings with double glazing in less than half of the glazed area had the lowest median SIT of 22.3 °C (CI: [21.8, 22.7] °C).

The type of local area also has a statistically significant association with the bedroom and living room SIT. For both rooms, city centre and other urban centre were associated with the highest median SIT. This might be due to the higher ambient temperatures in urban areas associated with the Urban Heat Island (UHI) effect (Heaviside et al., 2017), different proportions of dwelling types in urban compared to other regions, and potentially greater barriers to natural ventilative cooling (e.g. reduced wind speeds, greater noise levels or security concerns) in urban areas.

A significant difference was observed in the median SIT for both rooms according to the type of main heating system used. Specifically, the greatest difference was observed for the presence of communal heating. The median SIT of dwellings with communal heating was 2.0 °C higher for the bedroom and 2.8 °C for the living room (Figure 3) compared to the more common gas boiler. Even though the number of dwellings with communal heating was small (12), resulting in wide confidence intervals, there was still no overlap with the CI of the dwellings with a boiler. The

difference was also supported with a p-value < 0.01 resulting from the Pairwise Mann-Whitney U-test between the two types of heating system. This result is supported by previous work that studied the link between communal heating and indoor overheating (McLeod and Swainson, 2017), and reinforces the importance of careful planning when designing and implementing communal heating systems.

A pattern of decreasing median bedroom SIT is observed with increased levels of loft insulation, although no such pattern exists for the living room. This is likely due to bedrooms being more frequently located directly under the roof and hence influenced more by the heat transfer through that surface. Thus, adding thermal insulation to a dwelling's loft may only reduce indoor overheating risk for the top-floor rooms. For the bedroom, a difference of 0.7 °C can be observed between dwellings with no loft insulation (median SIT of 24.2 °C) and dwellings with loft insulation of 150 mm or more (23.5 °C), although it did not reach statistical significance.

A statistically significant difference between SAP 09 ratings and SIT was only observed for the living room, where a SAP rating greater than 70 was associated with the highest median SIT (23.8 °C); this was 1.4 °C higher than dwellings with SAP rating less than 30 and 0.9 °C greater than ones with rating between 51 and 70. The median bedroom SIT for the dwellings with SAP rating greater than 70 was also the highest amongst all levels, although the difference was not statistically significant. In agreement with these findings, Lomas et al. (2021) identified a significantly greater prevalence of indoor overheating risk in homes in more efficient dwellings (EPC bands A to C, compared to D to F). The findings on the association, or the lack of, between SIT and energy efficiency in this work contributes to the ongoing discussion on the potential impact of energy efficiency on indoor overheating (Chappells and Shove, 2005; Shrubsole et al., 2014).

4.2.4 Correlations Between Dwelling and Household Characteristics

Through the use of Kruskal-Wallis and Pairwise Mann-Whitney U-tests, several associations were revealed between household (Tables 4.5–4.7) and dwelling (Tables 4.8–

Table 4.11: Summary of the p-values of the Fisher's exact test that tests the significance of association between categorical variables. A statistically significant association is assumed for p-values ≤ 0.05.

	AO	AY	Inc.	Ill/dis	Emp.	Ten.	HC	Ben.	NP
Const.	**0.04**	**0.03**	0.57	**0.05**	**0.04**	**<0.01**	**0.02**	**<0.01**	0.18
DG	0.60	0.82	0.37	0.71	**0.03**	**<0.01**	0.89	**0.02**	0.61
DA	**0.05**	0.09	0.53	**0.02**	0.08	**<0.01**	**0.05**	**0.01**	**0.04**
DT	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
FA	**<0.01**	**<0.01**	**<0.01**	**0.03**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
LI	**0.04**	**0.04**	0.26	0.45	**0.01**	**0.01**	0.24	**<0.01**	**0.04**
HS	**0.02**	**<0.01**	**<0.01**	0.10	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
Area	0.53	0.92	0.12	0.68	0.09	**<0.01**	**0.01**	**0.02**	0.08
SAP	0.20	0.19	**<0.01**	0.45	0.27	**<0.01**	0.06	**0.02**	**<0.01**
Stor.	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
Tr.	0.90	0.74	0.29	0.40	0.83	0.68	**0.01**	0.43	**<0.01**

Row name disambiguation: Const. = Construction, DG = Double Glazing, DA = Dwelling Age, DT = Dwelling Type, FA = Floor Area, LI = Loft Insulation, HS = Main Heating System, Area = Nature of Area, SAP = SAP 09, Stor. = Storey, Tr. = Traffic.
Column name disambiguation: AO = Age band of Oldest occupant, AY = Age band of Youngest occupant, Inc. = Income, Ill/dis = Illness or Disability, Emp. = Employment Status, Ten. = Tenure, HC = Household Composition, Ben. = Means tested or certain disability related benefits, NP = Number of People.

4.10) characteristics with the bedroom and living room SIT. Prior to any causation being attributed to individual variables analysed, any correlation between variables should be explored.

Table 4.11 provides a matrix of p-values resulting from the Fisher's Exact test with the null hypothesis of independent variables (see Table 4.3 for the null hypothesis). By assessing the association of dwelling characteristics against household characteristics, a statistically significant association is obtained for each variable with at least one other variable. For the dwelling type (DT), floor area (FA) and storey (Stor), a significant association was observed with every household variable. Tenure (Ten.) was associated with every dwelling characteristic, with the exception of traffic (Tr.). This was equally true for occupants on means tested or certain disability related benefits (Ben.). A significant association between SAP 09 rating, income (Inc.) and tenure (Ten.) was established, although the same was not true for employment status (Emp.). The presence and level of loft insulation (LI) was associated with both age-related categories (AO, AY), employment status, tenure, benefits and no of

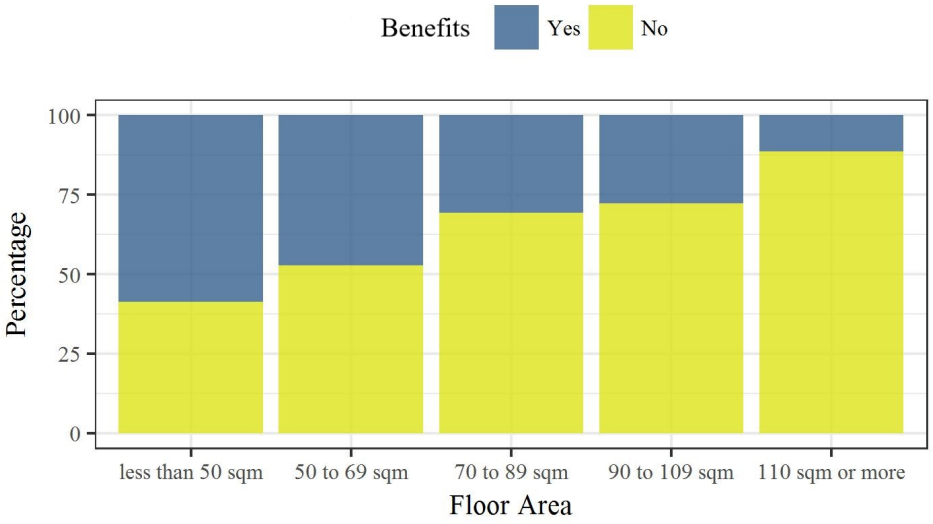


Figure 4.3: Bar plots of association between floor area and households on means tested or certain disability benefits.

people in the dwellings (NP).

A further investigation of the relationship between household occupants on means tested or certain disability benefits and floor area is displayed in Figure 4.3.[2] With increased floor area, the percentage (and probability) of a dwelling's occupants being on means tested or certain disability benefits decreases. As described in Sections 4.2.2–4.2.3, both variables were associated with the living room SIT. Increased floor area was associated with a decrease in living room SIT (Table 4.9). This might be expected, since given the same solar and internal gains, a smaller room will reach a higher internal temperature. Table 4.7 showed that the median living room SIT of households with occupants on means tested, or other disability benefits was higher than that of households with no occupants on such benefits. A plausible explanation for this association is that individuals with disabilities may spend more time at home, resulting in increased internal gains, and their limited mobility may lead to reduced window operation (Vellei et al., 2017). In addition, for some disabilities, there might be emphasis in keeping the home warm (Snell et al., 2015). Since the two explanatory variables are correlated (Figure 4.3), even if only one of these variables has a causal effect on indoor temperature, an association

[2]Similar analysis could not be performed for most variables due to data protection restrictions imposed by the data provider.

between the SIT and other variable would be expected. Thus, it is not possible to attribute causality without further investigation, that would likely necessitate the collection of data on occupant behaviour.

4.2.5 On the Use of Standardised Indoor Temperature

The monitored indoor temperature was standardised prior to the investigation of its association with dwelling and household variables to control for the confounding effect of regional weather, an approach informed by previous studies (Hamilton et al., 2017; Oreszczyn et al., 2006; Wilkinson et al., 2001). The standardisation relied on a regression-based approach, whose use in controlling for confounding is well-established. The efficacy of twelve regression models was evaluated using R^2, and a model was selected based on its performance and simplicity (Section 4.1.1). A pertinent question to the use of this approach is: *how can it be checked or validated?*

The topic of validation has not been discussed in previous studies that utilised this method. A metric may be used to quantify the predictive performance of the regression model when evaluated against the empirical data. Such a metric, R^2, was used to compare the candidate regression models. However, a threshold that indicates whether the model is valid, or good enough, does not exist. Developing such a threshold would require a thorough investigation that might rely on detailed empirical data where other factors that can influence the indoor temperature are also monitored. This could enable a researcher to attribute the effects of different variables on indoor temperature, and thus deduct the effect associated with the ambient conditions. Collecting such a detailed dataset from in-situ measurements in homes is challenging and expensive. Synthetic data that adequately represent occupant behaviour could provide an alternative. Another form of validation is the comparison of the results derived from this method with findings from other empirical and modelling studies, and against established building physics theory. Such a comparison was carried out within this chapter and findings from this work were in good agreement with existing knowledge and published studies, providing evidence in support of the selected statistical procedure.

4.2.6 Categorical Variable Classification

The statistical analysis (Sections 4.2.2–4.2.3) revealed that whether a parameter was associated with the SIT depended on the room being assessed. Therefore, the choice of classifiers depends on the room being modelled.

In Section 4.2.6.1 a set of classifiers are identified and discussed for each room, based on the method outlined in Section 4.1.4. Since the rest of this work will focus on the living room, as discussed in Section 3.3, the living room classifiers are used to select of a homogeneous group of 4M dwellings and complete Step 2 of the calibration framework in Section 4.2.6.2.

4.2.6.1 Classifier Selection

For the living room, significant differences in SIT were discovered for the dwelling type, dwelling age, floor area, storey, construction, presence of double glazing, nature of area, main heating system and SAP 09. Differences in dwelling types, wall construction and glazing type are modelled explicitly. Since all three variables were found to be significantly associated with the living room SIT, they should be considered as classifiers where each cluster consists of a single wall, glazing and dwelling type. Differences in the nature of the area are partly represented through the terrain model input, which assumes differences in local wind speed due to the surrounding urban form. However, it does not capture differences in outdoor temperature due to the Urban Heat Island (UHI) effect. This should again be used as a classifier, and the appropriate local weather data should be used for each region to capture variation due to UHI. The storeys of archetypes are fixed within UK-HSM to the most common number of storeys for each dwelling type in the building stock. Therefore, the use of different dwelling types partly captures differences in storeys that vary between archetypes, but not within archetypes. This may be used as a secondary classifier in calibration, together with the dwelling type. Dwelling age and SAP 09 are not modelled explicitly, but their effects are partly represented by the building fabric and air permeability properties. Differences might still remain, especially in the case of dwelling age and this may also be used as a secondary classifier, if enough data is available for the calibration. Differences in the main

heating system are not modelled explicitly, and the heating system is assumed to be inactive over summer. Given the significant differences in SIT of the various heating systems, this should also be used as a classifier. The effect of varying floor area or loft insulation thickness is modelled explicitly through the use of continuous model inputs, thus further classification is not needed.

The same process of identifying classifiers applies for the bedroom. However, the statistical analysis did not reveal a significant association between bedroom SIT and the wall or glazing type. Following this process, one may choose not to use these variables as classifiers when modelling the bedroom SIT. Alternatively and with an abundance of caution, they may still be used as classifiers since the non-significant result could be an artefact of the analysis; for example, the extent of double glazing in the dwelling (Table 4.9) does not specify which rooms are double-glazed.

Similarly to the dwelling characteristics, several household characteristics were found to be significantly associated with the living room or bedroom SIT (Tables 4.5–4.7). While many dwelling characteristics are modelled explicitly, most household characteristics are not. UK-HSM offers a choice between two occupancy types: (i) A family of four and (ii) Two pensioners. Each option is associated with a different set of assumptions regarding the number of occupants, their presence and activity schedule, in the effort to represent some of the occupant diversity that exists within the stock. However, it does not allow for the presence of just a single pensioner or a family of three. Looking at household composition, the modelled occupancy types match the two largest categories (assuming that people aged 60 or over are pensioners); yet, these two categories only account for 45 % of the sample. Classifying based on all household characteristics that were shown to be significantly associated with summer indoor temperatures, given the current structure of UK-HSM, would significantly reduce the empirical sample that can be used. On the other hand, ignoring the household characteristics from the classification process would result in some of this variation inappropriately falling under the electrical gains and window operation uncertainty, while it is the result of structural assumptions of the model. As an example of inappropriate uncertainty attribution, fewer occupants than assumed in

the model will possibly result in less metabolic gains. Since this will not be captured by the model assumptions, this uncertainty might be lumped on the electrical gains model input.

4.2.6.2 4M Classification

The process of using the classifiers identified in the previous section to segment the 4M dwellings into potentially homogenous groups is visualised in Figure 4.4.

The first classifier was the *dwelling type*, shown to be a classifier for the bedroom and living room, and a categorical model input in the UK Housing Stock Model (UK-HSM). This results in four clusters with 22 detached, 29 end-terrace, 55 mid-terrace and 87 semi-detached homes with temperature data in at least one of the two rooms. Due to its larger sample size, the rest of the classification focused on the group of semi-detached dwellings. While the exact location of the homes was not available, they were scattered within the city of Leicester, as shown in Figure 3.5(b), and it was assumed that the same *terrain*, Urban, applied to the entire sample. Another potentially important classifier is the presence of communal heating, however, this heating system was not present in any of the semi-detached dwellings. Due to the differences revealed in the factors that may influence the living room and bedroom temperatures, data from the two rooms were treated separately. For the rest of this process, only classifiers relating to the living room were considered due to the focus on MDLRT for the rest of this thesis (see Section 3.3 for a discussion on this choice).

The *wall type* and *glazing type* were both significantly associated with the living room SIT and were thus used as classifiers, resulting in five clusters. Due to their small sample size, the single glazing clusters were not considered further. The data within each double glazing cluster were cleaned following the procedure described in Section 3.4.3.2, resulting in the three final clusters: (i) Filled Cavity Wall (N = 26), (ii) Unfilled Cavity Wall (N = 8) and (iii) Solid Wall (N = 24). Two classifiers, the *number of storeys* and *dwelling age*, identified to be potentially important in Section 4.2.6.1 were not used. This was a pragmatic choice since further classification would result in groups of dwellings that were too small, and where extreme values within these groups could significantly influence the calibration process. The number

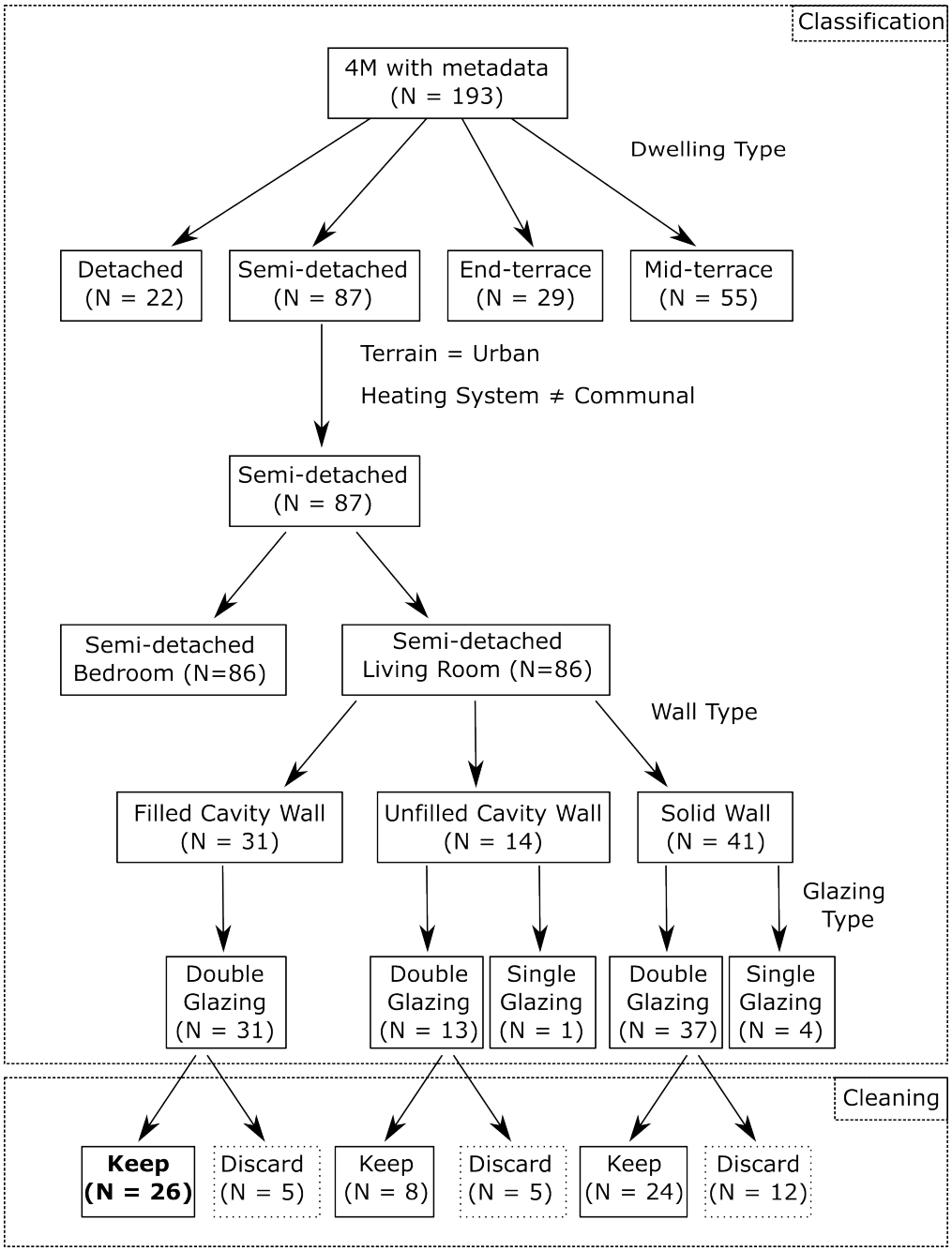


Figure 4.4: Flowchart of the first stage of the classification process and the subsequent cleaning. In bold font is the cluster selected to be used for the Bayesian calibration step.

of storeys was thought to have a small effect, since the calibration focused on living rooms which in most cases are located on the ground floor. While it is possible that a difference might exist between having one or two storeys above the living room, it was not thought to be as significant as that of other variables considered. The effects

of the dwellings' age were partly captured through the wall and glazing type, and further segmentation would reduce the number of dwellings per group, therefore, it was not considered necessary.

Classification based on household variables was not performed. The only household variable that may be partly captured in UK-HSM is the *household composition* which was not shown to be significantly associated with summer living room indoor temperatures. As with the number of storeys and dwelling age, it was preferred to not segment the clusters further due to their small size. From the final three clusters, the rest of this thesis will concentrate on group of dwellings with filled cavity walls.

4.3 Limitations

To study the statistical associations between summer indoor temperature and key dwelling and household characteristics, and subsequently segment the 4M dataset in accordance with the Bayesian calibration framework proposed in Section 3.2, this work has relied on established methods of statistical analysis and one of the most comprehensive datasets available (2011 EFUS). Yet, this study has a number of limitations.

Since the standardisation of indoor temperature focused on a single outdoor temperature during the summertime, it is unclear whether the trends observed at a daily-mean temperature of 20 °C would be similar for other standardisation temperatures. This limitation can be addressed in future work and would be most informative with datasets collected during summers warmer than that of 2011.

By design, the statistical analysis did not seek to identify causal relationships, only associations. As highlighted in Section 4.2.4, correlations between explanatory variables exist, and it is expected that some of the statistically significant association revealed in this work are the result of confounding. Resolving confounding and establishing causal relationships would have improved this work, and would have resulted in better-informed classification. However, this would require a different set of methods and data.

Another limitation, common amongst studies that utilise hypothesis testing, is

the presence of Type I and II errors (Greenland et al., 2016). For example, performing the Kruskal-Wallis test repeatedly at a significance level of 5 % suggests that the null hypothesis may falsely be rejected (type I error) in 5 % of the cases. Thus, statistically significant results should be treated with caution, and it is best that the outcome of a hypothesis test is considered together with the associated median SIT values and their confidence interval.

In carrying out the statistical analysis and overheating risk assessment, although local weather data were used, they did not necessarily represent the ambient weather conditions at the exact location of each dwelling. This is especially true for dwellings located in urban areas, as the weather data may not effectively capture the influence of the urban heat island effect or the local microclimate (Mavrogianni et al., 2009).

On the comparison carried out between stated thermal discomfort and predicted overheating risk, it should be highlighted that for the dataset analysed in this thesis (and the one analysed by Lomas et al. (2021) which was compared with the results of this work), data collection on stated thermal comfort was not carried out systematically during the summer period and could, thus, be influenced by *recall bias*. Within the field of epidemiology, *recall bias* can be defined as a "differential misclassification bias and the risk estimate may be biased away from or towards the null" (Coughlin, 1990). In the case of reported summer thermal comfort, occupants might be more likely to underreport summer thermal discomfort if they were satisfied with the thermal environment (or feeling cold) during or preceding the interviews.

The 2011 EFUS dataset was used for the statistical analysis, while the 4M dataset was used for the remaining stages. Therefore, an assumption was made that the findings of the statistical analysis also applied to the 4M dataset. Given that 2011 EFUS is considered to be a representative sample of the English housing stock with approximately 800 dwellings, this assumption seemed reasonable.

Finally, as discussed in Sections 4.2.6.1-4.2.6.2, not all categorical variables found to be associated with summer indoor temperatures were used for classification. This was partly due to the UK-HSM model structure, and as a result of the limited empirical data available for calibration in the 4M dataset.

4.4 Summary

The statistical analysis of the 2011 Energy Follow-Up Survey, presented in this chapter, revealed a statistically significant association between the standardised summer indoor temperature and several of the dwelling and household characteristics assessed (Sections 4.2.2–4.2.3). Whether variables exhibited such a relationship depended on the room being assessed. For the living room, a significantly greater indoor temperature was associated with purpose built flats, a high SAP 09 rating, the presence of double glazing in more than half of the dwelling, and homes with at least one occupant on means tested or certain disability benefits. Bedroom summer indoor temperature was higher in mid-terrace dwellings, households with young children, more than three occupants and where one or more occupants work full time. The summer indoor temperature in both rooms was significantly higher for homes rented from a local authority, located in the city centre or other urban centre, and with communal heating. The correlation between household and dwelling variables was explored, and its implications were discussed (Section 4.2.4).

The indoor overheating risk was also quantified based on the criteria defined in CIBSE's Technical Memorandum 59 (Section 4.2.1). It was demonstrated that for the relatively cool summer of 2011, the prevalence of indoor overheating according to Criterion 1 was 2.5 %. However, when considering Criterion 2 almost 26 % of dwellings were found to overheat. These results were not in good agreement with the occupants' stated thermal discomfort, reinforcing concerns regarding the effective quantification of indoor overheating risk (for example, the overheating threshold was not exceeded in 32 out of 61 homes whose occupants reported thermal discomfort).

Beyond its contribution to the field of indoor overheating research, the statistical analysis has also resulted in a set of classifiers (Section 4.2.6). The classifiers (dwelling, wall, glazing type, terrain and main heating system) were used during the categorical variable classification to segment the 4M dataset and select a single group of potentially homogeneous dwellings; this outcome informs Steps 3 and 4 of the Bayesian calibration framework, whose implementation is described in Chapters 5 and 6 respectively.

Chapter 5

Stochastic Characterisation

This chapter focuses on Step 3 of the Bayesian calibration framework, the stochastic characterisation of the UK Housing Stock Model (UK-HSM) inputs. Its aim is to identify probability distributions for all continuous model inputs of UK-HSM, specific to the group of semi-detached dwellings identified in Chapter 4. As visualised in the Figure 5.1, the probability distributions will inform Chapters 6 and 7, and contribute to the completion of the second research objective (Section 1.3).

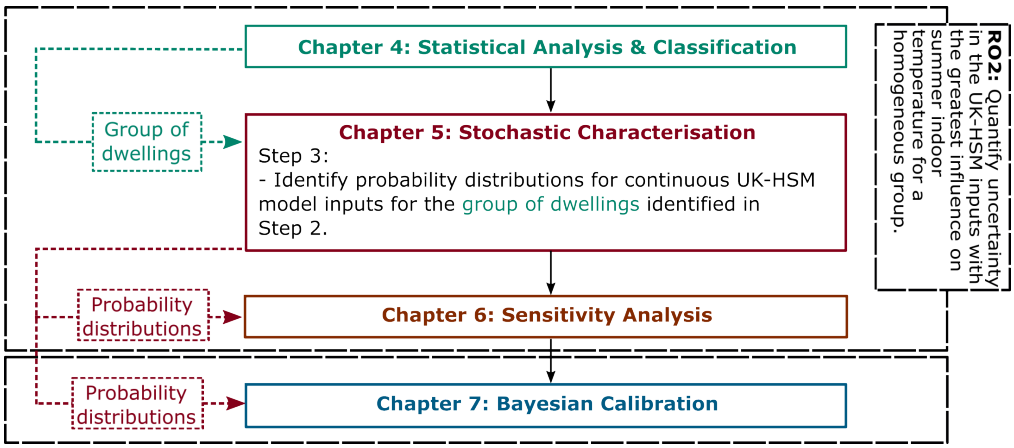


Figure 5.1: Chapter 5 flowchart. This is an abridged version of Figure 1.3, focusing on Chapter 5 and its outputs. RO2 is a shortened version of Research Objective 2. UK-HSM stands for UK Housing Stock Model.

To carry out the stochastic characterisation, a set of novel methods for specifying model input probability distributions are introduced in Section 5.1. The choice of approach depends on whether empirical data that could inform the probability distribution exist, and in what format (Figure 5.2). Where possible, the probabil-

ity distributions were based on empirical evidence. A description of the dataset and method used to identify a suitable distribution, along with a discussion on the appropriateness of this distribution, is provided for each model input of UK-HSM in Section 5.2. The twelve probability distributions are also summarised in Section 5.2.13. The strengths and limitations of this work are discussed in Section 5.3, while a summary of the chapter is offered in Section 5.4.

5.1 Methods

The flowchart in Figure 5.2 depicts the process of determining the method used for distribution identification depending on the data available. To identify the probability distribution that best describes an empirical dataset whose tabulated values are available, a distribution-fitting method was implemented. This novel method, within the field of building modelling, is described in Section 5.1.1 and formed the basis for a publication (Petrou et al., 2021b). In the case that empirical data were available only in a graphical form, the process described in Section 5.1.2 was used. If empirical data is not available, the probability distributions had to be assumed based on judgement, experience and the information known about the uncertain variable (Mun, 2012) as summarised in Section 5.1.3.

In all cases where empirical data were available, they were selected using the same classifiers identified in Section 4.2.6.2. This is to capture any associations that might exist between the continuous and categorical model inputs. However, this was not always possible, either because of the lack of metadata or because of the limited sample size.

Categorical variables were used to inform the distribution of the continuous model inputs in some cases. When this was the case, a weighted distribution was constructed based on the prevalence of those categorical features within the homogeneous group of dwellings. For example, the Roof U-value is a continuous model input of UK-HSM. One categorical variable that informs the distribution of Roof U-value is the loft insulation thickness. A question within the 4M household survey asked occupants to state which of the following five categories describes

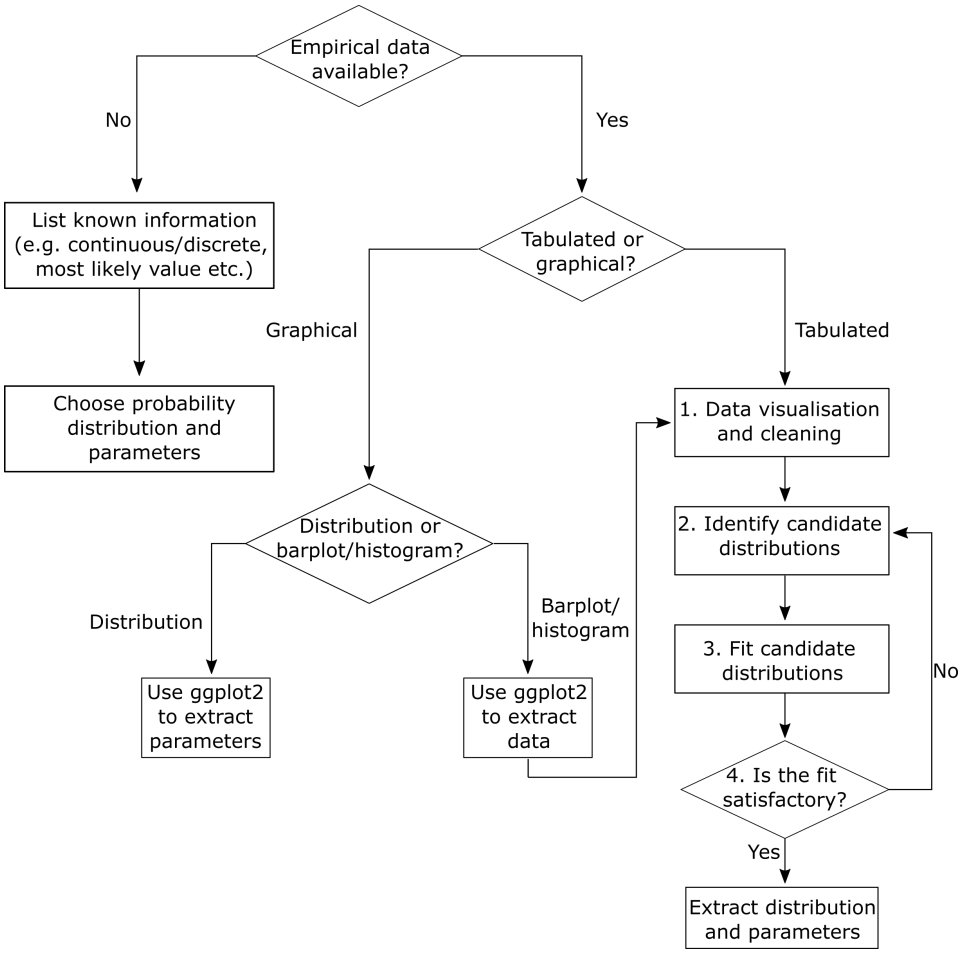


Figure 5.2: Workflow diagram of the stochastic characterisation process.

their loft insulation level: (i) No insulation, (ii) Up to 50 mm, (iii) 50–100 mm, (iv) 100–200 mm and (v) greater than 200 mm. As will be discussed in more detail in Section 5.2.3, a probability distribution was assigned to each category. A weighted distribution was then constructed by sampling a large number of times (10,000) from the five probability distributions in ratios equivalent to the relative ratios of loft insulation thicknesses within the filled cavity wall dwellings in the 4M dataset.

5.1.1 Inferring Probability Distributions with Tabulated Empirical Data

This section will discuss a novel method within the built environment developed for the identification of appropriate distributions for model inputs when tabulated empirical data are available. The code for implementation has been made publicly

available (Petrou, 2021). The proposed method consists of four main steps:

1. Data visualisation and cleaning: A histogram is used to visualise the data and inform the cleaning of the data. This is preferred over automated procedures based on the data's interquartile range or standard deviation when the data do not appear to be normally distributed. It is easier to identify and reject outliers when there is already an established model of the measured variable and its distributional form is known. However, this is often not the case and automatic methods of outlier detection, such as the Chauvenet's Criterion[1] that assumes a normal distribution would be inappropriate (Hughes and Hase, 2014). Given that many of the model parameters within the built environment field have a physical meaning, it might be better to compare measured extreme values with their theoretical equivalents derived from the understanding of the physical system being studied. For example, the measured U-value of a wall may be compared to the U-value calculated based on its construction and thickness.

2. Identify candidate distributions: Once outliers are removed, the data's empirical distribution, together with the "Cullen and Frey" graph of kurtosis against the square of skewness are used to identify candidate distributions (Figure 5.4). *Skewness*, is a measure of symmetry, with a value of zero indicating a fully symmetric distribution (Reimann et al., 2008). *Kurtosis*, indicates how heavy the tails of a distribution are (i.e. how flat or peaked the distribution is) with a value of three for a normal distribution (Reimann et al., 2008).[2] By plotting the kurtosis and square skewness of the collected data on a graph and overlaying the values that common distributions would take, one can infer the candidate distributions that may best describe the data. Since skewness and kurtosis may easily be affected by extreme values, one can employ a bootstrap technique of random sampling (at least 1000 samples) with replacement to plot multiple possible values on the Cullen and Frey graph (Hesterberg, 2011; Delignette-Muller and Dutang, 2015). If none of the

[1] As described by Hughes and Hase (2014), Chauvenet's Criterion relies on the assumption that the data follow a normal distribution, and "a data point is rejected from a sample if the number of events we expect to be farther from the mean than the suspect point, for the sample's mean and standard deviation, is less than half".

[2] The normalised kurtosis of a normal distribution is 0, it takes a negative value for a "peaked" and a positive value for a "flat" distribution (Reimann et al., 2008).

distributions that appear on the Cullen and Frey Graph provide an adequate fit to the data, other distributions could be explored. An alternative to this approach is to fit multiple distributions (e.g. the ones typically included on the Cullen and Frey Graph) and use step 4 to choose the most appropriate one.

3. Fit candidate distributions: The candidate distributions were fitted to the data using the R package **fitdistrplus** (Delignette-Muller and Dutang, 2015). To fit the candidate distributions to the data, several methods exist. In this work, the commonly used Maximum Likelihood Estimation (MLE) method was used which is the default option in the library **fitdistrplus** (Delignette-Muller and Dutang, 2015).[3] A probability distribution function, specified as $f(x_1|\boldsymbol{\phi})$, quantifies the probability of observing data point x_1, given the distribution parameters $\boldsymbol{\phi}$ (i.e. assuming that $\boldsymbol{\phi}$ are known) (Portet, 2020). Fitting a distribution to a set of known data points is the inverse problem, where the observations are known, and the parameters are unknown. Assuming $\mathbf{x} = x_{i=1}, \ldots, x_N$ independent and identically distributed (i.i.d.) observations, the likelihood function is a function of parameters $\boldsymbol{\phi}$ defined as (Smith, 2013):

$$\mathcal{L}(\boldsymbol{\phi}|\mathbf{x}) = \prod_{i=1}^{N} f(x_i|\boldsymbol{\phi}). \tag{5.1}$$

The likelihood function quantifies the probability of obtaining the observed data $\mathbf{x}$, if the parameters $\boldsymbol{\phi}$ had a specific value (Portet, 2020). By employing an optimisation algorithm, for any candidate distribution ($f(\cdot|\boldsymbol{\phi})$) and observed data ($\mathbf{x}$), parameters $\boldsymbol{\phi}$ are optimised in order to maximise the log of the likelihood function (Delignette-Muller and Dutang, 2015). This process is repeated for all candidate functions separately to identify the parameters and density function that best describes the data.

4. Identify the candidate distribution with the best goodness-of-fit: Finally, drawing from Information Theory, the Akaike Information Criterion (AIC) and its derivates are used to identify the best fitting distribution (Burnham and Anderson,

[3]Other methods may be preferred under specific circumstances, such as when it is desirable to place more weight on one of the tails of the distribution. One such case might be when the interest is in the least energy efficient dwellings whose building characteristics (e.g. permeability) are described by the tails of the distributions.

2004). AIC is defined as (Burnham and Anderson, 2004):

$$\text{AIC} = -2\log(\mathcal{L}(\hat{\boldsymbol{\phi}}|\mathbf{x})) + 2K, \tag{5.2}$$

where $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimate of parameters $\boldsymbol{\phi}$, $\log(\mathcal{L}(\cdot|\cdot))$ is the log likelihood and K is the number of distribution parameters (as an example, the normal distribution has two parameters: the mean and the standard deviation). For a collection of R candidate distributions (or models more generally), the best distribution given the data $\mathbf{x}$ is the one with the minimum AIC value (Portet, 2020). For a small number of observations, where $K > (N/40)$, the corrected AIC may be used instead (Portet, 2020):

$$\text{AICc} = \text{AIC} + \frac{2K(K+1)}{N-K-1}, \tag{5.3}$$

with AICc approaching AIC as N approaches infinity. While Equations 5.2 – 5.3 enable the ranking of candidate distributions, the actual values AIC or AICc are not themselves easily interpretable. However, some more interpretation is possible through the manipulation of the estimated AIC values. Rescaling the AIC (or AICc) of each candidate distribution j, with regard to the minimum AIC (AIC_{min}) results in an estimate (Δ_j) of the information loss when distribution j is selected instead of the best candidate distribution; effectively quantifying the strength of the AIC differences (Burnham and Anderson, 2004):

$$\Delta_j = \text{AIC}_j - \text{AIC}_{min}. \tag{5.4}$$

Burnham and Anderson, 2004 suggested that:

- Models with $\Delta_j < 2$ have substantial support (evidence)
- Models with $4 < \Delta_j < 7$ have considerably less support
- Models with $\Delta_j > 10$ have almost no support

Therefore, an alternative to the best candidate distribution (the one with the lowest AIC) with a Δ_j less than 2 may be considered a good alternative while one with Δ_j

greater than 10 should not be used. Portet (2020) warns that these guidelines should be treated with caution if, for example, a large number of candidate distributions are assessed. Instead, one can go further and estimate the Akaike weights (or "weight of evidence") (Burnham and Anderson, 2004; Portet, 2020):

$$w_j = \frac{\exp(-\Delta_j/2)}{\sum_{j=1}^{R} \exp(-\Delta_j/2)}, \tag{5.5}$$

where $\exp(-\Delta_j/2)$ is the distribution likelihood. The quantity, w_j is the probability that distribution j is best amongst the candidate distributions given the observations $\mathbf{x}$. Finally, a direct comparison between two candidate distributions can be carried out by computing their evidence ratio w_j/w_k, quantifying the strength of evidence of model j over model k.

While the AIC and its derivatives can help a modeller determine which distribution (or model) is best amongst the candidates, they do not provide any information on whether a distribution's fit is sufficient for its intended purpose. This may be decided by visualising the theoretical data (originating from the best distribution) against the empirical data in four plots (for an example see Figure 5.5):

1. Histogram with theoretical densities
2. Quantile-Quantile plot (Q-Q plot)
3. Empirical and theoretical Cumulative Distribution Function (CDF) plots
4. Percentile-Percentile plot (P-P plot)

A histogram of the data superimposed by the theoretical densities provides a quick and comprehensive check of the distribution fit. In a Q-Q plot, the theoretical quantiles from the assumed distribution are plotted against the empirical quantiles and a straight line would provide support for the assumed distribution. A P-P plot, will instead have the probabilities of the hypothetical distribution plotted against the probabilities of the empirical data at fixed quantiles. While a Q-Q plot is useful for exposing discrepancies in the tails of the distributions, a P-P plot is more focused on the main body (Reimann et al., 2008). The empirical CDF is a step function, where as the number of data points increase it should approximate the underlying distribution function (Reimann et al., 2008).

5.1.2 Inferring Probability Distributions with Graphical Empirical Data

In some cases, tabulated empirical data may not be available, but a visualisation is available (e.g. bar plot or density plot). One can extract useful information from such a graphical representation of the empirical data, and this was required in a few instances within the stochastic characterisation stage.

The general process required overlaying the visualisation onto a set of axes within the R package **ggplot2** (Wickham, 2016). The axes' scale was set to match that of the visualisation. In the case of trying to extract the parameters of a distribution, a set of distributions were trialled and their goodness-of-fit was visually assessed (see Section 5.2.11). If instead the aim was to extract the values from a bar plot, a digital "ruler" was used to estimate the number of counts for each bar (see Section 5.2.12).

While this is arguably an approximate method, in the absence of tabulated empirical data the use of graphical information in this manner was considered a more thorough and accurate method than ignoring their presence and postulating a distribution in the manner described in Section 5.1.3.

5.1.3 Inferring Probability Distributions without Empirical Data

As Mun (2012) suggests, to choose the appropriate probability distribution for an uncertain variable (in this case model input), one must first list important known information about it, review the descriptions of probability distributions and select one that characterises the variable based on their best understanding of it. In the last step, both the *distributional form* and *parameters* need to be chosen. In doing so, a series of questions may need to be answered:

1. Is the variable continuous or discrete?
2. Is some value of the uncertain variable most likely?
3. Is the value of the uncertain variable equally likely to be above or below the most likely value?
4. Is the uncertain variable more likely to be in the vicinity of the mean than further away?

5. Is value of the uncertain variable equally probable within a fixed range?

Once the form of the probability distribution is chosen, the parameters that define it must be selected. For example, if the answer to (1) is "continuous" and the answers to (2), (3), (4) are "Yes", a normal distribution is likely to be a good option. Based on any available information, a mean and standard deviation are chosen as the distributional parameters. On the other hand, for a "continuous" variable where only (5) is true, a uniform distribution would be a better choice and based on the values that define the uncertain variable's fixed range, the distributional parameters can be specified.

Identifying probability distributions in such a way is far from perfect, and hence the empirical route was preferred where possible. However, whether the chosen distribution will have a significant impact on the QoI will be revealed by the sensitivity analysis and important model inputs will subsequently be calibrated.

5.2 Results

The following sections present the outcomes of identifying a probability distribution function for each model input of UK-HSM. Where possible, the characterisation was based on data collected from groups of dwellings that adhered to the same set of classifiers used in Section 4.2.6.2. To demonstrate the novel method of identifying probability distributions when tabulated data are available, the entire process (as described in Section 5.1.1) was implemented for the wall U-value, with outcomes of each step presented and discussed in Section 5.2.1. For other model inputs where this method was applied, the histograms used to visualise and clean the data were placed in the appendices (Section H) for brevity. In addition, since the computational burden of fitting multiple distributions in this analysis was relatively small, Weibull, gamma, normal and lognormal distributions were fitted in all model inputs, without consulting the Cullen and Frey graph.

5.2.1 Wall U-value

During 2012-2013, fieldwork commissioned by the UK Department of Energy and Climate Change (which in 2016 became part of the Department for Business, Energy

and Industrial Strategy (BEIS)) aimed to provide an assessment of the thermal performance of walls in English dwellings and compare it to the theoretical values (Hulme and Doran, 2014). The Building Research Establishment (BRE) led the

Table 5.1: Summary statistics, mean and percentiles, of wall U-value measurements (Hulme and Doran, 2014). Other (or non-standard) solid walls are solid brick walls with thickness $\geq$ 330 mm or non-brick solid walls, whereas standard solid walls are brick walls with thickness less than 330 mm (Hulme and Doran, 2014).

	Wall Type	**No.**	**Wall U-value (W/m^{2}K)**				
			2.5 %	25 %	50 %	75 %	97.5 %
1	Filled cavity	109	0.3	0.6	0.7	0.8	1.2
2	Other solid	33	0.6	1.1	1.4	1.6	2.1
3	Standard solid	85	1.0	1.4	1.7	1.9	2.2
4	Unfilled cavity	50	0.8	1.3	1.5	1.7	2.0

data collection. For approximately 300 dwellings, in-situ measurements were taken using heat flux plates (Hukseflux HFP01) and surface temperature measurements for a period of two weeks. The homes were a sub-sample of the 2010/11 English Housing Survey (EHS). Two measurements were taken for each dwelling as far away as possible from any thermal bridges. For this work, a 6 % adjustment was applied to the raw data, since following the publication of the original report it was discovered that the heat flux plates read 4-8 % lower than intended (BRE, 2016). Table 5.1 provides a summary of the data, where the U-value is the arithmetic mean of the two measurements taken at each dwelling and following the 6 % correction.

The corrected arithmetic mean of the measured filled cavity wall U-values is visualised in Figure 5.3. The solid vertical line represents the empirical median based on the data collected, while the dashed line is the theoretical value based on Appendix S (RdSAP) of SAP 2012 (BRE, 2014). In an update specifically to Appendix S of SAP 2012 (BRE, 2019), the theoretical value has now been changed to reflect the findings of Hulme and Doran (2014). However, the theoretical value prior to the update was included in Figure 5.3 to highlight the magnitude of differences between theoretical and empirical values for commonly used model inputs. Differences were even greater for the solid wall U-value (see Petrou et al. (2021a) and Appendix Section H.1).

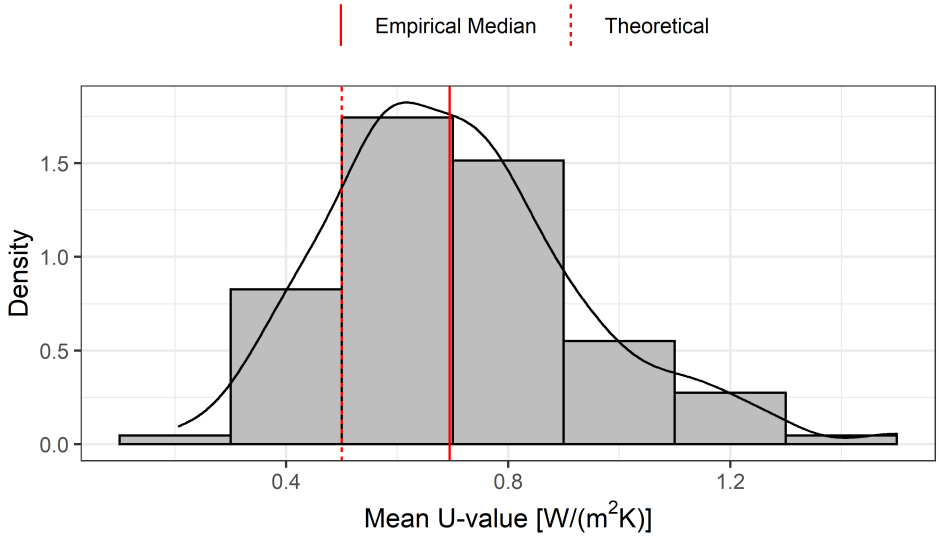


Figure 5.3: Histograms and density lines of the measured wall U-value, following a 6 % correction. Data from Hulme and Doran (2014). The theoretical line is based on RdSAP of SAP 2012 (BRE, 2014).

By simply inspecting the filled cavity wall histogram, the data distribution seems to be positively skewed, as the right tail is longer than the left. The lowest measured value is $0.2\,\mathrm{W/(m^2K)}$ while the largest value is $1.5\,\mathrm{W/(m^2K)}$. While a value of $0.2\,\mathrm{W/(m^2K)}$ is well within the expected theoretical values of well-insulated cavity walls (BRE, 2014), a value of $1.5\,\mathrm{W/(m^2K)}$ is rather high (BRE, 2014; BRE, 2016). This high value could be the result of surveyors incorrectly classifying the wall as filled-cavity or placing the heat flux over a thermal bridge. However, lower than nominal levels of insulation and poor workmanship could also lead to a worse than intended thermal performance.

To determine which distribution should be fit, the skewness and kurtosis of the filled cavity wall sample was estimated and plotted in a Cullen and Frey graph shown in Figure 5.4. The sample's values, indicated by the blue dot, suggest that the gamma, Weibull and lognormal distributions are likely to provide a good description of the collected data (this will be determined by the results associated with goodness-of-fit in Step 4 of the distribution fitting process). To account for the uncertainty in the sample skewness and kurtosis, a non-parametric bootstrap analysis was run 1000 times, with the results shown as yellow rings in Figure 5.4. Many of the points lie

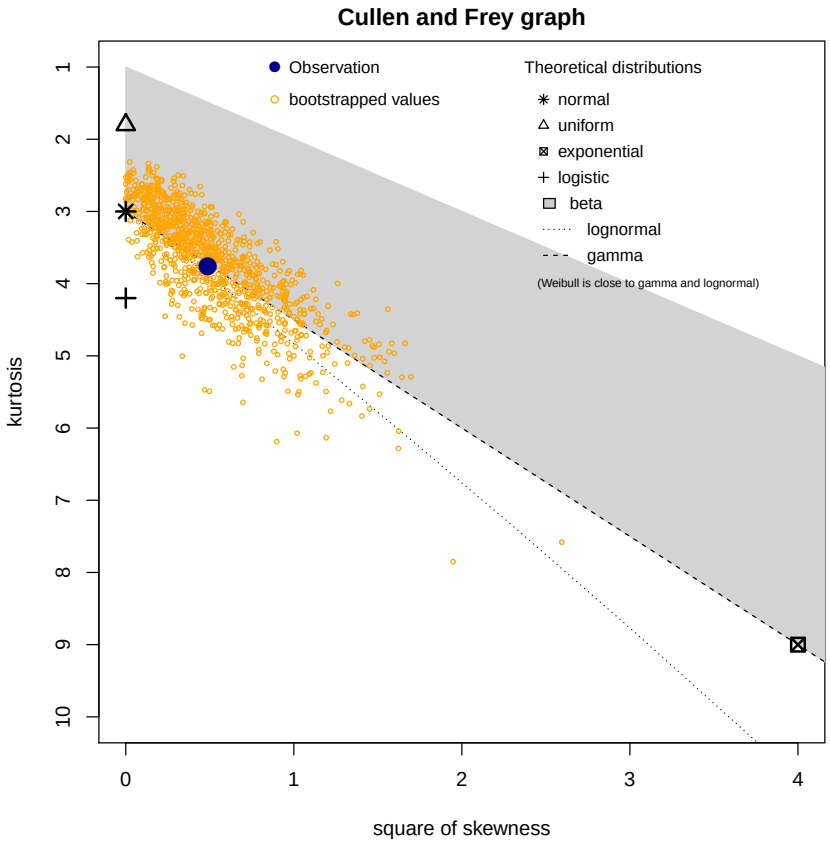


Figure 5.4: Cullen and Frey graph of kurtosis against square of skewness.

within the shaded area that represents possible kurtosis and squared skewness values the beta distribution can take. However, as the beta distribution is bound within the [0, 1] interval it was not chosen as a candidate distribution. A few bootstrap points concentrated close to the kurtosis of three and squared skewness of zero, encouraging the inclusion of the normal in the candidate distributions.

Table 5.2: Distributions for wall U-value ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j). P1 and P2 represent the parameters of the fitted distribution, stated to two significant figures.

Wall Type	Distr.	AIC	Δ_j	w_j	P1	P2
Filled cavity	gamma	-16.07	0.00	0.75	shape = 9.5	rate = 13
	lnorm	-13.72	2.34	0.23	meanlog = -0.4	sdlog = 0.33
	norm	-8.04	8.03	0.01	mean = 0.71	sd = 0.23
	weibull	-7.59	8.47	0.01	shape = 3.2	scale = 0.79

A summary of the AIC derivatives for all the fitted distributions is provided in

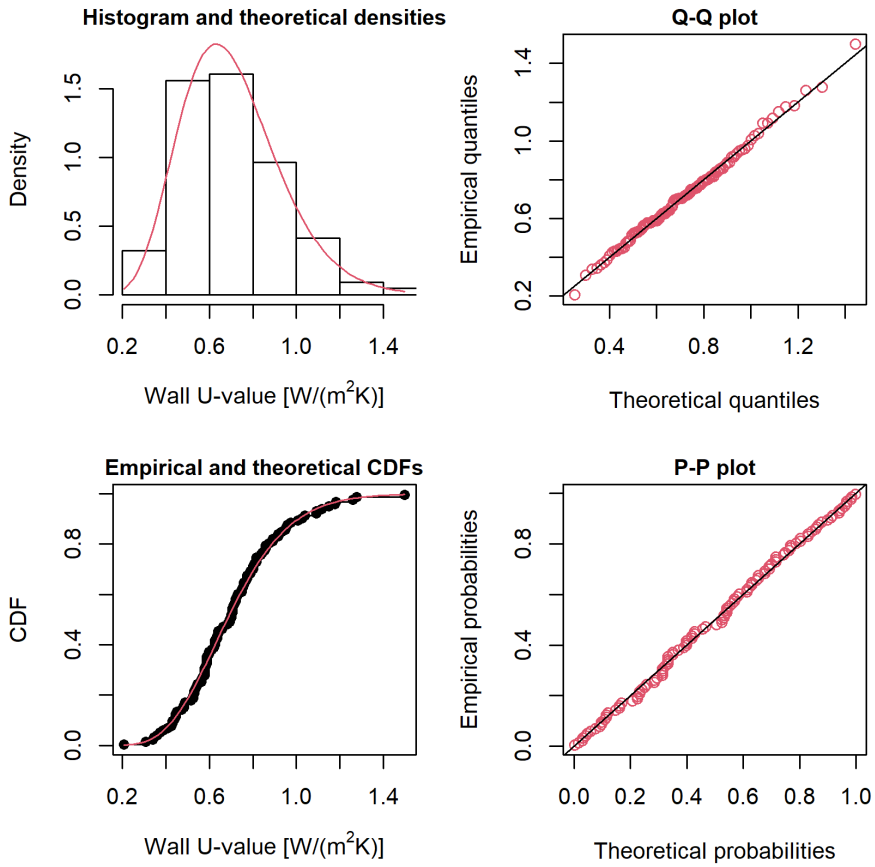


Figure 5.5: Goodness of fit plots for the BRE dataset of filled cavity wall U-values, assuming a gamma(9.5, 13).

Table 5.2. The gamma distribution has the lowest AIC value (-16.07), indicating that it can best represent the data amongst the candidate distributions. It is followed by the lognormal with an AIC of -13.72, the normal (AIC = -8.04) and Weibull (AIC = -7.59). To enable some further interpretation of the results, equations 5.4-5.5 were used to determine the AIC differences (Δ_j), and the Akaike weights (w_j). Based on the suggestions by Burnham and Anderson, 2004, with a $\Delta_j = 2.34$ there is some support for the lognormal as an alternative to the gamma, with considerably less support for the normal and Weibull distributions. This is further supported by the Akaike weights, with a 0.75 probability that the gamma distribution is the best distribution among the candidates given the observed wall U-values. A significantly lower probability of 0.23 is assigned to the lognormal while an almost negligible probability of 0.01 was assigned to the normal and Weibull distributions.

While the AIC-based statistics provided by Table 5.2 enable the modeller to choose the best fitting distribution amongst the candidates, the goodness of fit plots provided in Figure 5.5 are necessary to determine whether the best-fitting distribution is good enough for its intended use. The goodness of fit plots for the filled cavity wall construction are shown in Figure 5.5 where a gamma distribution was assumed. The empirical and theoretical densities and CDFs seem to align well while the points on the Q-Q and P-P plots align with the diagonal well. The diagonal is the line that indicates a perfect agreement between empirical and theoretical values. The Q-Q plots enables a closer inspection of the extremes. At the lower end, the point lies below the diagonal, suggesting the theoretical prediction is not as low as the empirical evidence, while at the upper end the theoretical value is not as high as the empirical. With the P-P plot, more attention is given to the body of the curve. There is small variation around the diagonal, yet no sizeable deviation is observed. Given these results, the gamma distribution is considered to describe the data adequately. Although small deviations were observed, especially in the Q-Q plot, this was at the extremes and differences were not large. The gamma distribution may therefore be considered a good approximation of empirical data of filled cavity wall U-values.

5.2.2 Window U-value

For the windows, no large-scale dataset with U-value measurements could be identified. Thus, the distribution of double-glazed window U-values had to be assumed. Based on the most recent RdSAP (BRE, 2019), the U-value of double glazing is expected to vary between 2.0–3.1 $\mathrm{W}/(\mathrm{m}^2\mathrm{K})$ as shown in Table 5.3.

Table 5.3: Assumed window U-value and probability distributions. The left-hand side is based on Table S14 in BRE (2019). The right-hand side lists the model input distributions assumed for window U-value in this work.

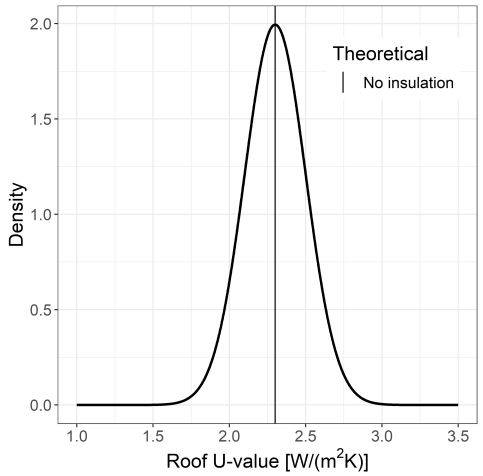
Glazing	Installed	Glazing gap	U-value ($\mathrm{W}/(\mathrm{m}^2\mathrm{K})$)	Distribution
Double	Pre-2002	6 mm in PVC frame, or any in non-PVC frame	3.1	normal(2.5, 0.3)
		12 mm in PVC frame	2.8	
		16 mm or more in PVC frame	2.6	
Double	2002 or later	Any	2.0	

The distribution will depend on the frequency of different double glazing types within the stock, along with how much the U-value of each double glazing type can vary. Information on the prevalence of each double glazing type, or the variation of U-value per window type, was not available. While building construction age can be a proxy of the double glazing type, it is likely an unreliable one, especially for older dwellings that may have been refurbished. Since the only information available were the theoretical U-values (Table 5.3), a normal distribution with a mean of $2.5\,\mathrm{W}/(\mathrm{m}^2\mathrm{K})$ was assumed and a standard deviation of $0.3\,\mathrm{W}/(\mathrm{m}^2\mathrm{K})$. The assumption of a normal distribution suggests an equal probability of being above or below the mean value. This is likely a simplifying assumption, and the true empirical distribution may be multimodal. However, the lack of information on the prevalence of each window type prevented a more detailed specification.

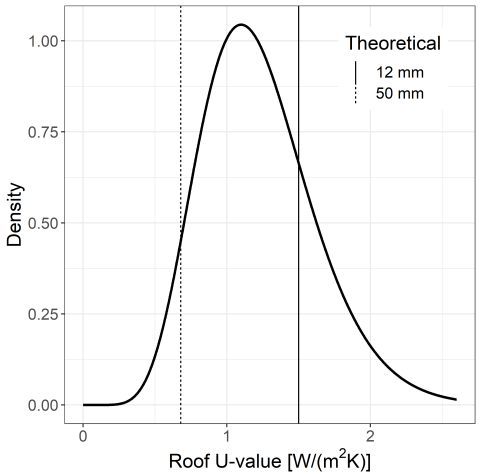
5.2.3 Roof U-value

The roof U-value for a single house will depend on the roof type, the extent and quality of roof insulation. Similar to windows, no large-scale resource of measured roof U-values could be identified. Therefore, the distributions and distributional parameters were once again informed by RdSAP (BRE, 2019), with a summary of roof U-values provided by Table 5.4. A question within the 4M household survey asked occupants to state which of the following five categories describes their loft insulation level: (i) No insulation, (ii) Up to 50 mm, (iii) 50–100 mm, (iv) 100–200 mm and (v) greater than 200 mm. For the purposes of archetype-based stochastic modelling, these categories may be used to cluster dwellings into separate groups during the Categorical Variable Classification step and assign each group with a distinct probability distribution. However, if the key factor that varies with each category is the roof U-value then it might be possible to merge some of these categories into larger groups and represent them with a distribution weighted based on each group's frequency. If roof U-value has no significant impact on the QoI, then all the groups can be merged together and the central value can be used.

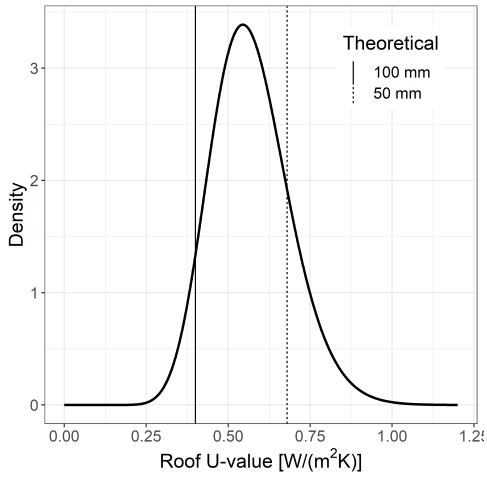
For each category where roof insulation was present, most probability density was within the U-values associated with the insulation thickness. For example, in



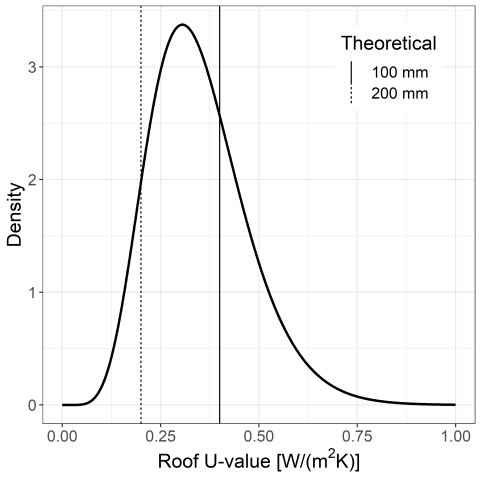
(a) No insulation



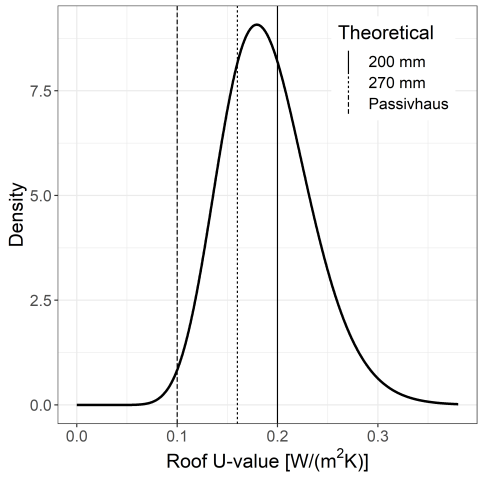
(b) Less than 50 mm



(c) 50–100 mm



(d) 100–200 mm



(e) 200 mm or more

Insulation Level	Distribution
No Insulation	normal(2.3, 0.2)
Less than 50 mm	gamma(9.5, 7.7)
50-100 mm	gamma(22.6, 39.6)
100-200 mm	gamma(7.8, 22.4)
200 mm or more	gamma(17.8, 93.8)

(f) Assumed distributions

Figure 5.6: Probability density functions of roof U-value assumed for each group of loft insulation thickness provided in the 4M dataset. The theoretical values were informed by RdSAP (BRE, 2019).

Table 5.4: Assumed insulation level and U-values based on dwellings age band in England and Wales. Adapted from Table S10 in BRE (2019).

Age Band	Assumed Insulation	Pitched, slates or tiles, insulation between joists or unknown ($W/(m^2K)$)
Before 1900, 1900-1929, 1930-1949, 1950-1966	None	2.3
1967-1975	12 mm	1.5
1976-1982	50 mm	0.68
1983-1990	100 mm	0.40
1991-1995	150 mm	0.30
1996-2002	150 mm	0.26
2003-2006	270 mm	0.16
2007-2011	270 mm	0.16
2012 onwards	270 mm	0.16

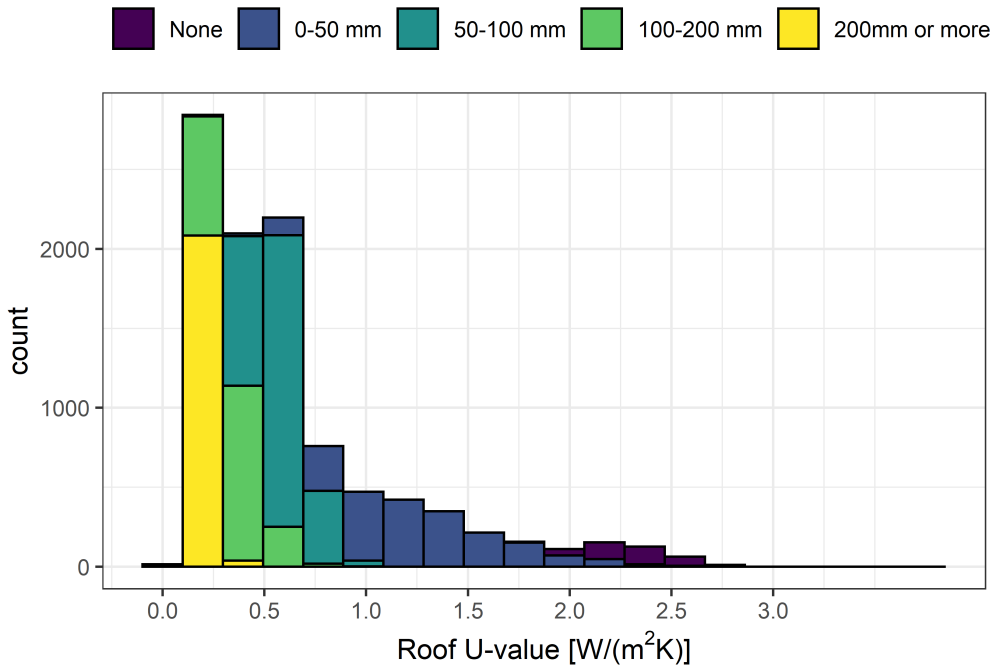


Figure 5.7: Weighted distribution of roof U-values based on the prevalence of loft insulation levels within the group of semi-detached dwellings with filled cavity wall in the 4M dataset.

the group with loft insulation thickness ranging between 50–100 mm, most of the probability density will be within the range of 0.40–0.68 $W/(m^2K)$. A peak was expected to exist within this range, representing the average loft insulation thickness. The probability density would decrease in both directions from the peak. A non-zero probability will exist beyond the reference values, especially in the case of a U-value

greater than 0.40 $W/(m^2K)$; as with filled cavity walls, factors such workmanship and performance degradation over time will likely result in U-values greater than those quoted in Table 5.4. Therefore, a positively-skewed probability distribution may capture the possibility of a subgroup of dwellings performing significantly worse than expected. One distribution that satisfies the requirements is the gamma, which was also the probability distribution that provided the best-described U-values of the filled cavity wall construction in Section 5.2.1. For roofs without any loft insulation, the key feature assumed to characterise their distribution is the existence of a central value. Workmanship might have some effect, and some variability in roof construction will likely exist, yet the impact on U-value is not expected to be as significant as for the insulated lofts. A normal distribution was assumed, suggesting that values above the mean are equally likely as below and that most of the probability density is around the mean.

Following this reasoning and using the RdSAP values in Table 5.4, a set of probability density functions were defined, visualised in Figure 5.6. For the "No Insulation" group, a normal distribution was assumed with a mean = 2.3 and standard deviation = 0.2. For the "Less than 50 mm", a gamma(9.5, 7.7) distribution was assumed. This distribution places most of the probability density (0.71) between a U-value of 0.68 $W/(m^2K)$ and 1.50 $W/(m^2K)$, and only allows for a small probability (0.01) of performing similarly to having no insulation at all. Similarly, the sub-groups 50–100 mm and 100–200 mm had most of their probability density within the range of associated reference values and with a long right tail. For the group representing dwellings with loft thickness greater or equal to 200 mm, a U-value of 0.1 $W/(m^2K)$ in Figure 5.6(e) was used to represent a typical Passivhaus standard roof U-value (McLeod et al., 2013). It is possible to have a U-value greater than this, but it would be quite uncommon, hence the distribution assumed allows a probability of 0.01 for U-values less than 0.1 $W/(m^2K)$. In all four cases where insulation was present, the upper reference value was more probable than the lower reference value, capturing the assumption of worse than ideal performance.

To determine how the U-value varies within the group of semi-detached dwell-

ings identified within Section 4.2.6.2, a weighted distribution was constructed which is visualised in Figure 5.7. To construct this distribution, a large number (10,000) of samples were drawn from the five distribution (Figure 5.6(f)) in proportion with the prevalence of each insulation type within the chosen cluster of dwellings. This resulted in a multimodal distribution that may need to be segmented further (Step 4 of the framework described in Section 3.2) if this model input is shown to be significant in the sensitivity analysis.

5.2.4 Floor U-value

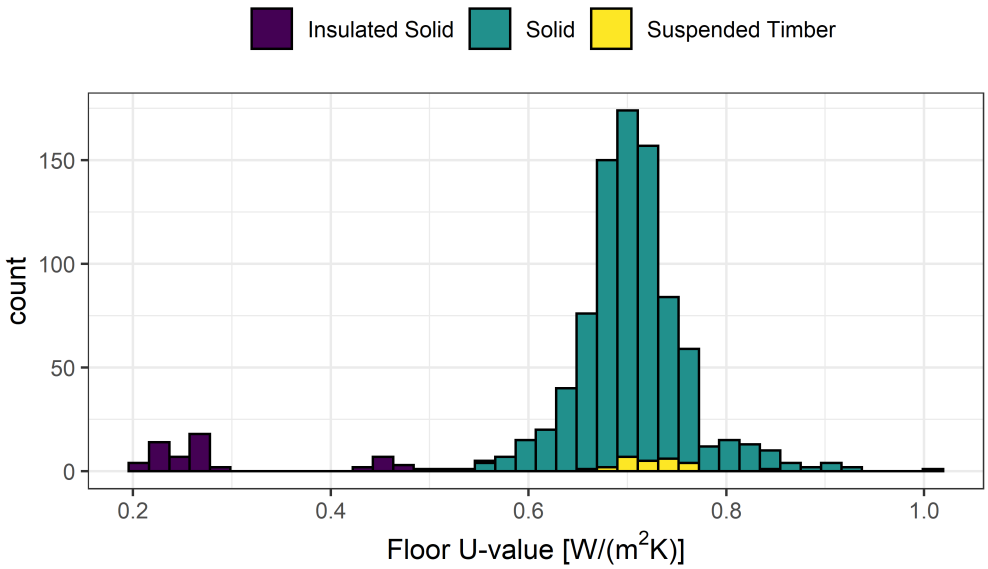


Figure 5.8: Floor U-values estimated using the RdSAP S5.5 guidance for semi-detached dwellings in the 2012 English Housing Survey.

The floor U-value will predominantly depend on the floor construction type and the presence of floor insulation. A strong association is expected between floor construction type and construction date, while the presence of floor insulation will likely be influenced by construction and refurbishment age. Since empirical measurements of floor U-value could not be identified, RdSAP was used to estimate plausible values. RdSAP (Section S5.5) provides an analytical method based on BS EN ISO 13370 to estimate floor U-values depending on the type of floor construction and the presence of insulation. In addition, it requires a few parameters to be known (although many default values are provided) including the dwelling's floor area and

exposed perimeter. Neither the EHS nor the 4M had information on the type of floor construction. Based on Table S11 from RdSAP (BRE, 2019), it was assumed that dwellings constructed until 1929 had suspended timber floors, dwellings built after 1929 had solid floors, with insulation present for dwellings constructed after 1995. While 4M included estimates of floor area, it did not include any information on the exposed perimeter. The EHS was thus preferred to inform this distribution. Other parameters required for this calculation, such as the soil conductivity, were taken to be the defaults provided by RdSAP.

A histogram of the estimated floor U-values per floor type according to the RdSAP methodology is shown in Figure 5.8. Solid floor is the most common floor type. The distributions of U-values for uninsulated solid floors (median = 0.70 $\mathrm{W}/(\mathrm{m}^2\mathrm{K})$) and suspended timber (median = 0.72 $\mathrm{W}/(\mathrm{m}^2\mathrm{K})$) floors have similar central values. In contrast, the median U-value for insulated solid floors is 0.26 $\mathrm{W}/(\mathrm{m}^2\mathrm{K})$. Since this distribution appears to be multimodal, and a unimodal distribution would not describe the data well, a theoretical distribution was not fitted to the dataset. Whether a distribution should be fitted for a subset of this dataset will be determined at Step 4 of the Bayesian calibration framework, described in Chapter 6.

5.2.5 Fabric Air Permeability

Approved Document L1 (ADL1) describes air permeability as the physical property used to measure the building fabric's airtightness and defines it as the "air leakage rate per hour per square metre of envelope area at the test reference pressure differential of 50 pascals" (HMG, 2016). A typical method of quantifying this is through pressure testing, also referred to as a "blower-door test" or "fan pressurisation test"; a fan fitted in the doorway of a house supplies or extracts air at steady pressure differences with the aim of evaluating the dwelling's air leakage characteristics (Stephen, 2000). During the pressurisation test, internal doors are kept open while windows are closed, chimneys, flues and all purpose-provided ventilation inlets are sealed, and mechanical ventilation is switched off. Air permeability is a model input of UK-HSM and a distribution must therefore be defined based on the best available evidence.

The largest and most comprehensive dataset of airtightness measurements in English dwellings constructed before 1995 is held by BRE (Perera and Parkins (1992), Stephen (1998) and Stephen (2000)). It consists of measurements of air leakage rate for 471 dwellings, of which 384 had enough information to estimate their air permeability. The mean air leakage rate was 13.1 ACH at 50 Pa, with an approximate range of 2–30 ACH. This large variability, with the greatest leakage rate being an order of magnitude larger than the smallest measurement, led Stephen (1998) to conclude that it was "impossible to make a realistic estimate of airtightness of a dwelling, newly built or otherwise, by simple inspection alone; some form of measurement being required". To explore the factors influencing airtightness, Stephen (2000) compared the mean leakage rate (ACH at 50 Pa) of different subgroups of the dataset. For brevity, this section will focus on construction age. Other key findings and limitations of the Stephen (2000) analysis are discussed in Section H.3 of the appendices.

Construction age can be a useful proxy of typical construction practices, materials used and the designed air permeability that may be dictated by building regulations and standards. With the caveat that chimneys were closed during the pressurisation tests, Stephen (2000) discovered that the mean leakage rate of dwellings pre-1920 was lower than for dwellings constructed between 1920–1980; this was contrary to a belief at the time that older dwellings are more draughty and less airtight. Dwellings built after the 1980s were more airtight with a mean air leakage rate of approximately 10 ACH at 50 Pa, although a large variability was still observed.

An overall trend of improved airtightness for post-1995 homes might be expected, since building regulations have been striving to reduce infiltration for a few years via approved document L (ADL). In 1995, ADL1 stated the aim of limiting leakage in dwellings and provided a list of measures to achieve this by reducing unintentional air paths HMSO (1995). The guidance in 2002 maintained the same aim and supplemented the list of measures that may be taken to reduce air leakage with an alternative method of demonstrating compliance; the use of pressure testing

Table 5.5: Datasets used to inform the model. input for air permeability

Dwelling age	Sample size	Dataset
Pre-1995	384	Stephen (2000)
2002–2006	63	BRE (2004)
Post-2006	110	Pan (2010)

to show that the dwelling's air permeability does not exceed $10\,m^3/h/m^2$ at 50 Pa (HMG, 2002). In the 2006 version of ADL1, new dwellings had to be designed with a permeability of $10\,m^3/h/m^2$ at 50 Pa and pressure testing had to be carried out to demonstrate its compliance (although not every dwelling has to be pressure tested in a development of identical dwellings) (HMG, 2006). The target of $10\,m^3/h/m^2$ at 50 Pa remained in ADL1 2016 (HMG, 2016), with pressure testing continuing to be used.

Given the association between construction age and air permeability, three datasets that cover different construction periods were used to inform this model input, summarised in Table 5.5. A few points are worth discussing regarding this choice. Firstly, although on average dwellings with different floor and wall type are expected to have different air permeabilities, they are not treated differently in this analysis. Although Stephen (2000) provided measured means of solid and cavity walls, and solid and suspended timber floors, there was no information about their distributional form. In addition, there was no discussion about the presence of insulation, which will likely have an impact on the building's air tightness. Furthermore, this dataset likely paints an outdated picture of the housing stock, since a proportion of dwellings will likely have received some retrofit measures that have changed their air leakage characteristics. However, this dataset is the most comprehensive dataset of blower-door measurements for pre-1995 dwellings and thus provides the best indication of air permeability for this group of dwellings. Another point to note is that no suitable dataset could be identified for dwellings built between 1995–2002. These were assumed to be represented by the BRE dataset of dwellings constructed between 2002–2006, since similar instructions were given to reduce air leakage in the 1995 and 2002 ADL1 version (HMSO, 1995; HMG, 2002). Although the 2002 version of ADL1 explicitly mentioned the target of $10\,m^3/h/m^2$

at 50 Pa while the 1995 did not, dwellings constructed between 1980–1994 and 2002–2004 had mean permeabilities close to 10 $m^3/h/m^2$ at 50 Pa. It is therefore assumed that similar values would be observed for dwellings constructed between 1995–2002.

5.2.5.1 Pre-1995

Table 5.6: Distributions for each construction period ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j). P1 and P2 represent the parameters of the fitted distribution, stated to two significant figures.

Const. Period	Distributions	AIC	Δ_j	w_j	P1	P2
Pre-1995	weibull	2298	0	0.98	shape = 2.5	scale = 13
	gamma	2305	8	0.02	shape = 4.8	rate = 0.42
	norm	2320	22	0.00	mean = 11	sd = 4.9
	lnorm	2334	37	0.00	meanlog = 2.3	sdlog = 0.49
1995–2006	lnorm	309	0	0.53	meanlog = 2.3	sdlog = 0.28
	gamma	310	1	0.41	shape = 13	rate = 1.3
	norm	314	5	0.04	mean = 10	sd = 2.8
	weibull	316	6	0.02	shape = 3.8	scale = 11

The permeability measurements of pre-1995 dwellings included in the BRE dataset are shown in Figure 5.9 (and in Figure H.4 of the appendices). These include cavity and solid wall constructions, along with solid and suspended floor types. By inspecting the histogram, the only extreme values are on the right tail with a cluster of dwellings with permeabilities ranging from 26–29 $m^3/h/m^2$. Although these values are approximately 2.5 times greater than the mean, the rest of the distributions seems to have a longer right tail which this cluster could be part of. Indeed, looking at the leakage rates (ACH at 50 Pa, n = 471) in Figure 2 of Stephen (2000) which the data in Figure 5.9 are a sub-sample of, there is a long right tail. The entire dataset was thus used to fit a set of candidate distributions, with the AIC and its derivatives summarised in Table 5.6. The best fitting distribution was Weibull, with shape = 2.5 and scale = 13. The goodness of fit plots in Figure 5.9 suggest that Weibull provides a satisfactory fit, as it describes the vast majority of data points well.

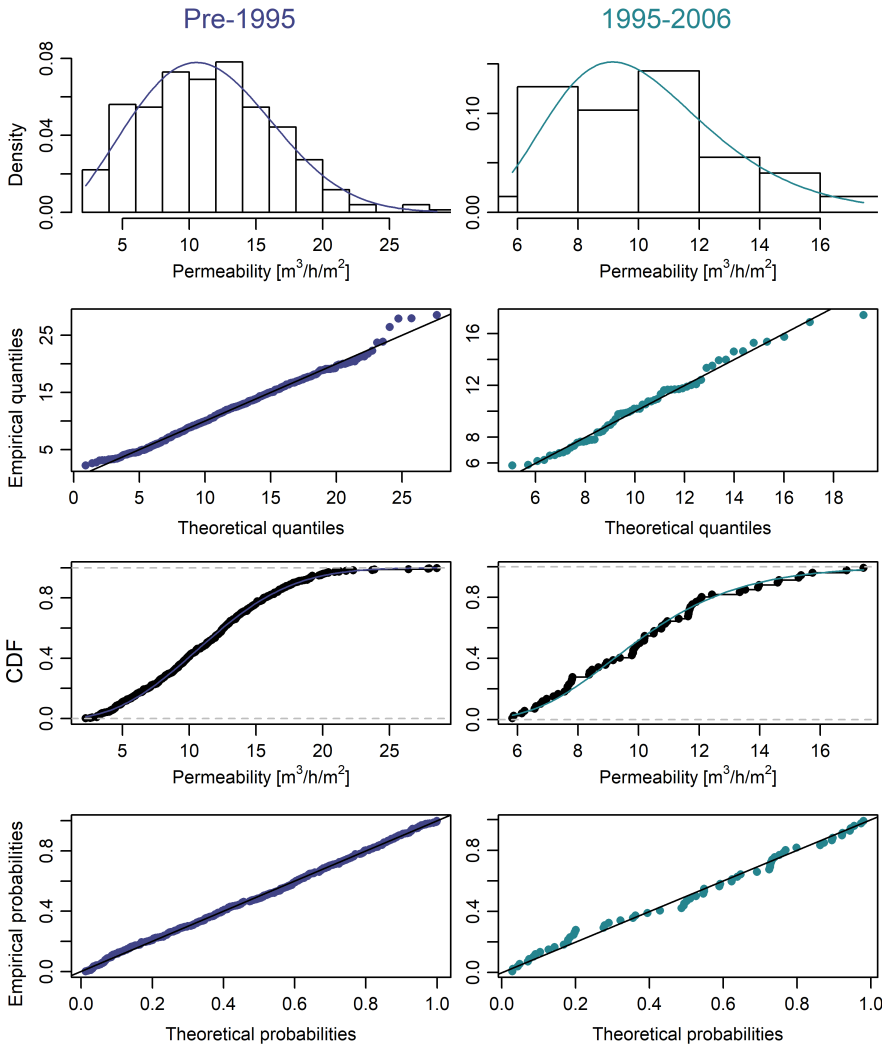


Figure 5.9: Goodness of fit plots for air permeability. The Pre-1995 dataset (Stephen, 2000) was fitted with a weibull(2.5, 13) while the 1995–2006 dataset (BRE, 2004) was fitted with a lognormal(2.3, 0.28).

5.2.5.2 1995–2006

The Energy Saving Trust, motivated by an earlier and smaller study that showed 2 in 3 dwellings failed to achieve air permeabilities of $10\,\mathrm{m^3/h/m^2}$, contracted BRE and the National Energy Services (NES) to examine the extent and effect of non-compliance with air permeability goals in dwellings constructed to the 2002 edition of ADL1 (BRE, 2004). Pressure testing was conducted in 99 dwellings, with the measurements presented as histograms for the entire sample, and separately for houses (66) and flats (36). Only the sub-sample of houses was used, as this was

deemed to be more representative of the semi-detached dwellings modelled in this study. No extreme values were observed (Figure H.5 in the appendices) and a set of candidate distributions were fitted. The best fitting distribution according to Table 5.6 was a lognormal with a meanlog = 2.3 and a sdlog = 0.28. This distribution was shown to describe the dataset well in the goodness of fit plots in Figure 5.9.

5.2.5.3 Post-2006

Pan (2010) analysed the air permeability tests of 287 post-2006 dwellings, comprised of 110 houses and 177 flats. A comparison of typologies revealed a statistically significant difference in the air permeability of houses and flats (lower and mid-ground), but no significant differences for different housing typologies.

From this study, it was not possible to extract tabular data but distributional parameters were provided by the author. The overall dataset was determined to be described well by a normal distribution with a mean of $5.97\,\mathrm{m^3/h/m^2}$, with a standard deviation of $2.29\,\mathrm{m^3/h/m^2}$ (Pan, 2010). While it would be possible to use the distributional information provided by Pan (2010) for the entire dataset, this would bias the model input towards the more prevalent and airtight flats' subgroup. Since the focus in this case is semi-detached dwellings, it was preferred to use only the houses' subgroup information. However, the appropriate distribution and its parameters were not provided for subgroups, nor were any histograms. Only bar plots with a mean and 95 % confidence interval of the air permeability were presented (Figures 2–5 in Pan (2010)). However, it was possible to estimate each subgroup's standard deviation given its mean and 95 % confidence interval by a normal distribution, and using the following equations:

$$(\bar{x} - 95LB) \times \sqrt{n}/(1.96) = \sigma_{LB}, \tag{5.6}$$

$$(\bar{x} + 95UB) \times \sqrt{n}/(1.96) = \sigma_{UB}, \tag{5.7}$$

$$\sigma = (\sigma_{LB} + \sigma_{UB})/2, \tag{5.8}$$

where σ_{UB} and σ_{LB} are the standard deviations estimated using the upper ($95UB$) and lower ($95LB$) bounds of the 95 % confidence interval, respectively. n is the number

of dwellings within each sample and σ is the mean standard deviation. Note that σ_{UB} and σ_{LB} may slightly differ due to rounding errors of the confidence intervals provided by Pan (2010). Using this method, the distributions of houses and flats are normal(7.14, 2.01) and normal(5.25, 2.21), respectively. Although a subgroup was provided for semi-detached dwellings, this was not used due to its small sample size of 19 and since there was no statistically significant difference between different typologies within the house subgroup.

A possible criticism of this method is the assumption that each subgroup is described by a normal distribution. To demonstrate that this is indeed a reasonable assumption, the following procedure was being used: by assuming that the house and flat subgroups are described by normal distributions, the entire dataset should also be described by a mixture distribution whose mean and variance can be estimated analytically given each subgroup's distributional parameters and prevalence as detailed by Behboodian (1970). Following this method, the estimated distributional parameters for the entire dataset are normal(5.97, 2.32) which are similar to the normal(5.97, 2.29) distribution identified by Pan (2010).

5.2.5.4 Weighted Distribution

The cluster of semi-detached homes being assessed includes dwellings constructed in all three constructions periods that permeability distributions were identified for. Depending on their prevalence, 10,000 samples were drawn from the three distributions to construct a weighted distribution that would be representative of the cluster's overall permeability levels. This is visualised in Figure 5.10. The weighted distribution is a positively-skewed unimodal, dominated by the Pre-1995 samples since they were more prevalent than the 2002-2006 and Post-2006 categories.

As before, a set of distributions were fitted with a summary of the AIC metrics, summarised in Table 5.7. A Weibull with shape = 2.6 and scale = 13 was identified to best describe the weighted distribution, with its fit supported by the goodness-of-fit plots in (Appendix Figure H.6).

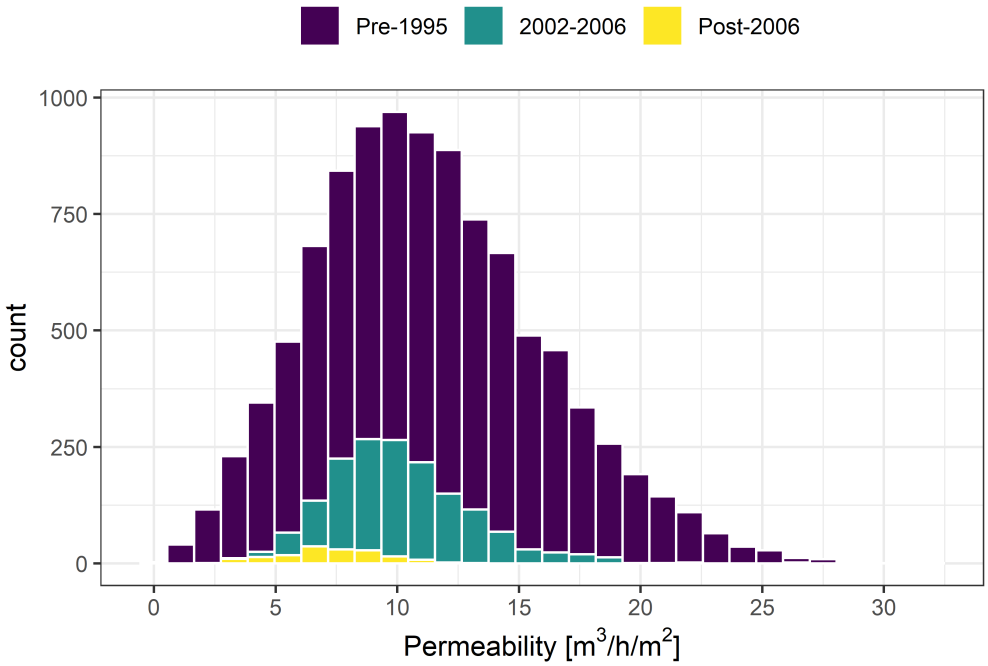


Figure 5.10: Weighted distribution of air permeability comprised of 10,000 samples drawn from the three distributions previously identified, depending on the prevalence of dwellings with different construction periods within the cluster.

Table 5.7: Distributions fitted to the weighted air permeability of a cluster of semi-detached dwellings. They are ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j).

Distributions	AIC	Δ_j	w_j	P1	P2
weibull	58847	0	1.00	shape = 2.6	scale = 13
gamma	59114	266	0.00	shape = 5.2	rate = 0.46
norm	59253	406	0.00	mean = 11	sd = 4.7
lnorm	60235	1387	0.00	meanlog = 2.3	sdlog = 0.48

5.2.6 Solar Absorptivity

Solar absorptivity of roof and external walls will depend on the material used for the building envelope, its colour and the presence of any reflective coating. Without any tabulated or empirical data on the prevalence of solar absorptivity values within the Leicester housing stock, a distribution had to be assumed. Since this material property is bound between 0 and 1, a beta distribution was deemed an appropriate choice since it is a flexible model used to represent probability over a fixed range, often 0–1 (Mun, 2012).

Table 5.8: Absorptivity and emissivity of frequently used construction material. Adapted from CIBSE (2015).

Material	Condition/Type	Absorptivity	Emissivity
Brick	Glazed/light	0.25–0.36	0.85–0.95
	Light	0.36–0.62	0.85–0.95
	Dark	0.63–0.89	0.85–0.95
Cement Mortar, screed	-	0.73	0.93
Clay tiles	Red, brown	0.60-0.69	0.85-0.95
	Purple/dark	0.81-0.82	0.85-0.95
Concrete	Tile	0.65–0.80	0.85–0.95
	Block	0.56–0.69	0.94

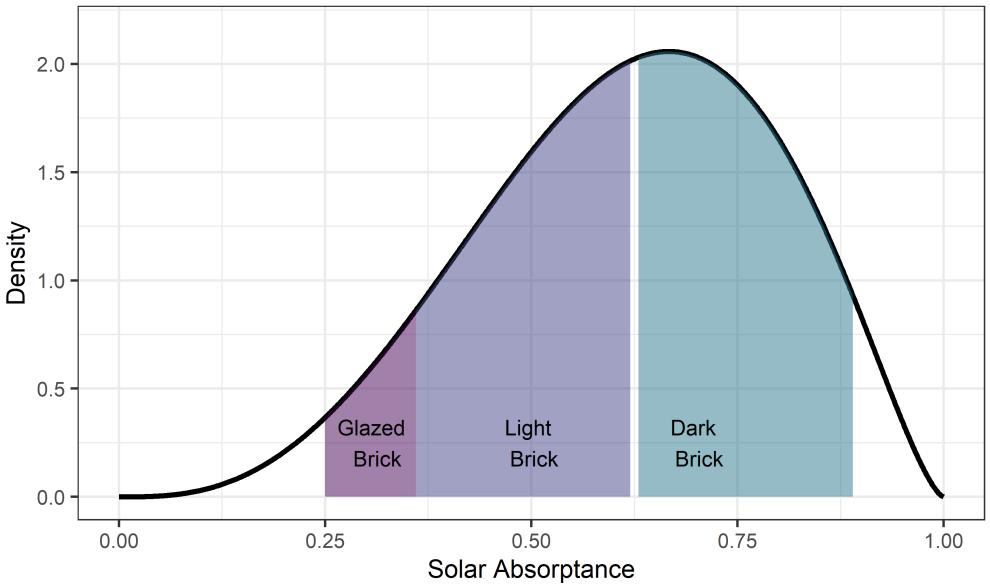


Figure 5.11: Density plot of the chosen probability density function for solar absorptance. The shaded areas mark the probability regions assigned to theoretical absorptivity values of different brick types.

The most common wall construction is masonry-based, while typical roof constructions of semi-detached dwellings have used concrete tiles (NHBC Foundation, 2019). The solar absorptivity of these materials can thus inform distributional parameters. Ranges of absorptivity and emissivity, adapted from CIBSE (2015), are shown in Table 5.8. The range of absorptivity is between 0.25–0.89, although a value less than 0.36 only refers to glazed/light brick. It is assumed that light and dark bricks have been more commonly used in the UK than glazed bricks, therefore a greater probability will be assigned in the interval 0.36–0.89. A smaller probability density

is expected between 0.25–0.36, and even smaller for below 0.25 and above 0.89. The absorptivity of concrete tiles, ranging between 0.65–0.80, providing further support for a high probability region in the range of 0.5–0.8. Other factors, such as dirt, might also influence the absorptivity, but they are not considered in this analysis.

A distribution that satisfies the above assumptions is a beta with shape 1 = 4 and shape 2 = 2.5 (beta(4, 2.5)). A visualisation of this probability density is shown in Figure 5.11. With this choice of distributional parameters, the probability of solar absorptance being below 0.26 is 0.03, while a value above 0.89 is assigned a probability less than 0.05. The probability assigned to glazed, light and dark bricks were 0.07, 0.40 and 0.45, respectively.

5.2.7 Glazing Fraction

Glazing fraction is a UK-HSM model input that controls the ratio of glazed area to external wall area. It was not possible to inform this model input based on the 4M dataset. However, there was enough information within the EHS to obtain a set of empirically-derived probability distributions based on the glazed and wall area of the front and back façade. Histograms of glazing fraction, separated by wall type (solid, cavity, filled cavity), can be found in the appendices (Figure H.7). The median value was similar between the three wall construction types, ranging from 0.26 to 0.27. In all three cases the distributions are positively skewed with a long right tail and no clear extremes.

A set of probability distributions were fitted to each group of dwellings, with the summary of AIC-based measures of fit for the filled cavity wall group summarised in Table 5.9. For all three wall types, the distribution that best describes the glazing fraction is a gamma, although the shape and rate differ. The goodness of fit plots (Appendix Figure H.8) suggest that the gamma distribution provides a satisfactory fit.

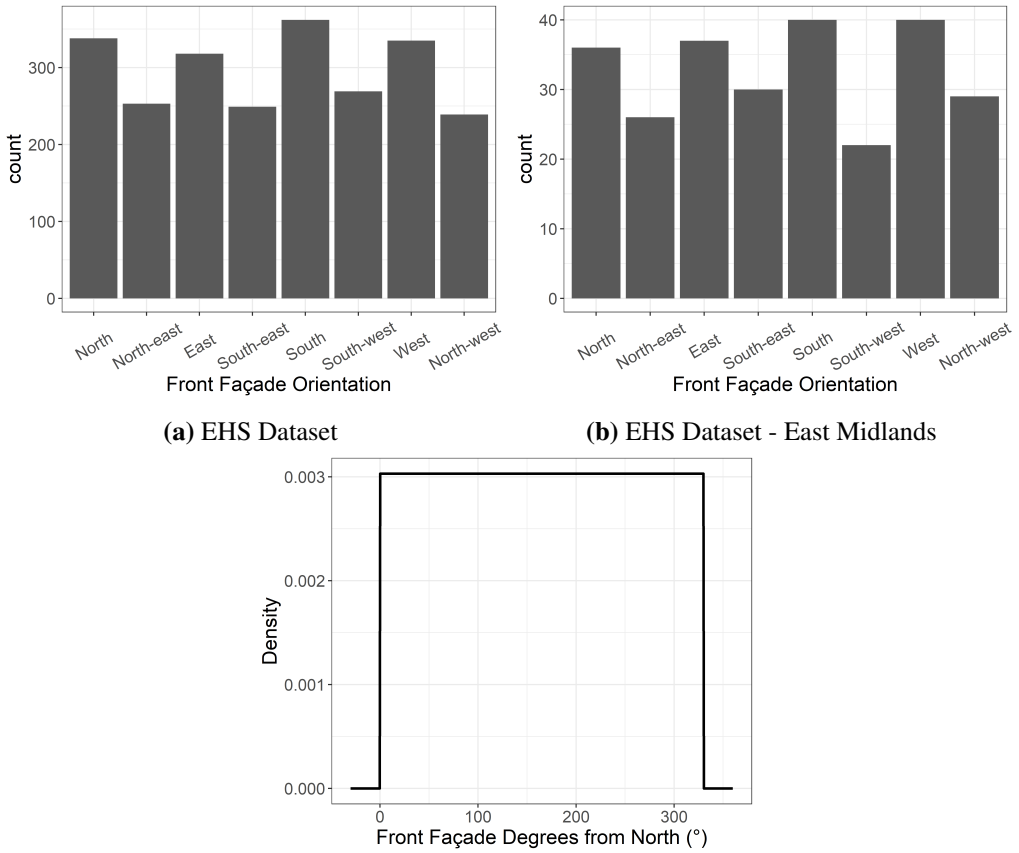
5.2.8 Orientation

The orientation of a dwelling can have a significant impact on the indoor environment. Since the 4M dataset did not include any information on the orientation of the

Table 5.9: Distributions fitted to the glazing fraction of semi-detached dwellings in the English Housing Survey, with filled cavity wall (FCW) construction. They are ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j).

Wall	Dist.	AIC	Δ_j	w_j	P1	P2
FCW	gamma	-3401	0	1.00	shape = 14	rate = 53
	lnorm	-3388	13	0.00	meanlog = -1.4	sdlog = 0.27
	norm	-3303	98	0.00	mean = 0.26	sd = 0.072
	weibull	-3220	181	0.00	shape = 3.7	scale = 0.29

buildings monitored, this model input was informed by the EHS. Only dwellings that satisfied the classifiers defined in Section 4.2.6.2 were used.



(a) EHS Dataset

(b) EHS Dataset - East Midlands

(c) Proposed Probability Distribution

Figure 5.12: Barplots (a) and (b) visualise the prevalence of different orientation within the EHS 2012 dataset at the national and East Midlands level, respectively. The density plot in (c) shows the proposed probability distribution used for the orientation model input of UK-HSM.

Figures 5.12(a)–5.12(b), show the prevalence of front façade orientations of semi-detached dwellings at the national and East Midlands level. In both cases, there

is a more frequent occurrence of the four cardinal directions than the intercardinal directions. It is unclear whether this is indeed the result of common building practices or an indication that surveyors are more likely to assign a cardinal than an intercardinal direction. No single peak exists and plotted over a continuous range, it is expected that the frequency of angles will be quite small. This relationship is likely approximated to a satisfactory degree by a uniform probability distribution, shown in Figure 5.12(c). This was chosen to range between 0-330°, since simulations at an angle greater than 330° would likely have a similar effect to that of 0°.

5.2.9 Floor-to-ceiling Height

The 4M dataset did not include measurements of the floor-to-ceiling height of each dwelling. To inform this model input the EHS was used, which provides measurements within the living room and main bedroom. These two measurements were averaged since UK-HSM assumes the two rooms to have the same height and specified by a single input. EnergyPlus, UK-HSM's calculation engine, assumes surfaces to be infinitely thin plates and thus the use of interior measurements of room dimensions results in smaller wall surface area than in the real building (DOE, 2020). While the differences might be small, EnergyPlus guidance suggests the use of outside dimensions for exterior surfaces and the midpoint for interior surfaces (DOE, 2020). To implement this advice, all floors above ground level were assumed to be 0.25 m thick which was modelled by adding 0.125 m to the averaged measurements of bedroom and living room height from the EHS.

From Figure 5.13, the floor-to-ceiling height of semi-detached dwellings within EHS is mostly between 2–3 m, although there is a long right tail. Only two dwellings had a value greater than 3.5 m, which is less than 0.04 % of the dataset. These were removed as they might have been the result of surveyor error or simply extreme values. The density plot appears multimodal; this is the result of the measurement resolution at 0.05 m.

To accommodate the data's long right tails and in addition to the four probability distribution fitted for other model inputs (normal, lognormal, Weibull and gamma), the inverse Weibull distribution was also fitted to the data. Inverse Weibull belongs

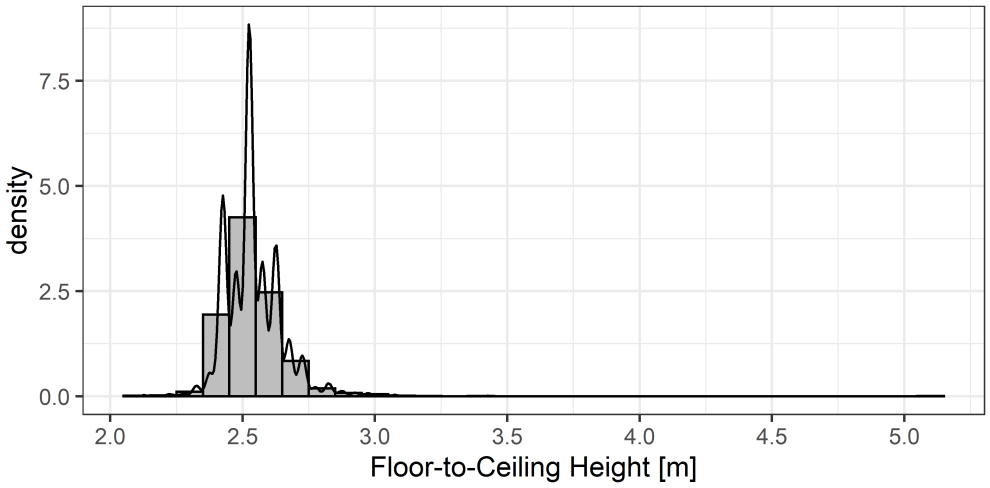


Figure 5.13: Histogram of the floor-to-ceiling height measurements for semi-detached dwellings in the EHS. This is the average of the main bedroom and living room measurement plus 0.125 m.

Table 5.10: Distributions fitted to floor-to-ceiling height measurements from 2012 English Housing Survey. The measurements were augmented with 0.125 m. They are ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j).

Wall	Distributions	AIC	Δ_j	w_j	P1	P2
	lnorm	-2777	0	0.96	meanlog = 0.93	sdlog = 0.034
	gamma	-2771	7	0.04	shape = 840	rate = 330
FCW	norm	-2755	22	0.00	mean = 2.5	sd = 0.087
	invweibull	-2525	252	0.00	shape = 27	scale = 2.5
	weibull	-2251	526	0.00	shape = 24	scale = 2.6

to the family of generalised extreme value distributions and is able to capture long tails well (de Gusmão et al., 2011).

Table 5.10 provides a summary of the AIC and its derivatives for every distribution fitted. The best fitting distribution for semi-detached dwellings with filled cavity walls is a lognormal, with a meanlog = 0.93 and a sdlog = 0.034. Based on the GOF plots in Figure 5.14, the theoretical model deviates from the empirical data for values greater than 2.7 m and less than 2.3 m. Since the range of 2.3–2.7 m includes 98.7 % of the data, it was considered an adequate description. The gaps between points on the P-P plot are the result of the measurement resolution.

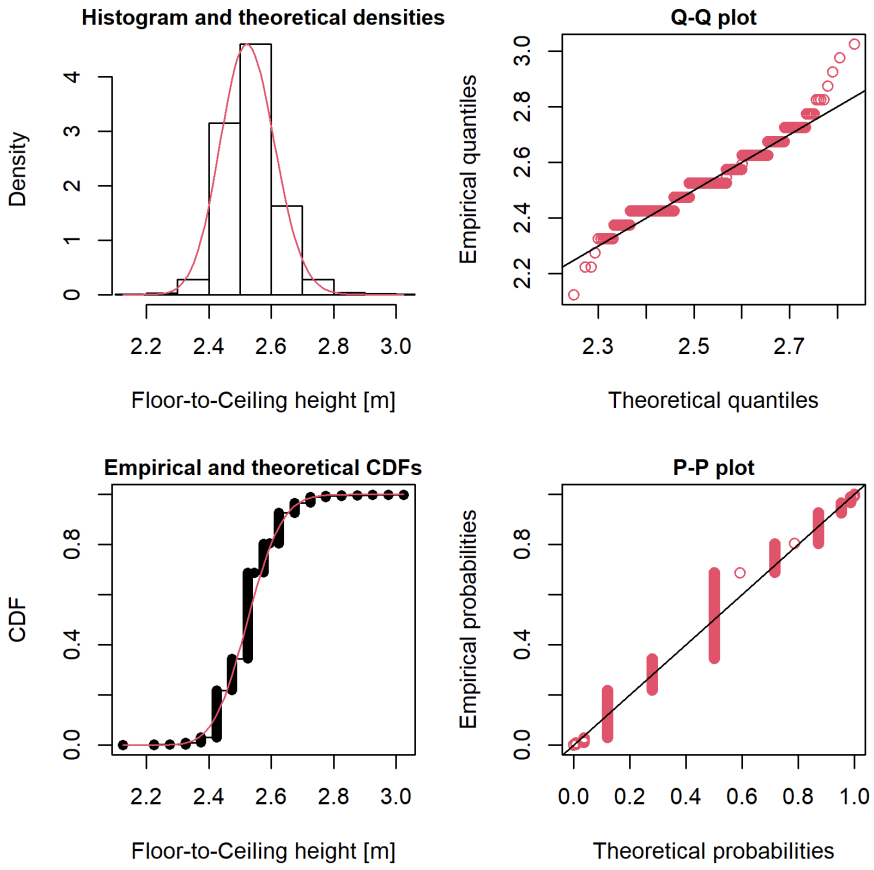


Figure 5.14: Goodness of fit plots for a lognormal distribution fitted to the floor-to-ceiling height measurements from the 2012 English Housing Survey. The measurements were augmented with 0.125 m.

Table 5.11: Distributions fitted to the floor area factor dataset. These are ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j). The corrected AIC was used for both groups. P1 and P2 represent the parameters of the fitted distribution, stated to two significant figures.

Wall	Dist.	AIC	Δ_j	w_j	P1	P2
	invweibull	-17.00	0.00	0.94	shape = 5.5	scale = 0.74
	lnorm	-11.00	6.00	0.05	meanlog = -0.2	sdlog = 0.23
FCW	gamma	-8.00	9.00	0.01	shape = 17	rate = 20
	norm	0.00	17.00	0.00	mean = 0.84	sd = 0.23
	weibull	3.00	20.00	0.00	shape = 3.5	scale = 0.93

5.2.10 Floor Area Factor

The floor area in UK-HSM is set by the floor area factor. For a semi-detached model, a floor area factor of 1 is equivalent to a ground floor area of 51.6 m^2. By

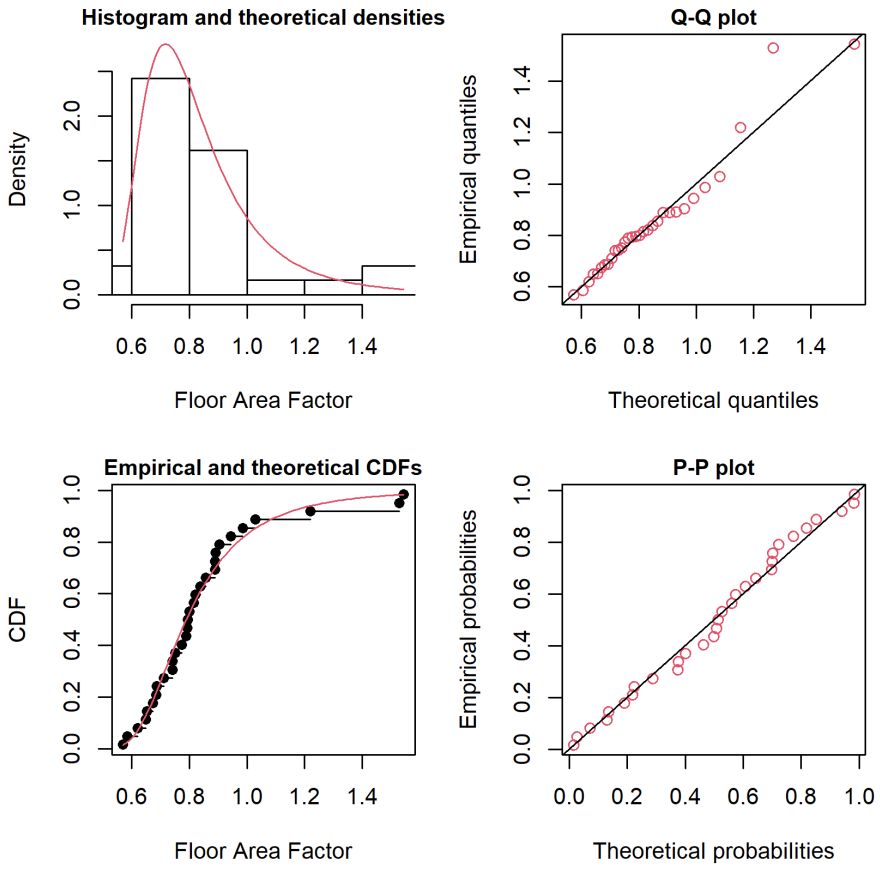


Figure 5.15: Goodness of fit plots of floor area factor, estimated using the semi-detached homes with filled cavity wall in 4M. An inverse Weibull distribution was assumed.

varying this model input, the floor area varies linearly. The 4M dataset included estimates of ground floor area for 87 semi-detached dwellings obtained using the OS MasterMap analysis with Google Imagery. These estimates should be considered approximate, given the challenges in accurately inferring a home's floor area using such a technique; it can be difficult to know where the partition of different homes in the same building is, features such as balconies and overhangs may be included in the building footprint, and differences in floor area for different levels could be hard to differentiate.

The floor area was transformed to floor area factor and five candidate distributions were fitted for the group with a filled cavity wall constructions (see Appendix Section H.6 for other construction types). The AIC and its derivatives for each fitted

distribution is summarised in Table 5.11. The best fitting distribution is the inverse Weibull. The goodness of fit plots (Figure 5.15) suggest that the inverse Weibull provides a good fit for all but one data point.

5.2.11 Window Opening Threshold

The model input that controls window operation is the indoor temperature at which the windows would fully open within a specified availability schedule (Section 3.5.3). Any value specified for this model input will be fixed for the entire simulation period, therefore the distribution of long-term average window-opening threshold temperatures is of interest. As discussed in Section 3.5, window opening behaviour depends on several factors (including, smoking behaviour, noise levels, and security concerns) and varies between and within households (Fabi et al., 2012; Mavrogianni et al., 2016). Thus, the UK-HSM approach to modelling window opening is a simplification of a complex phenomenon.

No large-scale datasets of window opening behaviours in UK homes could be identified. To inform the probability distribution of this model input, the data collected by Rijal et al. (2007) were used. Rijal et al. (2007) conducted field surveys in 15 UK office buildings between March 1996 - Sep 1997. The findings from their longitudinal survey were used, where the indoor temperature was recorded close to 219 subjects. The subjects were asked to record their thermal satisfaction, clothing and activity level along with their use of building controls four times a day, with a total of 35,764 thermal satisfaction records collected. A strong correlation between the window state and the thermal environment was noted. From the data collected, logistic regression models were developed to predict the probability of the window being open given indoor and outdoor temperatures as predictors. A scatterplot of proportion of windows open at different globe temperatures is shown in Figure 5.16(a). The "uncorrected logit" is the result of a logistic regression model fitted to the data described by the following equation (Rijal et al., 2007):

$$logit(p) = 0.354T_g - 8.53, \tag{5.9}$$

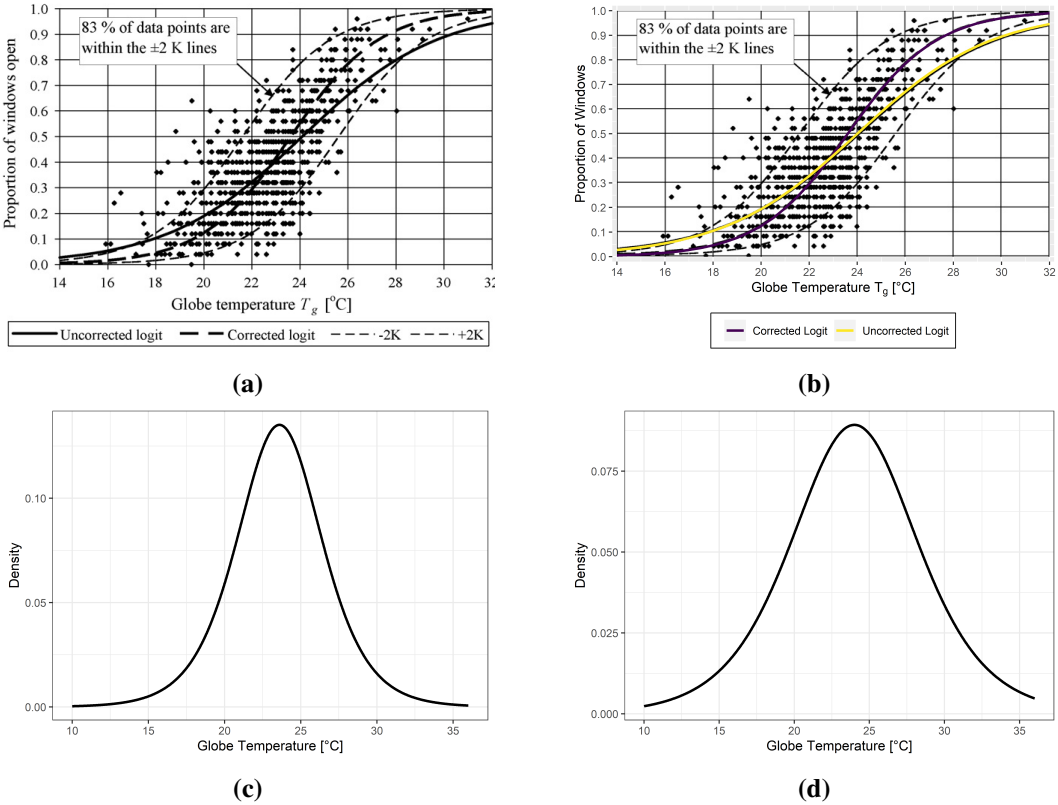


Figure 5.16: Sub-figure (a) shows the proportion of windows stated to be open at different globe temperatures by Rijal et al. (2007). In sub-figure (b), this plot was overlaid in R with best fit lines. Sub-figures (c) and (d) show the probability density functions for corrected and uncorrected logit fits, respectively.

where p is the proportion of windows being open and T_g was the globe temperature. According to Rijal et al. (2007), T_g could be considered equivalent to the operative temperature for all practical purposes. While this is an unbiased estimator of window opening, it does not describe the cloud of observations well and it does not take into account the "deadband" of temperatures; the range of temperatures over which the proportion of open windows remains unchanged (Rijal et al., 2007). For this reason, Rijal et al. (2007) fitted a "corrected logit" model, with 83 % of the data being within $\pm$2K of the regression line.

For both models in Figure 5.16(a), as T_g increases, the proportion of windows opened or the probability of any single window opening increases until a saturation level. An analogy may be made with the cumulative distribution function of the window opening threshold input of UK-HSM when treated probabilistically. Based

on Figure 5.16(a), probability distributions of temperatures at which windows would be opened were derived. Since the full dataset was not made available, Figure 5a of Rijal et al. (2007) was overlaid on an axis within the **ggplot2** library of the programming language R. Subsequently, a set of logistic distributions were fitted and the goodness of fit was visually assessed. A good fit was identified with the logistic distribution parameters of: mean = 23.6 and scale = 1.85 for the corrected logit and mean = 24 and scale = 2.8 for the uncorrected logit, visualised in Figure 5.16(b). The probability density functions for the corrected and uncorrected logit can be seen in Figures 5.16(c)–5.16(d). The corrected logit has 95 % of its probability density within the interval 16.8–30.4 °C, narrower than the equivalent probability interval of the uncorrected logit which lies between 13.7–34.3 °C. Since the aim with this model input was to identify a distribution of average window opening thresholds, the corrected logit was deemed more appropriate. This model better describes the cloud of data points and appears less extreme than the uncorrected logit which suggests that on a regular basis 2.5 % of occupants will open their windows when the indoor temperature is less than 13.7 °C.

5.2.12 Electrical Gains Factor

As discussed in Section 3.5, a single model input is used to define the intensity of electrical equipment usage, including showering and cooking. To inform this model input, the 2010-2011 Household Electricity Survey (HES) was used. Electrical power demand and energy consumption were monitored in 251 English Households, 26 for a calendar year and the rest for periods of one month through different intervals of the year (Intertek, 2012). This dataset was selected since it contained detailed information on electricity usage, collected at around the same time as the 4M survey. Thus, it was assumed to be representative of typical usage at that time. Of interest is a subset of 55 semi-detached dwellings without electric heating whose annual consumption is visualised in Figure 66 of Intertek (2012) and is reproduced in Figure 5.17(a).

While the data collected were analysed and presented in useful ways within the HES report, the raw data were not available. To infer this dataset, bar chart was

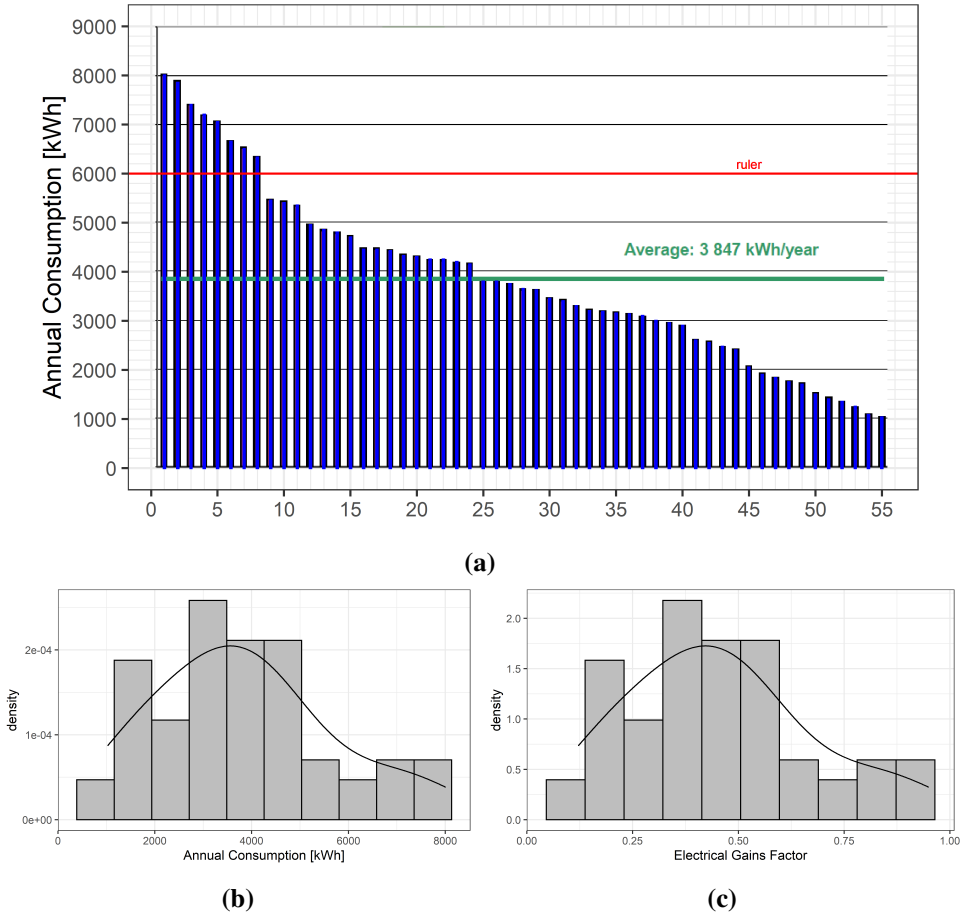


Figure 5.17: Figure (a) shows the annual consumption of electricity in semi-detached dwellings without electric heating, reproduced from Intertek (2012) and overlaid in R. Figures (b) and (c) show the distributions of annual consumption and electrical gains factor derived from (b).

Table 5.12: Distributions fitted to the inferred electrical gains factor. They are ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j).

Distributions	AIC	Δ_j	w_j	P1	P2
gamma	-15.76	0.00	0.44	shape = 4.3	rate = 9.5
weibull	-15.71	0.05	0.43	shape = 2.3	scale = 0.52
lnorm	-12.68	3.08	0.09	meanlog = -0.9	sdlog = 0.51
norm	-10.80	4.96	0.04	mean = 0.46	sd = 0.21

overlaid on a graph in R with the axes being aligned (Figure 5.17(a)). Then, one by one the values of annual consumption were measured using a digital "ruler"; a horizontal line specified within the R package **ggplot2**. Each blue vertical line indicates the inferred annual consumption per dwelling. An upper estimate of this

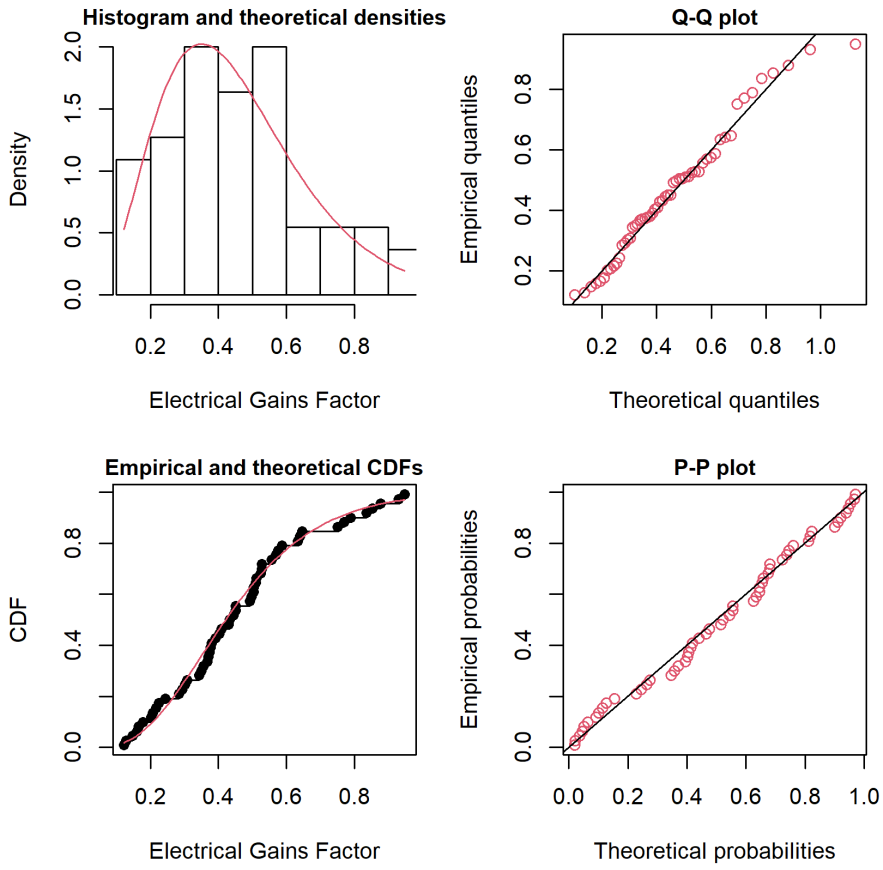


Figure 5.18: Goodness of fit plots when assuming a gamma distribution as the statistical model of electrical gains factor derived from the Household Electricity Survey.

process' uncertainty is ± 100 kWh/year, although it's likely less than that. Following this process, the mean of the inferred annual consumption was 3846 kWh/year which only differs by 0.03 % from the stated mean annual consumption of 3847 kWh/year. Figure 5.17(b) visualises the inferred distribution of annual consumption. This was converted to electrical gains factor, shown in Figure 5.17(c), by dividing it with the number of kWh associated with an electrical gains factor of 1 (8431 kWh). No extreme values appear and a long right tail is observed. Four candidate distributions were fitted to the inferred dataset, with a summary of the AIC-based measures of goodness of fit provided in Table 5.12. The best-fitting distribution was gamma with shape = 4.3 and rate = 9.5. The goodness of fit plots in Figure 5.18 suggest that gamma provides a good description of the dataset. From the Q-Q plot, only one extreme point appears while the P-P plot indicates very small difference between

empirical and theoretical probabilities.

It is notable that the average consumption of 3.847 kWh/year is less than half of the value assumed by an electrical gains factor of one. Some of this difference may be explained by the fact that not all homes in the dataset had an electric oven, stove and shower as assumed in UK-HSM. However, part of this discrepancy is likely due to the assumed power and schedules of cooking appliances being at the upper end of the spectrum. For example, the average annual consumption of oven use for multiple pensioner households was reported by Intertek (2012) to be 211 kWh, while UK-HSM assumed a value of 2.792 kWh. Following from this result, future work aims to review the schedules of UK-HSM, although this was not within the scope of this thesis.

5.2.13 UK-HSM Distributions

A summary of the distributions identified for the cluster of filled cavity wall dwellings with double glazing is shown in Table 5.13. Ten data sources were used, with eight of the twelve model inputs informed by empirical data. Two model inputs could not be described adequately with a unimodal distribution, the Roof and Floor U-value. For these model inputs, a distribution was not fitted. As described in Section 3.2, if either of these model inputs are found to be influential during the sensitivity analysis, they would be used as classifiers and the group would be further segmented.

5.3 Discussion

This chapter described in detail the process of identifying probability distributions for UK-HSM model inputs based on the best available evidence. Contrary to published literature reviewed in Section 2.4.5.1, the distributions were not constrained to be uniform (Cerezo et al., 2017; Sokol et al., 2017; Wang et al., 2020), since the use of detailed distributions is thought to offer important benefits to building stock modelling. Well-informed probability distributions used as priors in Bayesian calibration can improve posterior parameter identifiability (Smith, 2013), and are likely to result in better predictions. Even if the model input distributions are not used for calibration purposes, they can inform the model sensitivity and forward

Table 5.13: Model input distributions identified for the group of semi-detached dwellings with filled cavity wall in the 4M dataset. The empirical distributions of Roof and Floor U-value were multimodal, and theoretical distributions were not identified for either model input.

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-to-Ceiling Height	lnorm(meanlog = 0.93, sdlog = 0.034)	[4]
Floor Area Factor	invweibull(shape = 5.5, scale = 0.74)	[3]
Window Opening Threshold	logis(location = 23.6, scale = 1.85)	[9]
Electrical Gains Factor	gamma(shape = 4.3, rate = 9.5)	[10]

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

uncertainty analysis (Tian and Choudhary, 2012; Tian et al., 2018). When empirical data are available, then more trust can be placed in the model input distributions, and their use instead of theoretical values can result in substantial differences in the model output (Petrou et al., 2021a). In the absence of empirical observations, a distribution that captures the assumed uncertainty around a theory-based value is better than the use of a single, fixed, value. A thorough description of the approach taken was provided for increased transparency and to enable other researchers to define appropriate distributions for their modelling work

To characterise the model inputs using the available empirical data, the distribution fitting method introduced in Section 5.1.1 was applied several times. Based on this experience, the method is flexible, quick and easy to implement. The greatest difficulty is finding alternatives to the commonly used distributions when they do not provide an adequate description of the empirical data. This could be addressed by reviewing the properties of different distributions and selecting candidates according to their characteristics, similarly to how the Fréchet was identified for this application. Alternatively, one can leverage the great number of distributions already provided

by R and fit several (or all) of them in a "brute-force" approach (see Petrou (2021) for a demonstration). In either case, a careful evaluation of the goodness-of-fit must follow

It is also important to reflect on what stochastic characterisation achieves. It enables modellers to identify *theoretical* distributions, in the form of parametric probability models, which appropriately describe the *empirical* distributions available or the modeller's best understanding if empirical data is not available (see Wild (2006) for a broader discussion around distributions). Even if empirical data are available, it is not possible to claim that a mathematical definition of the data generating process has been identified but merely a satisfactory description of the best indication we have of the data generating process, the empirical distribution. Taking the example of wall U-values, the data generating process is the mechanism behind the distribution of true filled cavity wall U-values within the English housing stock. This process would have multiple components, such as the effect of using different insulation materials, the change of building practices and regulations over time and workmanship. Since it is infeasible to accurately model this mathematically, a snapshot of this process' output is used by taking measurements for a sample of filled cavity walls. This results in the empirical distribution of true wall U-values augmented with measurement and sampling errors shown in Figure 5.3. By following the procedure described in Section 5.1.1, a theoretical distribution was identified that adequately describes the dataset. The appropriateness of the assumed distribution will vary depending on the application and the potential impact that a misrepresentation of the data might have on the model output.

5.3.1 Limitations

Despite the efforts to identify detailed, empirically-informed distributions for each UK-HSM model input, data availability, quality and limitations with the approach followed have influenced the results.

For inputs relating to the archetype's spatial characteristics, such as the Floor-to-Ceiling Height or the Glazing Fraction, distributions could be informed by the EHS. Due to the relatively large and representative sample of homes within the EHS,

well-informed distributions were developed. Floor Area Factor was directly informed by the 4M dataset, and could potentially be used as an explanatory variable in the calibration if the influence of its measurement error is not large; this is explored in Chapter 6. For model inputs where there was a lack of large-scale empirical datasets (such as, Solar Absorptivity, Floor, Window and Roof U-value) probability distributions had to be assumed based on information from RdSAP and CIBSE Guide A (BRE, 2019; CIBSE, 2015). In some cases, observations were used to inform the model input distributions, yet the confidence in them is limited.

A noteworthy example is the probability distribution selected for the Window Opening Threshold. This distribution was chosen by identifying a best fit line on a graph of empirical data through visual inspection (Section 5.2.11). Such an approach is not without uncertainties, and the quality of the fit partly depends on the figure resolution. Nevertheless, the major limitation in this case is expected to be the data used. Measurements were made in offices, which are thought to offer less adaptive potential to the indoor thermal environment than homes (Oseland, 1995; Pathan et al., 2017; Rupp et al., 2015). Thus, a distribution based on measurements from homes could differ. Another limitation, likely of lesser importance, is that measurements were based on globe (or operative) temperature, while the model operates the windows based on air temperature.

A further limitation of this work is that model input correlations were not quantified. Several building characteristics, and thus UK-HSM model inputs, are expected to be correlated. One such example is likely the wall U-value and fabric air permeability. As a result of building regulations prescribing both parameters (HMG, 2016), and due to the fact that wall insulation tends to reduce infiltration (Hong et al., 2004), homes with a lower wall U-value are more likely to have a lower than average fabric air permeability. Since such relationships were not explored, the distributions identified will be used in later chapters while assuming no correlation. This is a common limitation amongst building stock modelling studies (Booth et al., 2012; Sokol et al., 2017). The main challenge in addressing this limitation is the lack of large-scale unified datasets of building parameters from which correlations could

be established. A workaround, at least for the subset of calibration variables, is the quantification of correlations in the posterior distributions (Kristensen et al., 2018).

5.4 Summary

A novel approach for identifying appropriate probability distributions for building energy model inputs was introduced. The method, which can be adapted depending on the format and availability of empirical data, was used to characterise the UK-HSM continuous model inputs (Section 5.2.13). Empirical data from ten sources were used to inform eight out of the twelve model inputs, although the sample sizes and associated assumptions varied. With the exception of Floor and Roof U-value, all other model inputs were adequately described by unimodal distributions. The only model input where sufficient data were available to be informed by the 4M survey was Floor Area Factor.

The outcomes of this chapter inform the later chapters. The following chapter will identify which model inputs have the greatest influence on the model output, given their distribution. Subsequently, in Chapter 7, the uncertain and influential model inputs will be calibrated with their prior based on the distributions identified in this chapter.

Chapter 6

Sensitivity Analysis

One of the main outputs of Chapter 4 was the selection of a potentially homogeneous group of 26 semi-detached dwellings, monitored during the 4M survey. Determining whether they are indeed homogeneous, according to the definition introduced in Section 3.2, firstly required a probability distribution to be defined for each continuous model input of the UK Housing Stock Model (UK-HSM); this was accomplished in Chapter 5. To complete the last part of work needed to address the second research objective (Section 3.1), and inform the fifth and final step of the Bayesian calibration framework (Section 3.2), two key tasks must be completed: (1) determine whether any model inputs described by multimodal distributions have a large influence on the model output, thus, requiring further segmentation to identify a homogeneous group of dwellings, and (2) select the UK-HSM inputs with the greatest influence on summer indoor temperature of the homogeneous group. As depicted in Figure 6.1, this chapter describes the use of sensitivity analysis to achieve these tasks.

The Morris method (also known as the method of elementary effects) was used to carry out the sensitivity analysis. The theory underpinning this approach is summarised in Section 6.1.1, while details about its implementation are provided in Section 6.1.2. A way of determining whether the Floor Area Factor, the only UK-HSM input that could be informed by the 4M survey, should be used as an explanatory variable in the calibration is described within the same section. The results from this work are presented in Section 6.2 and discussed in Section 6.3. A summary is offered in Section 6.4.

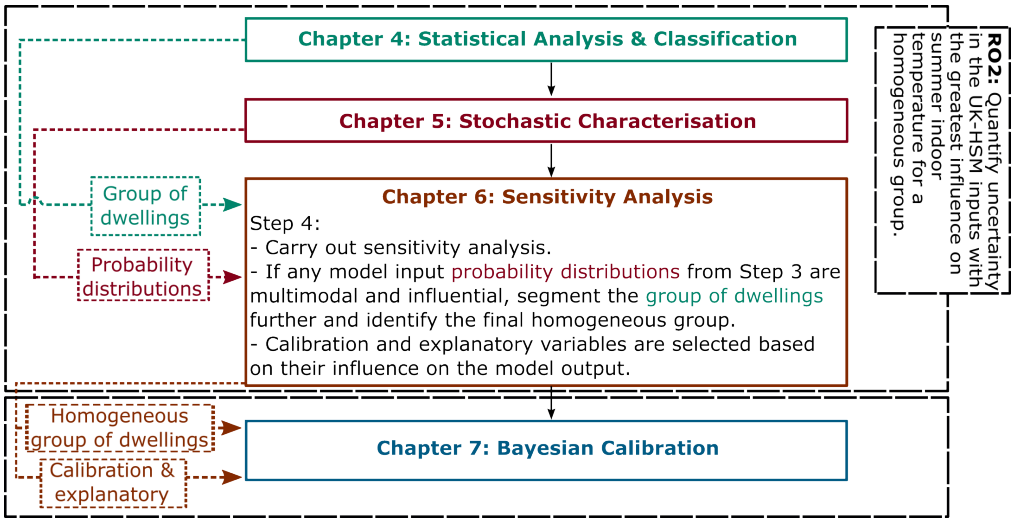


Figure 6.1: Chapter 6 flowchart. This is an abridged version of Figure 1.3, focusing on Chapter 6 and its outputs. RO2 is a shortened version of Research Objective 2. UK-HSM stands for UK Housing Stock Model.

6.1 Methods

The Morris method is a screening technique whose use in selecting building energy model inputs to be calibrated using Bayesian inference was first suggested by Heo et al. (2012), and since then it has been the most frequently used method for such purposes according to the review by Hou et al. (2021). For this chapter's brevity, and since the theory of the Morris method has been discussed at length within the built environment literature (e.g. Kristensen and Petersen, 2016; Petersen et al., 2019), the following section will provide a short summary of this technique, while an in-depth description is provided in Section G of the appendices.

6.1.1 The Morris Method

The Morris method is a global sensitivity analysis technique[1] with a low computational cost compared to other global approaches (Tian, 2013). This technique relies on repeated one-at-a-time sampling from a p-level grid of the input space, where the baseline set of inputs changes with each repetition (Morris, 1991; Tian, 2013; Saltelli et al., 2008). For each *trajectory*, an *elementary effect* is estimated for each model input, which is the difference in the model output due to a change in the model input,

[1] Tian (2013) defines global sensitivity analysis as "interested in the influences of uncertain inputs over the whole input space"

divided by the change in model input (Equation G.1 in the appendices). By repeating this process for r trajectories, a distribution of r elementary effects is generated per model input (Kristensen and Petersen, 2016). The original method suggested using the mean (μ) of the elementary effects to assess the overall influence of the factor on the output. As an improvement, Campolongo et al. (2007) proposed the use of the mean (μ^*) of the absolute elementary effects, defined as:

$$\mu_i^* = \frac{1}{r}\sum_{t=1}^{r} |EE_{it}|, \tag{6.1}$$

where, $|EE_{it}|$ is the absolute value of the elementary effect for model input i in trajectory t. Contrary to μ, μ^* is not sensitive to sign of the elementary effects and provides a reliable ranking of parameters (Campolongo et al., 2007). The standard deviation of elementary effects (σ) was proposed as a quantity indicative of the model inputs' non-linear effect, or interaction with other inputs (Morris, 1991). Building on this idea, Garcia Sanchez et al. (2014) used the σ/μ^* ratio to identify non-linear and interaction effect based on the following categories: (i) Almost linear effects are identified in the region of $\sigma/\mu^* < 0.1$, (ii) monotonic effects at $0.1 < \sigma/\mu^* < 0.5$, (iii) almost monotonic at $0.5 < \sigma/\mu^* < 1$ and (iv) non-monotonic or with interactions at $\sigma/\mu^* > 1$.

A limitation of the Morris method is that it does not take into consideration the distributional form of the model inputs, only their upper and lower bound (Kristensen and Petersen, 2016). The impact of this limitation was assessed by Kristensen and Petersen (2016) where the more computationally expensive Sobol's method was used as a reference point for two ISO 13790 models. While the ranking of model inputs was near-identical when uniform distributions were assumed for the Sobol's method, the ranking differed when non-uniform distributions were used. However, the Morris method was able to identify the group of six most influential model inputs at roughly 2 % of the computational cost. Furthermore, in an investigation of high-fidelity models (including EnergyPlus-based models), it was demonstrated that the ranking of variables according to the Sobol's technique had not converged after 260,000 simulations (Petersen et al., 2019), but groups of high and low influence on the model

output could be identified, and they were largely in agreement with those offered by the Morris method. Given these findings, the use of the Morris method is considered reasonable if the purpose is to identify groups of influential model inputs, since it has been shown to achieve similar results as the Sobol's method, at a fraction of the computation cost, and the exact ranking of model inputs using the Sobol's method may not converge even after hundreds of thousands of simulations.

Two important considerations when implementing the Morris method are the number of levels (p) and trajectories (r) used. This was investigated by Petersen et al. (2019), who concluded that for high fidelity models, and when the aim is to identify a group of important input factors, p values equal or greater than 4 and r values ranging from 100 to 1000 may be needed (Petersen et al., 2019). However, a higher number of levels, closer to 12, was thought to be beneficial when cooling energy was the output of interest.

6.1.2 Implementation

Following the recommendation of Petersen et al. (2019), the experiment design was based on a trajectory number (r) of 500, but the simulations were run in batches of $r = 25$. With every batch, the absolute mean of the elementary effects was qualitatively assessed for convergence. The batches are smaller than Petersen et al. (2019) recommended in the effort of reducing computational cost. With no prior information regarding the appropriate level number (p) for this study, p was set to 12 since this was the value that Petersen et al. (2019) determined to be best when modelling cooling energy. While it cannot be assumed that the same value of p would be best for both models, $p = 12$ was the upper limit assessed by Petersen et al. (2019) and it results in a relatively large sampling resolution, albeit at a higher computational cost for the same value of r. The Python package **SALib** (Herman and Usher, 2017; Iwanaga et al., 2022) was used to sample the model inputs according to the Morris method and analyse the results following the simulations.

All continuous model inputs were sampled, with their upper and lower bounds summarised in Table 6.1. For each input, the upper and lower bound was defined to be the 0.5th and 99.5th percentile of the corresponding distributions listed in Table 5.13,

Table 6.1: Lower and upper bounds used for the two stages of sensitivity analysis.

Parameter	Parameter Name	Stage 1 Range	Stage 2 Range
P1	Wall U-value	0.26-1.48	0.26-1.48
P2	Window U-value	1.70-3.27	1.70-3.27
P3	Roof U-value	0.10-2.53	0.10-2.53
P4	Floor U-value	0.22-0.90	0.22-0.90
P5	Permeability	1.7-24.7	1.7-24.7
P6	Solar Absorptivity	0.16-0.96	0.16-0.96
P7	Glazing Fraction	0.12-0.48	0.12-0.48
P8	Orientation	0.0-330.0	0.0-330.0
P9	Floor-to-Ceiling Height	2.32-2.77	2.32-2.77
P10	Floor Area Factor	0.55-1.94	0.9-1.1
P11	Window Opening Threshold	18.0-32.0	18.0-32.0
P12	Electrical Gains Factor	0.08-1.21	0.08-1.21

with only two exceptions: the window opening threshold and the orientation. For the window opening threshold, the 0.5th and 99.5th percentiles corresponded to 13.8 °C and 33.4 °C, respectively. Since this model input represents the temperature at which occupants will always open their windows if the schedule allows and the indoor temperature is greater than the outdoor, both values were considered extreme, especially the lower bound. Instead, the window opening threshold was sampled within the range of 18–32 °C. For the orientation, the lower and upper bounds of the specified uniform distribution (0–330°) were used to capture the impact of dwellings with a front façade facing south. For the rest of the continuous model inputs, a 99 % interval was used since it was considered to provide a reasonable coverage of the input space. It was not possible to use the maximum range because some input distributions were unbound in one or both directions (i.e. extending with a non-zero probability to infinity). For the two multimodal distributions (Roof and Floor U-value), the 99 % interval was estimated using their empirical distribution.

As explained in Section 3.2, the cluster of dwellings might have been further segmented and a second stage of the sensitivity analysis would have been carried out if any of the multimodal model inputs were found to have a significant influence on the model output. As will be discussed in Section 6.2, this was not considered necessary for this application of the framework. However, a second stage of the sensitivity analysis was implemented in order to determine whether the Floor Area

Factor (FAF), the only continuous model input for which data from the 4M survey were available, should be calibrated or used as an explanatory variable. While in the first stage of the sensitivity analysis FAF was allowed to vary within the 99 % interval (0.55–1.94) of its distribution, in the second stage it was constrained to the interval of 0.9–1.1. The Stage 2 interval captures the idea that the FAF of a typical house (FAF = 1) has a measurement error of 10 %. If FAF had a relatively small influence in Stage 1, it would be kept fixed in the calibration. If FAF had a relatively large influence in Stage 1, but a small influence in Stage 2, it could be used as an explanatory variable since its effect on the model output is substantial but the impact of its measurement error is not. Finally, if the influence of FAF in Stage 2 is large, relative to other model inputs, the uncertainty surrounding this model input would be considered substantial, and the model input would have been calibrated.

Table 6.2: Model inputs kept fixed during the sensitivity analysis

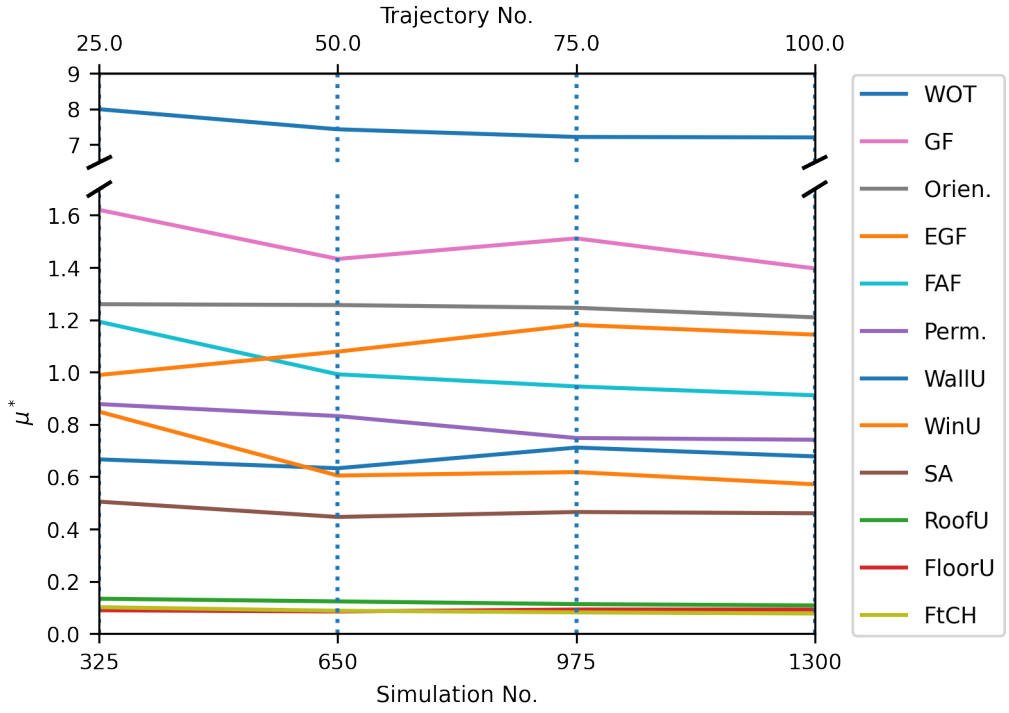
Parameter	Value
Dwelling Type	Semi-detached
Wall Type	Filled cavity
Occupancy Type	Pensioners
Terrain	Urban

Simulations were carried out using the 2009 weather data obtained for Leicester; further information on the weather file was provided in Section 3.4.4. The model output of interest was the summer-averaged Mean of the Daytime Living Room Temperature (MDLRT). The choice of MDLRT, largely based on the significance of this model output for heat-mortality estimation, was explained in Section 3.3. Since the Morris method requires a single output per simulation, the summer-averaged MDLRT was considered a reasonable choice. The summer period in this case was assumed to be July–August 2009, since this was the period with sufficient monitored temperature data for the calibration stage. The values for the categorical model inputs of UK-HSM are summarised in Table 6.2. The *dwelling type*, *wall type* and *terrain* were based on the classifiers used to select the cluster of 4M dwellings (Section 4.2.6.2). The *occupancy type* was informed by the most frequent household composition in the same cluster.

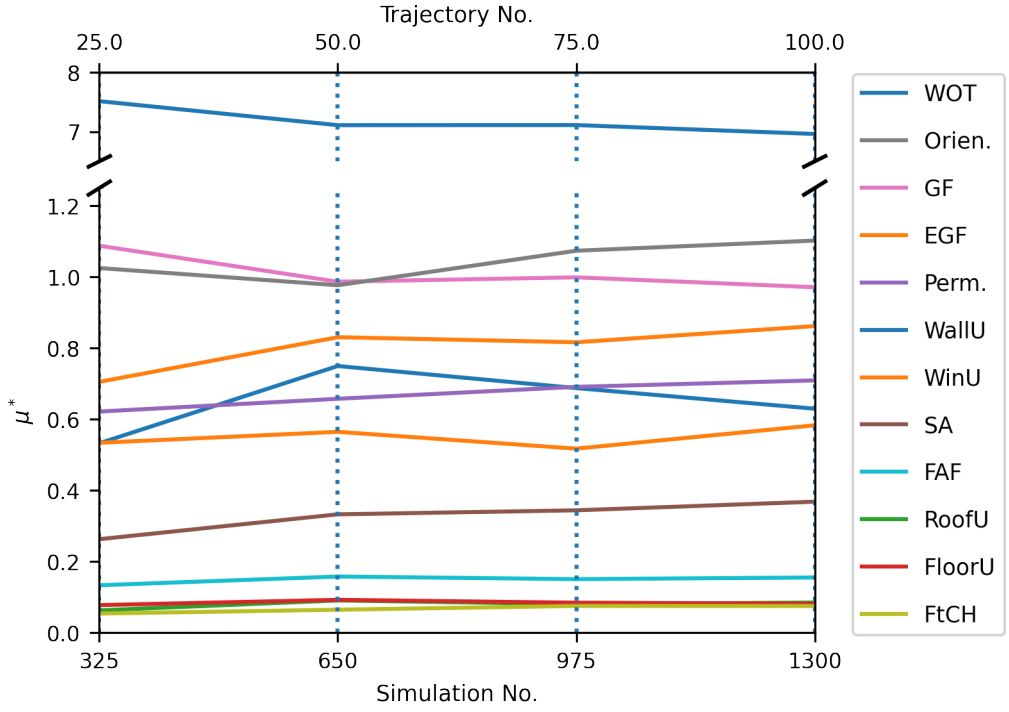
6.2 Results

The convergence plots for the two stages of sensitivity analysis are displayed Figure 6.2. The absolute mean (μ^*) was quantified after 325, 650, 975 and 1300 simulations, corresponding to 25, 50, 75 and 100 trajectories. The relative change in μ^* depending on trajectory number, visualised by the trend lines, allows the evaluation of convergence for each parameter. Note that for both stages, a "broken axis" approach to plotting was used where a discontinuity exists on the y-axis. This enabled the visualisation of convergence for all parameters on the same plot, despite the large range of μ^* values. Fluctuations of μ^* were observed in both stages, although distinct groups could be identified. Given the level of fluctuations, the within-group order may change if more simulations were carried out, but the between-group order is expected to remain unchanged. Since convergence cannot be guaranteed at a specific trajectory number, and a specific μ^* value cannot guarantee whether a parameter should be used for calibration or kept fixed, running further simulations was deemed unnecessary.

In Stage 1, the Window Opening Threshold is in the first group with $\mu^* = 7.20$ after 1300 simulations, making it the most influential parameter. Glazing Fraction, Orientation, Electrical Gains Factor, Floor Area Factor, Permeability, Wall U-value, Window U-value and Solar Absorptivity are part of the second group. Although this group includes eight model inputs, and their exact order varies depending on the trajectory number, a few useful distinctions can still be made. The Glazing Fraction and Orientation, with a μ^* of 1.40 and 1.21 after 1300 simulations, are always more influential than Permeability, Wall U-value, Window U-value and Solar Absorptivity with a $\mu^* < 0.9$ regardless of the trajectory number. The Floor Area Factor, Permeability, Solar Absorptivity, Window and Wall U-value follow in the same group, ranging in values between $\mu^* = 0.8$ and $\mu^* = 0.5$. The final group includes the Floor U-value, Floor-to-Ceiling Height and Roof U-value, which have largely converged with a μ^* ranging between 0.08–0.11. Given their relative importance, all parameters in the last group can be considered non-influential and will be kept fixed in the calibration stage. This group includes both multimodal



(a) Stage 1



(b) Stage 2

Figure 6.2: Convergence plots for the two stages of sensitivity analysis for the living room. Disambiguation: WallU = Wall U-value; WinU = Window U-value; RoofU = Roof U-value; FloorU = Floor U-value; Perm. = Permeability; SA = Solar Absorptivity; GF = Glazing Fraction; Orien. = Orientation; FtCH = Floor-to-Ceiling Height; FAF = Floor Area Factor; WOT = Window Opening Threshold; EGF = Electrical Gains Factor.

parameters (Roof and Floor U-value), hence further classification is not necessary.

The second stage of sensitivity analysis aims to quantify the relative importance of measurement uncertainty in the Floor Area Factor model input. Similarly to Stage 1, fluctuations are observed for most parameters. As expected, the Floor Area Factor μ^* has reduced (from $\mu^* = 0.91$ to $\mu^* = 0.16$), with an influence comparable to that of the third group of model inputs. While it was only the sampling range of the Floor Area Factor that differed between the two stages, the relative importance of some other parameters has also changed in Stage 2. As an example, although the relative importance of Glazing Fraction remains high compared to other model inputs, its μ^* had decreased from $\mu^* = 1.40$ to $\mu^* = 0.97$.

While the sampling process could have contributed to the differences in μ^* between the two stages for parameters other than the Floor Area Factor, another potentially important factor is parameter interaction. Figure 6.3 visualises the standard deviation (σ) and μ^* of each parameter for the two stages of sensitivity analysis for a trajectory number of 100. With the exception of the Window Opening Threshold which falls within the area of "monotonic" behaviour, all other parameters are either "almost monotonic" or "non-linear and non-monotonic". The latter category is also an indication of potential parameter interaction (Garcia Sanchez et al., 2014). It is therefore likely that an interaction exists between the Floor Area Factor and one or more other parameters. By reducing the range of Floor Area Factor in Stage 2 compared to Stage 1, it is likely that the relative importance of other parameters also changes since their influence might depend on the floor area. Further investigation of these interactions were out of the scope of this research.

A summary of the relative importance of each model input and its categorisation in regard to the calibration stage is provided in Table 6.3. Floor U-value, Roof U-value and Floor-to-Ceiling Height are all classified as non-influential and will be kept fixed at the calibration stage. Assuming an uncertainty of $\pm 10\%$ around the Floor Area Factor value 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 considered to be

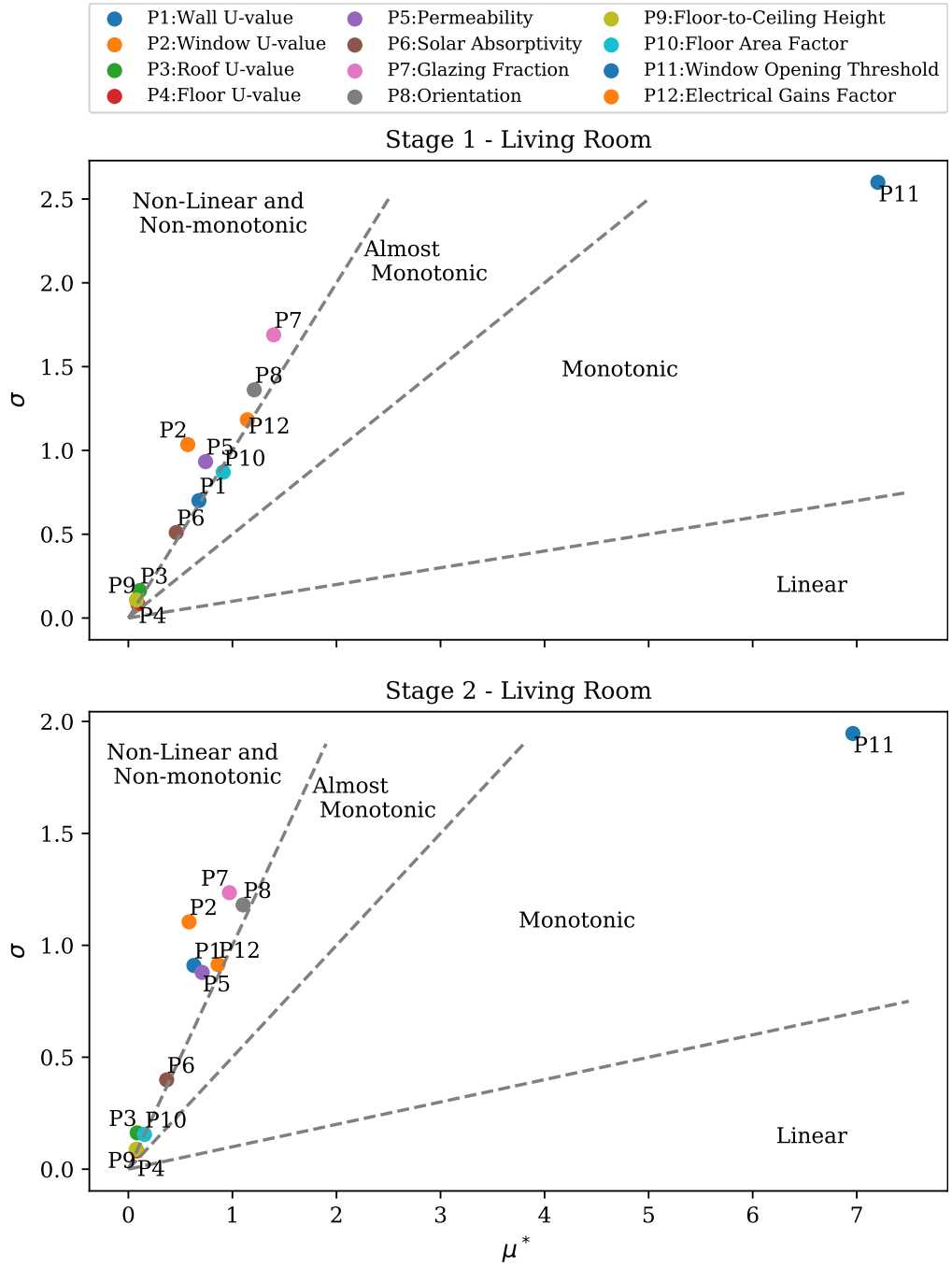


Figure 6.3: Scatter plots of the standard deviation (σ) of elementary effects for each parameter against their absolute mean (μ^*) for the two stages of sensitivity analysis for the living room. The metrics are based on 100 trajectories, equivalent to 1300 simulations.

relatively unimportant and this parameter may be used as an explanatory variable. In both stages of the sensitivity analysis, Window Opening Threshold was the dominant model input and will be calibrated, together with the Orientation, Glazing Fraction

Table 6.3: Summary of the rank and absolute mean of elementary effects (μ^*) for each parameter, ordered in ascending order of Stage 2 rank. Type corresponds to how each parameter will be treated at the calibration stage, with variables that will be calibrated in bold font.

Parameter Name	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)	Calib./Fixed
Wall U-value	7 (0.68)	6 (0.63)	Calib./Fixed
Window U-value	8 (0.57)	7 (0.58)	Calib./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

and Electrical Gains Factor. Whether other parameters will be calibrated or kept fixed will depend on whether their addition results in over-parametrisation, in line with the advice of Chong and Menberg (2018).

6.3 Discussion

The application of the Morris method in Step 4 of the Bayesian calibration framework (Section 3.2) led to several important outcomes. It has demonstrated that the relative influence of the two UK-HSM model inputs described by multimodal distributions (Roof U-value and Floor U-value) on the summer-averaged MDLRT is small, when compared to that of other model inputs. Thus, further segmenting the group of 26 semi-detached dwellings identified in Section 4.2.6.2 is not necessary, and this group can be considered homogeneous according to the definition provided in Section 3.2.

A further outcome was the selection of calibration variables to be used in the fifth and final step of the calibration framework. For the assumed upper and lower bounds, informed by the findings of Chapter 5, the Window Opening Threshold was the dominant model input. While the importance of window opening in determining summer indoor temperature has been previously suggested (Mavrogianni et al., 2014;

Taylor et al., 2018b), this is the first time its impact was demonstrated through a simultaneous assessment of all other UK-HSM continuous model inputs (when varying through extreme values). Following from this result, it is expected that window opening behaviour will have a critical effect on summer indoor temperature, at least in semi-detached homes. In cases where window opening is restricted, the influence of other building or occupant characteristics is likely to be secondary. Furthermore, this finding also highlights the need to carefully specify this model input in UK-HSM, and window opening specification more generally in models of naturally ventilated homes. Given the scarce empirical data available to inform this modelling input, model calibration of this variable would be especially beneficial. Also in agreement with previous studies, building orientation was shown to be important (Taylor et al., 2014). Glazing Fraction, which impacts both the level of solar gains and the ventilative potential of the modelled homes, was also shown to be influential. Of comparable importance was the Electrical Gains Factor, followed by Permeability, Wall and Window U-value. Roof U-value was shown to have little influence on the model output, likely due to the analysis focusing on the living room temperature, and in agreement with the findings in Chapter 4.

There are several examples of the use of the Morris method for selecting the inputs of building energy models to be calibrated using Bayesian inference, as highlighted in the review of Hou et al. (2021). A novelty introduced in this work is the two-stage application of the Morris method to determine whether a model input, in this case Floor Area Factor, should be treated as an explanatory, fixed or calibration variable. It was determined that the Floor Area Factor should be treated as an explanatory variable, since the uncertainty surrounding individual values due to measurement error is small compared to its influence when considering the overall variation within the homogeneous group of dwellings.

Finally, the two-stage sensitivity analysis suggested an interaction between Floor Area Factor and other model inputs (for example, Glazing Fraction), and a non-linear relationship between most model inputs and the summer-averaged MDLRT. Due to such interactions, it is expected that the ranking obtained for most model

inputs will partly depend on the upper and lower bounds chosen for other variables. This would be the case for any model where interaction between inputs exists, and thus, it is important to carefully consider the bounds of all variables assessed in the sensitivity analysis.

6.3.1 Limitations

Although this chapter describes the successful implementation of Step 4 of the Bayesian calibration framework (Section 3.2), and the last component of work needed to address the second research objective (Section 3.1), limitations exist.

One such limitation relates to the use of the Morris method in determining whether a model input should be treated as explanatory in the calibration. In Bayesian calibration studies of building energy models, where the Morris method is frequently used, it is common practice to use weather variables as explanatory variables in the calibration process (Chong and Menberg, 2018; Kristensen et al., 2018; Menberg et al., 2019). Weather variables were also used as explanatory variables in this study, as will be discussed in Section 7.1.4. Since the Morris method requires a sampling procedure before the simulations are run, it is not possible to determine the influence of a time-dependent input, such as the hourly dry bulb temperature, contrary to a static model input (e.g. floor area). Thus, while the procedure introduced in this chapter informed the use of FAF as an explanatory variable, it did not enable the simultaneous investigation of weather-based variables as explanatory variables.

Through the two-stage sensitivity analysis, and the consideration of σ/μ^*, this study revealed the interaction and non-linear behaviour of some UK-HSM model inputs. While informative, this investigation is not complete, since only the interaction of FAF with other model inputs was investigated. Studying these behaviours was not required for the current research, but extending this analysis to study and quantify the most important model interactions and non-linearities could be instructive to future modelling applications of UK-HSM.

6.4 Summary

Through the two-stage application of the Morris method, it was determined that the influence of the Floor and Roof U-value on the model output of interest was small compared to that of other UK-HSM model inputs. Thus, further segmenting the group of 26 semi-detached dwellings selected in Section 4.2.6.2 is not necessary. In addition, this work revealed that the Floor Area Factor, the only model input that could be informed by the 4M survey, could be used as an explanatory variable in the calibration stage. The Window Opening Threshold was shown to be the most dominant model input and was selected to be calibrated together with the Orientation, Glazing Fraction and Electrical Gains Factor. The analysis also suggested that model input interactions exist, and it is likely that several model inputs exhibit a non-linear relationship with the model output.

Informed by the outcomes of Chapters 4–6, the following chapter will describe the Bayesian calibration of UK-HSM for the homogeneous group of 4M dwellings.

Chapter 7

Bayesian Calibration

For each input of the UK Housing stock Model (UK-HSM), a probability distribution was identified in Chapter 5, and its influence on the model output was ranked in Chapter 6. In addition, Chapter 6 confirmed that the group of 26 semi-detached dwellings, selected from within the 4M dataset in Section 4.2.6.2, can be considered homogeneous, according to the definition offered in Section 3.2. In response to the third research objective, and building on these outcomes, this chapter presents the improvement in UK-HSM's predictive ability following the application of the Bayesian calibration framework, and the reduction in model input uncertainty (Figure 7.1).

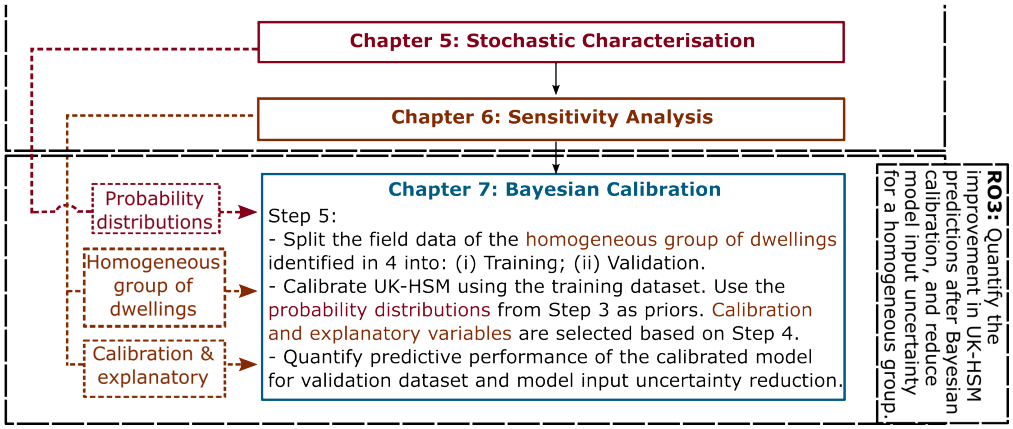


Figure 7.1: Chapter 7 flowchart. This is an abridged version of Figure 1.3, focusing on Chapter 7 and its outputs. RO3 is a shortened version of Research Objective 3. UK-HSM stands for UK Housing Stock Model.

In Section 7.1, the methods behind the calibration procedure are described in detail. Section 7.1.1 discusses the monitored and simulated data used for the

calibration, including the choice of temporal resolution and the method for data standardisation. Informed by the review of archetype-based Bayesian calibration studies (Section 2.4), which all focused on building energy performance, a statistical formulation was selected. This is detailed in Section 7.1.2, while the priors are summarised in Section 7.1.3. To quantify the impact that the choice of weather and calibration variables has on the calibration, a parametric experiment was performed, whose details are summarised in Section 7.1.4. A novel method for implementing Bayesian calibration is introduced in Section 7.1.6.

The results from this chapter are presented in Section 7.2. Following an initial exploratory analysis (Section 7.2.1) of the monitored and simulated data carried out prior to the calibration, the results of the parametric calibration experiment are presented (Section 7.2.2). A single instance of the parametric analysis is explored in more detail in Section 7.2.3, while the performance of the alternative implementation is quantified in Section 7.2.4. The results and the limitations of this chapter are discussed in Section 7.3. The chapter concludes with a summary in Section 7.4.

7.1 Methods

A detailed description of the methods used in this chapter is offered in the following sections, with a summary provided in Figure 7.2.

7.1.1 Data

7.1.1.1 Monitored Data

The monitored data consisted of hourly measurements of indoor temperature, collected as part of the 4M survey in Leicester, for the 26 semi-detached dwellings assigned to the same homogeneous cluster (Section 4.2.6.2). More information on the 4M survey is provided in Section 3.4.3. Following the data cleaning process described in Section 3.4.3.2, only the subset of homes whose indoor temperature was thought to be free-floating during the summer period was used, since UK-HSM assumes no heating or cooling between May and September. For each monitored dwelling, the Floor Area Factor was estimated using floor area data included in the 4M dataset.

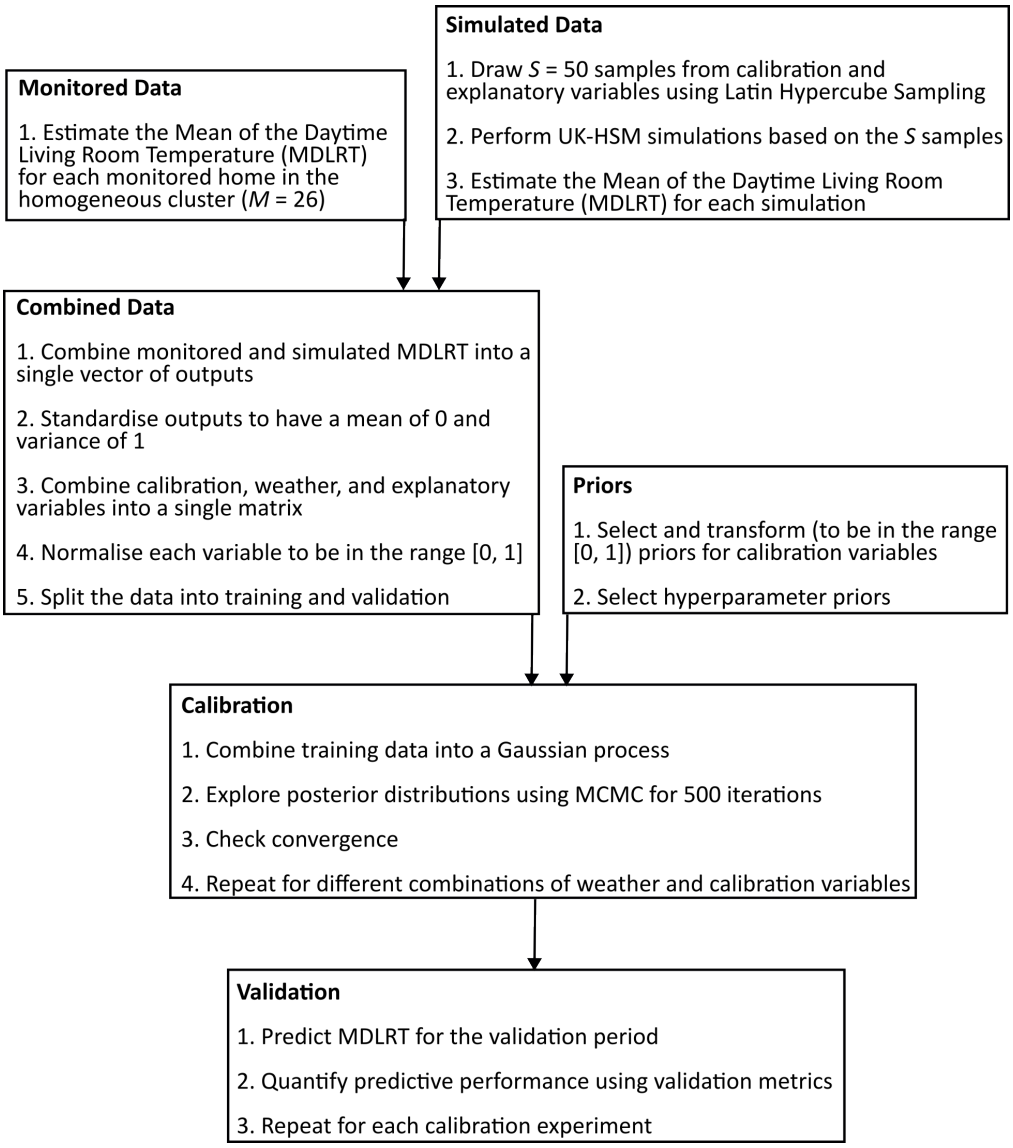


Figure 7.2: Workflow diagram for the Bayesian calibration.

7.1.1.2 Simulated Data

To generate the simulation data, 50 simulations of the semi-detached archetype in UK-HSM were run (i.e. $S = 50$) using inputs sampled with the commonly used Latin Hypercube Sampling (LHS) procedure (Tian et al., 2018), and following the suggestion of having at least ten samples per variable (Chong and Menberg, 2018). Only the selected calibration (Window Opening Threshold, Orientation, Glazing Fraction and Electrical Gains Factor) and explanatory variables (Floor Area Factor) were sampled. The samples were drawn from uniform distributions whose lower and upper bounds were the same as those used in the screening procedure, with

a summary offered in Table 7.1. Sampling uniformly ensured that the surrogate model was able to represent the computer model well across the entire range of input values. All other continuous model inputs, shown in Section 6.2 to have a smaller influence on the living room temperatures, were kept fixed at the median value of their respective distributions. Categorical model inputs were specified as the most frequent value in the cluster. Simulations were run using the 2009 Leicester weather file, developed from data made available by the Met Office Integrated Data Archive System (MIDAS), as described in Section 3.4.4.

Table 7.1: Lower and upper bounds of model inputs sampled to train the surrogate model, and values of model inputs kept fixed.

Model input	Type	Sampling range or Value
Window Opening Threshold	Calib.	18.0-32.0
Orientation	Calib.	0.0-330.0
Glazing Fraction	Calib.	0.12-0.48
Electrical Gains Factor	Calib.	0.08-1.21
Permeability	Fixed	11.29
Wall U-value	Fixed	0.71
Window U-value	Fixed	2.5
Solar Absorptivity	Fixed	0.63
Floor Area Factor	Explan	0.55-1.94
Roof U-value	Fixed	0.51
Floor U-value	Fixed	0.70
Floor-to-Ceiling Height	Fixed	2.53
Dwelling Type	Fixed	Semi-detached
Wall Type	Fixed	Filled cavity
Occupancy Type	Fixed	Pensioners
Terrain	Fixed	Urban

7.1.1.3 Data Resolution for UK-HSM Calibration

The calibration of UK-HSM was based on observations of the mean of the daytime living room temperature (MDLRT). The monitored and simulated hourly data points between 08:00-22:00 were averaged for each day of the month in July and August. As explained in Section 3.3, the MDLRT was chosen due to the use of a comparable metric in previous applications of UK-HSM where changes in heat-related mortality, associated with installation of energy efficiency measures or overheating interventions, were quantified (Taylor et al., 2015; Taylor et al., 2018b).

7.1.1.4 Data Transformation

As it is common practice in Bayesian calibration procedures, the calibration and explanatory variables were standardised 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 (Higdon et al., 2004; Chong et al., 2017; Menberg et al., 2019). To normalise a vector $\mathbf{v}'$ to be in the range [0, 1], the following equation was used (Chong and Menberg, 2018):

$$\mathbf{v} = \frac{\mathbf{v}' - \min(\mathbf{v}')}{\max(\mathbf{v}') - \min(\mathbf{v}')}. \tag{7.1}$$

To standardise a vector $\mathbf{v}'$ to have a mean of 0 and variance of 1, the following equation was used (Chong and Menberg, 2018):

$$\mathbf{v} = \frac{\mathbf{v}' - \text{mean}(\mathbf{v}')}{\text{sd}(\mathbf{v}')}. \tag{7.2}$$

7.1.2 Statistical Framework

The calibration approach employed within this work relies on the framework introduced by Kennedy and O'Hagan (2001). The adaptation of this framework for the problem of calibrating archetype-based building stock models of indoor temperature was inspired by the work of Booth et al. (2012) and Kristensen et al. (2017a). 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 et al. (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, as per the Cerezo et al. (2017) approach. It includes a model discrepancy term which could potentially reveal shortcomings of

UK-HSM. It also allows for the straightforward specification of non-normal priors for the calibration parameters, and the findings from Chapter 5 suggest that many such parameters are best described by non-normal distributions.

Given the large computational cost of UK-HSM (EnergyPlus) simulations, a Gaussian process (GP) is trained as a surrogate model on EnergyPlus simulations ($\mathbf{y}_c^{(S)}$) and monitored data ($\mathbf{y}_c^{(M)}$), as suggested by 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). If within the cluster there are M monitored dwellings, the total number of monitored data points used for the calibration are $N_c^{(M)} = M \times D_c$. Similarly, if S computer simulations were run to train the emulator, the total number of simulated data points used for calibration were $N_c^{(S)} = S \times D_c$. With regard to the statistical formulation, what differentiates each day is a set of weather variables. Day 1 ($d = 1$), is associated with weather variables $\mathbf{w}_1$, day 2 ($d = 2$) is associated with weather variables $\mathbf{w}_2$ and so on. What differentiates dwellings on the same day is the set of explanatory variables, which in this case only includes the floor area as other explanatory variables (e.g. window to wall ratio) were not accurately known. Thus, monitored dwelling $m = 1$ is associated with explanatory variable $x_{m=1}^{(M)}$, while simulated dwelling $s = 1$ is associated with explanatory variable $x_{s=1}^{(S)}$. 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, as explained in Section 7.1.1.2.

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

$$y_{md}^{(M)} = y(\mathbf{x}_m^{(M)}, \mathbf{w}_d) = \eta(\mathbf{x}_m^{(M)}, \mathbf{w}_d, \boldsymbol{\theta}) + \delta(\mathbf{x}_m^{(M)}, \mathbf{w}_d) + \varepsilon_{md}^{(M)}, \tag{7.3}$$

where:

- $y_{md}^{(M)}$ is the MDLRT for monitored home m on day d,
- $\eta(\cdot)$ is the surrogate model represented by a Gaussian process,
- $\delta(\cdot)$ is the discrepancy term (or model bias) represented by a Gaussian process,

- $\mathbf{w}_d$ 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
- $\varepsilon_{md}^{(M)}$ is the associated error term.

The error term, $\varepsilon_{md}^{(M)}$, 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 (Kennedy and O'Hagan, 2001); this might include stochastic occupant behaviour and violations of the cluster homogeneity assumption (Kristensen et al., 2017a). This is assumed to be normally distributed, with a mean of zero and a variance of $1/\lambda_\varepsilon$, as per Chong and Menberg (2018). For the $N_c^{(S)}$ model data points, the following statistical relationship was defined:

$$y_{sd}^{(S)} = y(\mathbf{x}_s^{(S)}, \mathbf{w}_d, \mathbf{t}_s) = \eta(\mathbf{x}_s^{(S)}, \mathbf{w}_d, \mathbf{t}_s) + \varepsilon_{sd}^{(S)}, \tag{7.4}$$

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 used in the UK-HSM simulations and
- $\varepsilon_{sd}^{(S)}$ is a simulation error (or noise) term.

The simulation error term $\varepsilon_{sd}^{(S)}$ 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 allowed the same set of UK-HSM simulations to be used in the parametric analysis discussed in Section 7.1.4, 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 7.3–7.4, 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 simula-

tion data (of length $N_c^{(M)} + N_c^{(S)}$) was constructed $\mathbf{z} = [\mathbf{y}_c^{(M)}, \mathbf{y}_c^{(S)}]$[1]. 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):

$$\mathcal{L}(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\xi}) \propto |\mathbf{K}_z|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T \mathbf{K}_z^{-1}(\mathbf{z}-\boldsymbol{\mu})\right\}, \tag{7.5}$$

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 Chapter E.1 of the appendices for further information). $|\mathbf{K}_z|$ and $\mathbf{K}_z^{-1}$ represent the determinant and inverse of the combined covariance matrix, respectively.

The posteriors were computed using the Markov Chain Monte Carlo (MCMC) approach, introduced in Section 2.3.8, which is most frequently used in Bayesian calibration studies (Hou et al., 2021). A short summary of the MCMC theory is provided in Section D.3. The No-U-Turn Sampler (NUTS) algorithm, an extension of Hamiltonian Monte Carlo, was used due to its superior performance compared to other commonly used MCMC algorithms (Chong et al., 2017).

7.1.3 The Choice of Priors

In Chapter 5, a probability distribution was identified for each model input based on the best available evidence. These may be used as the priors of the calibration parameters. However, since the calibration parameters were standardised to be in the interval [0, 1], the prior distributions had to be reparametrised to be on the same scale. This is trivial for a normal or uniform distribution, where a linear transformation can be analytically performed, but this is not the case for many other distributions. To transform such distributions, the following steps were taken:

1. A large number of samples is drawn from the probability distribution of each model input. In this case, a sample size of 100,000 was considered adequate.

[1]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 (Section 7.1.1.4).

2. The min-max transformation (equation 7.1) is applied to the sampled data, using the same minimum and maximum values as those used to standardise the model input data. This may result in some transformed priors extending below zero and above one, but places them on the same scale as the standardised model data.

3. In case the distribution is zero-bound, data points below zero are removed. This enables distributions to be fitted which do not have support for a value less or equal to zero.

4. The same probability distribution (e.g. Weibull($shape$, $scale$)) as the one used to generate the data is fitted to the standardised data, and a new set of distribution parameters are identified (Weibull($shape_{std}$, $scale_{std}$)).

While Steps 3–4 of the prior transformation process may introduce some bias, this is believed to be small. The standardised distributions for the four calibration parameters are listed in Table 7.2.

Table 7.2: Prior distributions used in the Bayesian calibration.

Parameter	Distribution
Glazing Fraction	gamma(shape = 3.43, rate = 8.58)
Orientation	unif(min = 0, max = 1)
Window Opening Threshold	logis(location = 0.4, scale = 0.13)
Electrical Gains Factor	gamma(shape = 2.35, rate = 7.18)

The choice of hyperparameter priors used within the Bayesian calibration is also important for the calibration process. The hyperparameter priors were defined based on the suggestions of Chong and Menberg (2018) and Menberg et al. (2019). For the parameters that define the two error terms, the following priors were used:

- $\lambda_\varepsilon \sim Gamma(shape = 10, rate = 0.03)$: This prior represents the belief that the influence of measurement errors (and the overall variance not explained by calibrated model and model bias) will be approximately 0.3 %. This is because the observations were standardised to have a variance of 1, and with an expectation value for λ_ε of approximately 333, the variance of ε_{md} (σ^2_{md}) is ≈ 0.003 (since the variance of ε_{md} is equal to $1/\lambda_\varepsilon$).

- $\lambda_{sim} \sim Gamma(shape = 10, rate = 0.001)$: With an expectation value for λ_{sim} of approximately 10,000, contribution of ε_{sd} to the simulation variance is expected to be small, roughly 0.01 %.

Other hyperparameter priors are summarised in Section E.2 of the appendices.

7.1.4 The Choice of Variables: Parametric Calibration

The calibration process assumes that the error terms, ε_{md}, are *independently and identically distributed* (*iid*); they come from the same distribution and are mutually independent (Smith, 2013). Intuitively, the daily mean indoor temperatures in a

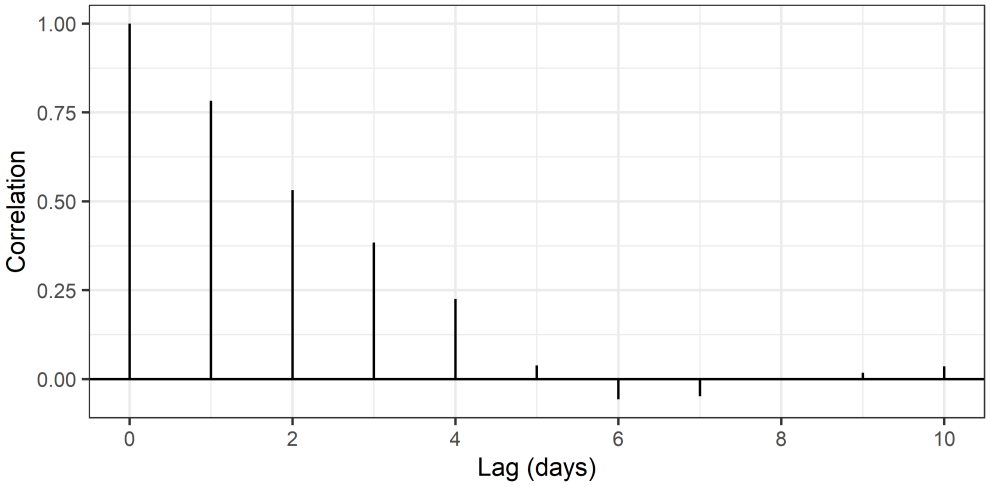


Figure 7.3: AutoCorrelation Function (ACF) plot of the daily-mean living room temperature, mean-averaged for all semi-detached dwellings of the homogeneous cluster.

free-floating building are not expected to be independent; the indoor temperature observed on day d ($y^{(M)}_{m,d}$) is more likely to be similar to the value of the previous day $y^{(M)}_{m,d-1}$, than that of the previous week $y^{(M)}_{m,d-7}$. This is demonstrated through the AutoCorrelation Function (ACF) plot in Figure 7.3, where the monitored indoor temperatures for the homogeneous cluster are autocorrelated for up to a lag of four days. For the purposes of statistical modelling and calibration, while $y^{(M)}_{md}$ (and $\varepsilon^{(M)}_{md}$) are not independent variables, they can be conditionally independent given the right selection of predictors $\mathbf{x}^{(M)}_m$ and $\mathbf{w}_d$. The explanatory parameters ($\mathbf{x}^{(M)}_m$) do not have an effect on the autocorrelation observed in Figure 7.3 since they do not vary between days. Therefore, it is the choice of weather variables that is crucial

to ensure that $\varepsilon_{md}^{(M)}$ are conditionally independent. To address this, in addition to the use of daily mean outdoor temperature and global horizontal irradiance, the use of lag components 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's and surrounding environment's thermal mass. For day d, with indoor temperature $y_{m,d}^{(M)}$ and associated weather variable $\mathbf{w}_d$, the one-day lag components are the weather observations of the previous day ($\mathbf{w}_{d-1}$), the two-day lag components are the weather observations of two days before ($\mathbf{w}_{d-2}$), and so on.

Table 7.3: Summary of weather, explanatory and calibration variable combinations assessed as part of the parametric calibration analysis.

Experiment	Weather	Explanatory	Calibration
EXP1	OT, GHI	FAF	WOT, Orien., GF, EGF
EXP2	OT, GHI	FAF	WOT, Orien., GF
EXP3	OT, GHI	FAF	WOT, Orien., 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, Orien.
EXP8	OT, GHI	FAF	WOT
EXP1L1	OT, GHI, OTL1	FAF	WOT, Orien., GF, EGF
EXP2L1	OT, GHI, OTL1	FAF	WOT, Orien., GF
EXP3L1	OT, GHI, OTL1	FAF	WOT, Orien., 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, Orien.
EXP8L1	OT, GHI, OTL1	FAF	WOT
EXP1L2	OT, GHI, OTL1, OTL2	FAF	WOT, Orien., GF, EGF
EXP2L2	OT, GHI, OTL1, OTL2	FAF	WOT, Orien., GF
EXP3L2	OT, GHI, OTL1, OTL2	FAF	WOT, Orien., 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, Orien.
EXP8L2	OT, GHI, OTL1, OTL2	FAF	WOT

Disambiguation: OT = Outdoor Temperature; GHI = Global Horizontal Irradiance, OTL1 = Outdoor Temperature with Lag of 1 day; OTL2 = Outdoor Temperature with Lag of 2 days; GF = Glazing Fraction; Orien. = Orientation; FAF = Floor Area Factor; WOT = Window Opening Threshold; EGF = Electrical Gains Factor.

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. For the same set of simulated and monitored data described in Section 7.1.1.2, the calibration process was carried out 24 times for a combination of weather and calibration parameters, as summarised in Table 7.3. The decision to only include lag components of the outdoor temperature, and not of the GHI, was informed by the exploratory analysis presented in Section 7.2.1. Due to its dominance during the sensitivity analysis, the Window Opening Threshold was included in all calibration runs. The calibrations were each run for 500 MCMC iterations.

7.1.5 Training & Validation

While LHS sampling allows an effective exploration of the input space for the calibration and explanatory variables, the same method could not be applied for the weather variables. The weather variables are based on the empirical data collected during July and August 2009 (see Section 3.4.4), and sampling in a way that provides effective coverage for one of them (e.g. Outdoor Temperature) may not necessarily result in an equally effective coverage for the other (e.g. Global Horizontal Irradiance). Performing the calibration during any specific period may result in fairly good predictive performance for an unseen period with similar weather conditions, but a poor performance if the weather conditions are different. This was apparent in the initial stages of this calibration work, where the calibration over the second week of July resulted in excellent predictions for the third week of the same month (due to their similarity in weather conditions), but relatively poor for other periods with less similar weather conditions.

To establish the model's performance, the 62-day period when field data was available was split into a 10-day (16.1 %) training period and a 52-day (83.9 %) validation period. The choice of a relatively short training period allowed for a

manageable computational cost and the undertaking of the calibration experiment outlined in the previous section. The 10 days of the training period were selected with the aim to cover as wide a range of Outdoor Temperature as possible. This was due to Outdoor Temperature being identified as likely the most influential weather variable in the exploratory analysis summarised in Section 7.2. The coverage of the other weather variables was not necessarily as good.

Following the calibration, the indoor temperature for the $D_v = 52$ unseen days of the validation period were predicted. To incorporate the uncertainty represented by the posterior distributions, $L = 500$ posterior samples were drawn and used for the prediction of the unseen period, resulting in a matrix of outputs:

$$\mathbf{Y}_v^{(P)} = \begin{bmatrix} y_{1,1}^{(P)} & y_{1,2}^{(P)} & \cdots & y_{1,D_v}^{(P)} \\ y_{2,1}^{(P)} & y_{2,2}^{(P)} & \cdots & y_{2,D_v}^{(P)} \\ \vdots & \vdots & \vdots & \vdots \\ y_{L,1}^{(P)} & y_{L,2}^{(P)} & \cdots & y_{L,D_v}^{(P)} \end{bmatrix}, \tag{7.6}$$

where $\mathbf{Y}_v^{(P)}$ is an $L \times D_v$ matrix of posterior predictions. Each row of the matrix is associated with a posterior sample, each column with a day. Thus, $y_{1,1}^{(P)}$ is the MDLRT predicted for day $d = 1$ from sample $l = 1$. For validation purposes, the predictions were mean-averaged 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 of validation metrics described in Section 7.1.5.1. The monitored data were also compared against the daily mean values of computer simulations used to train the surrogate model, providing a baseline for the improvement in predictive performance following the calibration.

7.1.5.1 Validation Statistics

To quantify the differences between predictions and monitored data, a combination of metrics commonly used for the empirical validation of building models were employed (Ruiz and Bandera, 2017). The normalised mean bias error (NMBE), is

an indicator of the overall behaviour of the model and is defined as:

$$\text{NMBE} = \frac{1}{\overline{y_v^{(M)}}} \cdot \frac{1}{D_v} \sum_{d=1}^{D_v} (\overline{y_d^{(M)}} - \overline{y_d^{(P)}}) \times 100\% \tag{7.7}$$

where $\overline{y_v^{(M)}}$ is the mean of $\overline{\mathbf{y}_v^{(M)}}$. NMBE is an estimate of the normalised mean of errors. It is subject to cancellation errors, and its sign indicates whether the model under or over-predicts compared to the monitored data. Together with NMBE, the Coefficient of Variation of the Root Mean Square Error (CV(RMSE)) is also commonly used as a measure of the errors' variability within the building modelling community, and is defined as:

$$\text{CV(RMSE)} = \frac{1}{\overline{y_v^{(M)}}} \sqrt{\frac{\sum_{d=1}^{D_v} (\overline{y_d^{(M)}} - \overline{y_d^{(P)}})^2}{D_v - 1}} \times 100\%. \tag{7.8}$$

For both the NMBE and CV(RMSE), the closer their value to zero, the better. If NMBE and CV(RMSE) are assumed to be of equal importance, a way of characterising the model's performance according to both metrics is the Goodness-of-Fit estimate (Ruiz and Bandera, 2017):

$$\text{GOF} = \frac{\sqrt{2}}{2} \sqrt{\text{CV(RMSE)}^2 + \text{NMBE}^2}. \tag{7.9}$$

Another commonly used metric that represents agreement in the patterns of the monitored and simulated data is the coefficient of determination (R^2):

$$R^2 = \left(\frac{D_v \cdot \sum_{d=1}^{D_v} \overline{y_d^{(M)}} \cdot \overline{y_d^{(P)}} - \sum_{d=1}^{D_v} \overline{y_d^{(M)}} \cdot \sum_{d=1}^{D_v} \overline{y_d^{(P)}}}{\sqrt{(D_v \cdot \sum_{d=1}^{D_v} (\overline{y_d^{(M)}})^2 - (\sum_{d=1}^{D_v} \overline{y_d^{(M)}})^2) \cdot (D_v \cdot \sum_{d=1}^{D_v} (\overline{y_d^{(P)}})^2 - (\sum_{d=1}^{D_v} \overline{y_d^{(P)}})^2)}} \right)^2. \tag{7.10}$$

R^2 ranges between 0 and 1, with a value of 1 suggesting a perfect agreement. While all aforementioned validation metrics quantify the performance in relative terms, a metric in the same unit as the model output can enhance the interpretation of results.

One such metric is the RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{D_v}\sum_{d=1}^{D_v}(\overline{y_d^{(M)}} - \overline{y_d^{(P)}})^2}. \tag{7.11}$$

7.1.6 Kronecker Product

A well-known limitation of Gaussian processes is their computational cost which rapidly increases with the number of data points. This large computational burden is the result of having to invert large covariance matrices – one of the two key components of a Gaussian process – potentially thousands of times.[2] It is sometimes possible to take advantage of the data's regular structure, if there is one, to reduce the computational cost of using a Gaussian process.

For the calibration application described in this chapter, the matrix of weather data ($\mathbf{W}_c = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D_c}]$) is the same across the simulated and monitored buildings, resulting in a regular structure within the input space. Due to this structure, one could define the covariance matrix as the Kronecker product of two smaller covariance matrices (see Appendix Section F.1 and Pollock (2013) for an explanation of the Kronecker product). Doing so may result in a reduction in computational cost, because it is computationally faster to invert two small matrices than a single big one. An in-depth explanation of this process can be found in Section F.2 of the appendices.

A small number of examples of Bayesian calibration that have employed this alternative formulation exist, but not within the field of building modelling (Bayarri et al., 2009; Bilionis et al., 2013; Hung et al., 2015; Williams et al., 2006). To investigate the potential benefits and drawback of this method in the Bayesian calibration of archetype-based housing stock models of summer indoor temperature, the Kronecker formulation outlined by Bayarri et al. (2009) was compared against the "traditional" approach described by Chong and Menberg (2018). The programmatic specifications of the "Kronecker" method were informed by Flaxman et al. (2015).

[2]Please see Appendix Section A for more information on covariance matrices. The inversion may need to take place thousands of times for the Markov Chain Monte Carlo process to converge.

7.2 Results

7.2.1 Exploratory Analysis

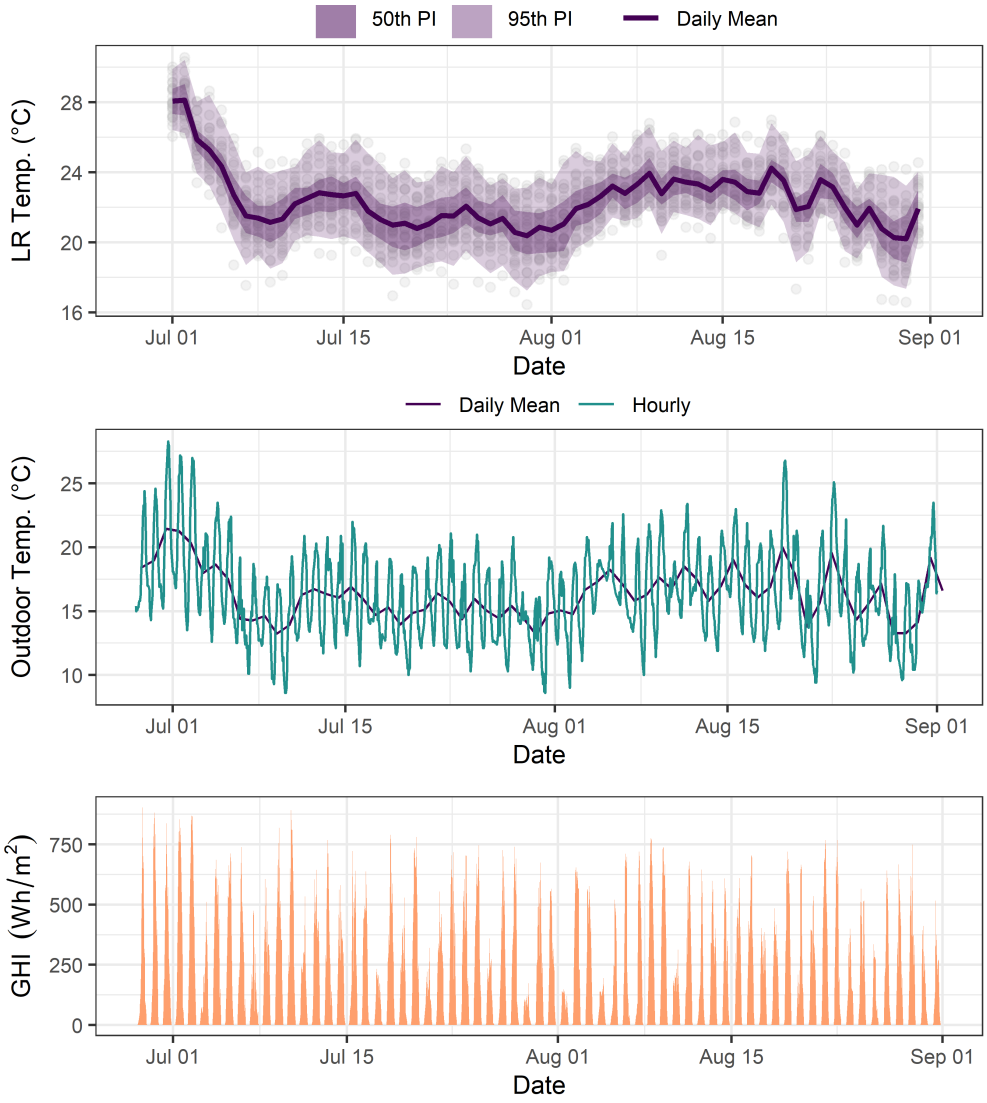


Figure 7.4: Timeseries plot of the daily monitored living room temperatures (LR Temp.) within the homogeneous group, the outdoor temperature and the global horizontal irradiance (GHI).

A timeseries plot of the MDLRT is presented in Figure 7.4. The outdoor temperature and global horizontal irradiance (GHI) are plotted in the same figure. The warmest period was observed during the first week of July, when the highest indoor temperature (group mean of 28.8 °C) was also recorded. That same week was also associated with high outdoor temperature and GHI levels. While there were

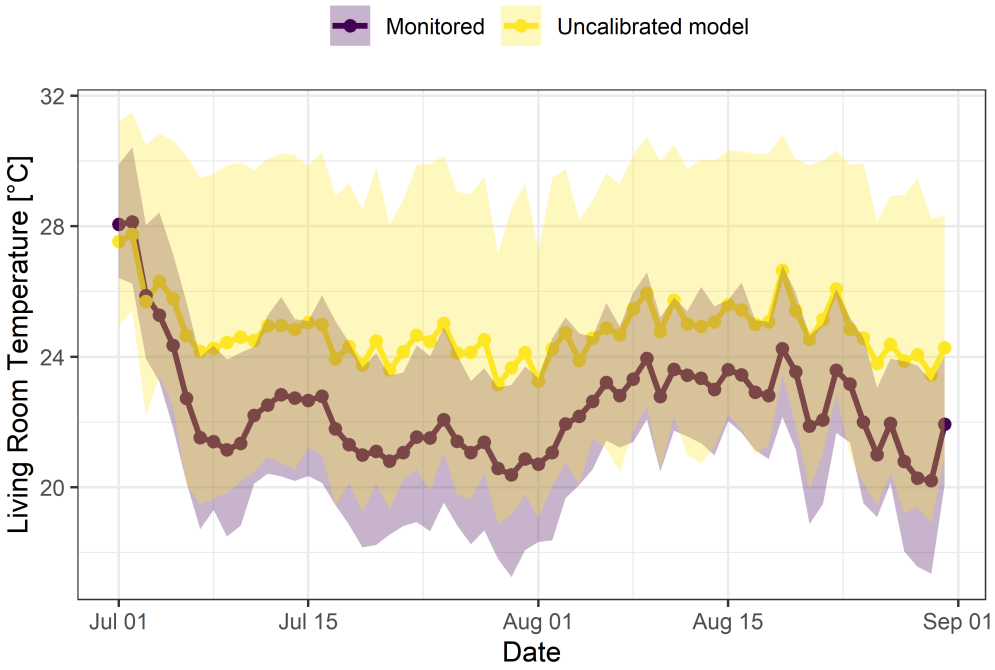


Figure 7.5: Timeseries plot of the daily-mean living room temperature for the monitored homes and uncalibrated computer simulations. The central line represents the mean-average across monitored homes and simulations, while the shaded area indicates the corresponding 95th percentile interval.

a few other days with high outdoor temperature, for example the 19th or 23rd of August, they were isolated incidents resulting in local peaks in indoor temperature that did not reach the levels of the first week of July.

A comparison of monitored (field) data and uncalibrated model simulations of living room temperatures is shown in Figure 7.5. The mean of the computer simulation predictions is consistently higher than that of the monitored data, except for the first week of July, where the monitored indoor temperatures were on average warmer than predicted. While a discrepancy between monitored and simulated data is clearly visible, the patterns of daily variation appear similar. At this stage, it is not possible to determine to what extent the differences between monitored and simulated data are due to model inadequacy or the incorrect specification of model inputs.

Figure 7.6 provides a matrix of scatterplots for a selection of weather and explanatory variables. The top row of plots visualises the relationship of MDLRT

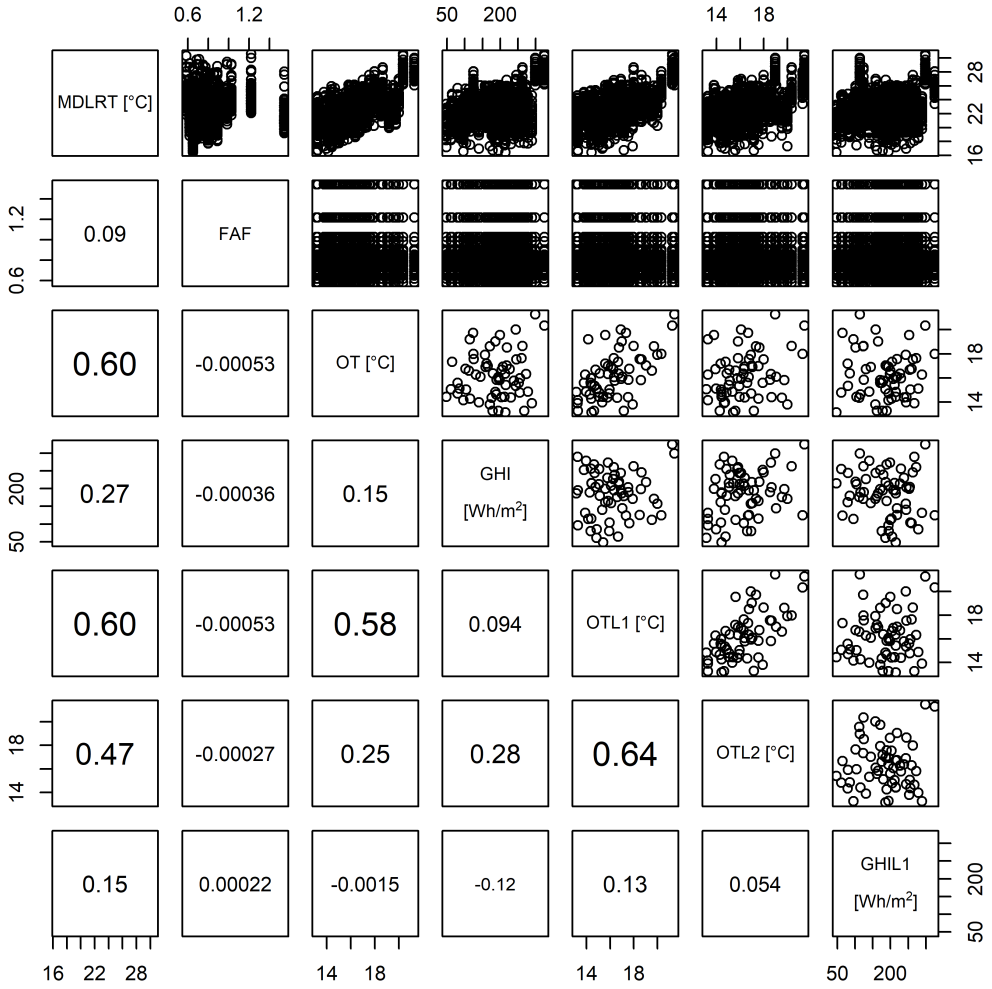


Figure 7.6: Scatterplots of the mean daytime living room temperature (MDLRT) monitored in 26 semi-detached dwellings over 62 days, the associated floor area factor (FAF), daily mean global horizontal irradiance (GHI), daily mean outdoor temperature (OT) and one and two day lag components (L1-2). The lower left panel provides the Pearson correlation coefficients. The diagonal offers the axes labels for each plot. Scatterplots of only weather variables have fewer points than those that include a dwelling variable (MDLRT and FAF) since weather variables do not vary between dwellings.

with each weather and explanatory variable. For all plots, there is large variance that cannot be explained by any single variable. A fairly strong positive relationship is observed between MDLRT and the daily mean outdoor temperature (OT), with a Pearson correlation coefficient (r) of 0.6. Stronger correlations (r = 0.91-0.95) have been reported in studies from other countries, possibly due to differences in dwelling characteristics and the use of the daily (24-hour) mean indoor temperature instead of MDLRT (Lee and Lee, 2015; Nguyen et al., 2014). The same level of linear

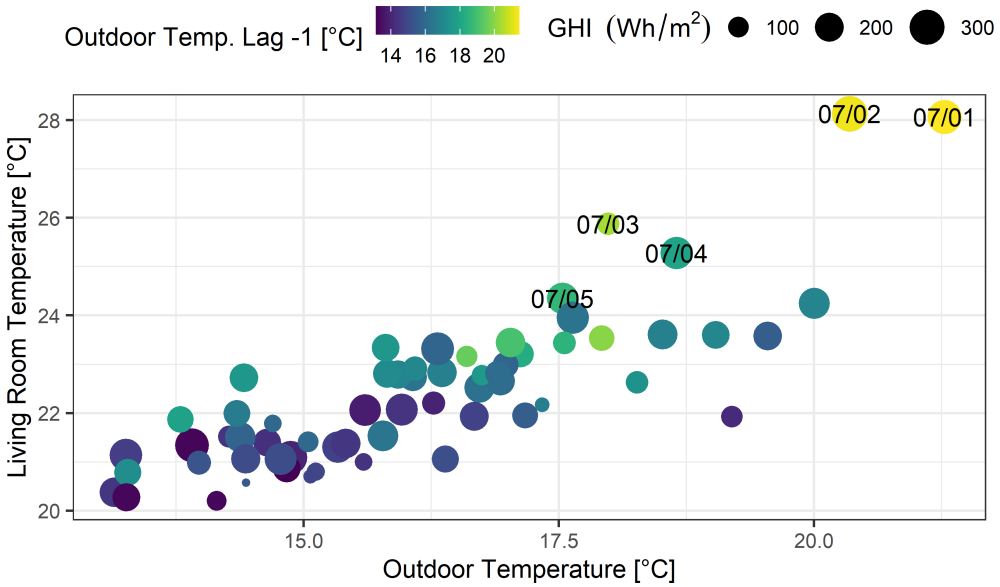


Figure 7.7: Scatterplot of the mean daytime living room temperature, averaged across the dwellings of the homogeneous cluster, against the daily-mean outdoor temperature. The colour and size of each marker is associated with the outdoor temperature lag component and global horizontal irradiance (GHI), respectively.

correlation as with OT is observed between MDLRT and the one-day lag component of the outdoor temperature (OTL1), with a slightly weaker correlation ($r = 0.47$) between the MDLRT and the two-day lag component (OTL2). As might be expected, the outdoor temperature components are themselves correlated, with a level of $r = 0.58$ for one-day lag and $r = 0.25$ for a two-day lag. The association of MDLRT with global horizontal irradiance (GHI) is lower than any outdoor temperature component included in Figure 7.6 ($r = 0.27$), with an even weaker relationship between the lag component of GHI (GHIL1) and MDLRT ($r = 0.15$). The scatterplot of the MDLRT against floor area factor shows no clear trend, with a very weak and positive correlation coefficient of $r = 0.09$. This goes against the expectation of a negative association suggested in Section 4.2. This suggests that the floor area has a relatively small effect on MDLRT, in comparison to other variables and within this homogeneous cluster of dwellings. Due to weak correlation between MDLRT and GHIL1, and to limit the computational cost of carrying out the calibration experiment, GHIL1 was not considered in the calibration

The moderately strong association between the lag components of the outdoor

temperature and the MDLRT supports their inclusion in the calibration process. However, it is not clear whether this is an artefact of the similarly strong correlation between the weather variables themselves. To explore this further, the mean MDLRT – averaged across the group of dwellings – was plotted against the daily mean outdoor temperature in Figure 7.7. In this plot, the colour and size of each marker also varies depending on the first lag component of the outdoor temperature and the GHI. As with Figure 7.6, there is a trend of increasing indoor temperature as the outdoor temperature increases. The two days with the highest indoor temperature (07/01 and 07/02) were associated with the highest recorded outdoor temperatures and high levels of GHI. The day with the third warmest indoor temperature (07/03) is associated with the 10th highest outdoor temperature and relatively low levels of GHI ($119\,\mathrm{Wh/m^2}$). However, the lag-component of the outdoor temperature associated with that day is one of the highest recorded, thus potentially providing an explanation of the high indoor temperature. Further evidence in support of the effect that lag components might have is provided by examining the conditions on 19th of August. On this day, the daily mean outdoor temperature ($20\,^\circ\mathrm{C}$) was higher than on the 3rd ($18\,^\circ\mathrm{C}$) and 4th ($18.7\,^\circ\mathrm{C}$) of July, while the GHI ($234\,\mathrm{Wh/m^2}$) was comparable to that on the 4th of July ($252\,\mathrm{Wh/m^2}$) and higher than on the 3rd of July ($119\,\mathrm{Wh/m^2}$). Yet, the mean MDLRT was lower on the 19th of August ($24.2\,^\circ\mathrm{C}$) than on the 3rd ($25.9\,^\circ\mathrm{C}$) or 4th ($25.3\,^\circ\mathrm{C}$) of July. This is likely associated with weather conditions on the previous days. The one-day lag component on the 19th of August ($16.9\,^\circ\mathrm{C}$) was lower than on either of the other two days ($20.4\,^\circ\mathrm{C}$ on the 3rd and $18.0\,^\circ\mathrm{C}$ on the 4th of July). Overall, the five days with the highest recorded indoor temperatures were the first five days of July.

7.2.2 Parametric Analysis

In the parametric analysis, the daily mean MDLRT ($\overline{\mathbf{y}_{\mathbf{v}}^{(\mathbf{M})}}$) was compared against the uncalibrated and bias-corrected calibrated model ($\eta(x,w,t)+\delta(x,w)$) predictions, as described in Section 7.1.5. The results of the parametric analysis are summarised in Table 7.4.

For the first set of parametric calibration experiments (EXP1–8), the outdoor

Table 7.4: Summary of out-of-sample validation metrics calculated over a 52-day period for the parametric calibration experiments. Time refers to the calibration time, hence a value was not provided for the uncalibrated model. Experiments EXP3, EXP1L1, EXP3L1, EXP7L1, EXP1L2 and EXP7L2 did not converge. Bold font indicates the best performing models.

Exp	CV(RMSE) [%]	NMBE [%]	GOF [%]	RMSE [°C]	R^2	Time [hrs]
Uncalib.	11.51	-11.20	11.36	2.53	0.79	-
EXP1	4.47	-0.73	3.20	0.98	0.43	1.77
EXP2	4.53	-0.75	3.25	1.00	0.41	1.02
EXP3	-	-	-	-	-	-
EXP4	4.55	-0.68	3.25	1.00	0.41	1.14
EXP5	4.39	-0.69	3.14	0.96	0.43	0.99
EXP6	4.36	-0.75	3.13	0.96	0.44	1.20
EXP7	4.34	-0.77	3.12	0.95	0.44	1.32
EXP8	4.33	-0.76	3.11	0.95	0.45	0.73
EXP1L1	-	-	-	-	-	-
EXP2L1	2.93	0.22	2.08	0.64	0.73	1.35
EXP3L1	-	-	-	-	-	-
EXP4L1	2.97	0.20	2.10	0.65	0.73	1.25
EXP5L1	3.17	0.22	2.25	0.70	0.70	1.36
EXP6L1	2.90	0.12	2.05	0.64	0.74	1.29
EXP7L1	-	-	-	-	-	-
EXP8L1	2.97	0.12	2.10	0.65	0.72	1.20
EXP1L2	-	-	-	-	-	-
EXP2L2	2.70	-0.08	1.91	0.59	0.77	1.53
EXP3L2	**2.65**	**-0.20**	**1.88**	**0.58**	**0.77**	**1.95**
EXP4L2	2.74	-0.21	1.95	0.60	0.76	2.13
EXP5L2	2.92	-0.20	2.07	0.64	0.74	1.68
EXP6L2	**2.67**	**-0.22**	**1.89**	**0.59**	**0.77**	**1.67**
EXP7L2	-	-	-	-	-	-
EXP8L2	2.71	-0.19	1.92	0.60	0.77	1.66

Disambiguation: RMSE = Root-mean-square error, CV(RMSE) = coefficient of variation of RMSE, NMBE = normalised mean bias error, GOF = goodness of fit and R^2 = coefficient of determination.

temperature and GHI were the only weather variables used. Except for EXP3, all other experiments converged within 500 MCMC iterations. The predictive performance improved following calibration according to CV(RMSE), NMBE, GOF and RMSE, but dropped according to R^2. Specifically, CV(RMSE) and NMBE reduced from 11.5 % and -11.2 % for the uncalibrated EnergyPlus model to 4.3–4.5 % and 0.1–0.2 %, respectively. However, R^2 reduced from 0.79 to 0.41–0.45.

Therefore, while all calibrated models are able to make more accurate predictions than the uncalibrated model, their ability to represent day-to-day fluctuations of indoor temperatures is worse.

In the second set of calibration experiments, following the addition of a one-day lag component of the outdoor temperature, five out of the eight experiments converged within 500 iterations. Looking at the out-of-sample prediction of experiments that converged (EXP2L1, EXP4L1, EXP5L1, EXP6L1, EXP8L1), the performance is comparable across all five experiments and better than of the first set of experiments (EXP1–8). In all cases, the CV(RMSE) and NMBE reduced from 11.5 % and -11.2 % for the uncalibrated EnergyPlus model to about 3 % and 0.1–0.2 %, respectively. The value of NMBE close to zero following the calibration suggests that the calibrated models do not, on average, over-predict the daily indoor temperature contrary to the EnergyPlus model. In terms of degrees Celsius, the RMSE improved from 2.53 °C to 0.64–0.70 °C. The calibrated model's performance as assessed by R^2 is lower (0.70–0.74) than that of the uncalibrated model (0.79), but higher than for the first set of calibrations (0.41–0.45). By using GOF, where CV(RMSE) and NMBE are combined, together with R^2 to rank the models, the best performing model is EXP6L1 (GOF = 2.05 %, R^2 = 0.74), where the only two calibration variables were the Window Opening Threshold and the Electrical Gains Factor.

The addition of a second outdoor temperature lag component (EXP1L2–EXP8L2) resulted in further improvement in predictive performance across most metrics. In general, CV(RMSE) is slightly lower for the second set (2.65–2.92 % compared to 2.79–3.11 %), while the magnitude of NMBE is comparable between the two sets of parametric experiments. Since the GOF values and RMSE are lower for this set of experiments, while the R^2 is higher (0.74–0.76 compared to 0.70–0.74), the addition of a second lag component has resulted in the best overall out-of-sample prediction. The best performing model is EXP3L2, with a marginally lower GOF (1.88 %) than EXP6L2 (1.89 %). Two experiments did not converge within 500 iterations.

A common characteristic amongst all experiments that did not converge (EXP3, EXP1L1, EXP3L1, EXP7L1, EXP1L2, EXP7L2) was the use of Orientation and Window Opening Threshold (WOT) as calibration parameters. Given that other calibration experiments that included WOT did converge, including EXP8, EXP8L1 and EXP8L2 where WOT was the only calibration parameter, it is suspected that the Orientation is responsible for the lack of convergence. This could be due to Orientation not being adequately described by a unimodal distribution, resulting in poor MCMC sampling. While it is not unlikely that there are more than one orientations that are frequently occurring in the group of monitored dwellings, whether this is the reason for poor convergence is unclear.

From Table 7.4 it may also be observed that the calibration time varied between experiments, generally increased as more weather variables were added, but did not always decrease as the number of calibration parameters reduced.

The model EXP3L2 included Window Opening Threshold, Orientation and Electrical gains factor as calibration parameters, while EXP6L2 did not include orientation. 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.

7.2.3 Detailed Analysis

Figure 7.8 presents a visual comparison of the mean MDLRT (averaged across the cluster's dwellings), uncalibrated model (averaged across simulations), bias-corrected calibrated model ($\eta(x,w,t)+\delta(x,w)$) and calibrated model without bias correction ($\eta(x,t)$) for experiment EXP6L2. The comparison is for the 52 days of the validation period. For the calibrated model predictions, the central line represents the mean of the posterior realisations, used to summarise the results since they are approximately normally distributed. The shaded region captures uncertainty around the mean estimate, equivalent to ± 1.96 standard deviations (or a roughly 95 % interval).

For most days of the validation period, the bias-corrected model performs best, as it has the smallest discrepancies from the monitored data. For 34 out of

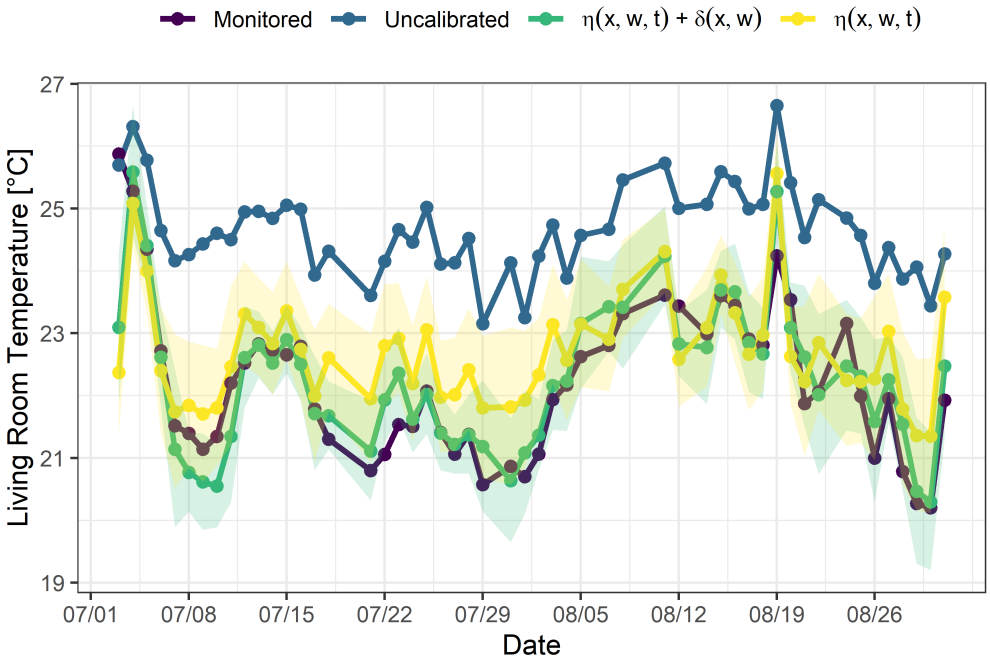


Figure 7.8: Timeseries plot of the mean daytime living room temperature for the: (i) Field data, mean-averaged across the cluster's dwellings, (ii) uncalibrated (EnergyPlus) model predictions mean-averaged across simulations, (iii) the bias-corrected calibrated model predictions ($\eta(x,w,t)+\delta(x,w)$), and (iv) the calibrated model predictions without model bias ($\eta(x,w,t)$). For the calibrated model predictions, the shaded area represents an uncertainty region of $\pm 1.96\sigma$ around the mean (central line).

the 52 days, the absolute difference between monitored data and calibrated model predictions is less than 0.5 °C, while for 18 days the differences are less than 0.2 °C. The discrepancies between the calibrated model predictions without bias-correction are overall smaller than the uncalibrated model (RMSE of 0.96 °C compared to 2.53 °C), but greater than for the bias-corrected model (RMSE = 0.59 °C)

A way of visually assessing the model performance is through the use of individual prediction error plots, similar in nature to the standardised residual plots. A form of these plots has been previously used for the analysis of Gaussian processes as emulators by Bastos and O'Hagan (2009). The modification here is that instead of comparing the emulator predictions to the simulator output, the bias-corrected model predictions were compared against the monitored data. As the error terms are expected to be normally distributed, approximately 95 % of individual (standardised)

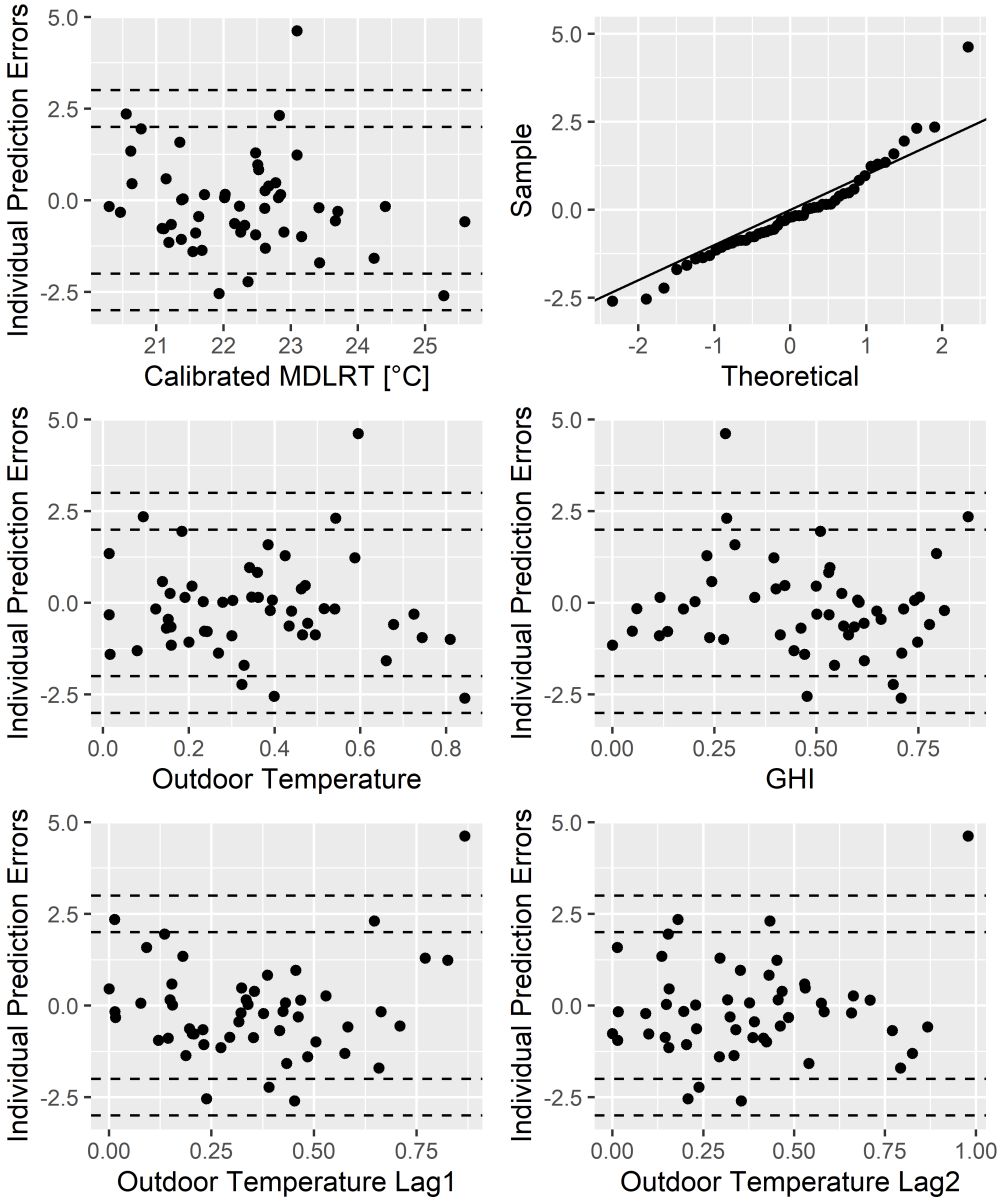


Figure 7.9: Diagnostic plots used to assess the calibrated model prediction against empirical data during the validation period. Each point represents a prediction error (or residual) for a particular validation day.

prediction errors are expected to be within the range of ± 2 and 99.7 % with a range of ± 3. In addition, for a well-performing model there should be no clear patterns and the sample and theoretical points should lie approximately on the diagonal in the QQ-plot. The scatterplots in Figure 7.9 did not reveal any clear patterns, with 90 % and 98 % of the individual prediction errors with an absolute value less than 2 and 3 respectively. In addition, most points lie close to or on the diagonal line of

the QQ-plot, suggesting that the normality assumption is reasonable. Only one data point lies outside the region of ± 3 which corresponds to the 3rd of July, when the mean calibrated prediction deviated by 2.6 °C as shown in Figure 7.8. 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), if a single extreme point exists it could be ignored, or it might indicate a local problem relating to those input values and investigated further with the addition of more data points.

Combining the analysis offered in the previous paragraph, and by visually assessing the performance of the calibrated, bias-corrected, predictions against the monitored data, it seems that 3rd of July measurement 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 that point, the R^2 would improve as a result of the calibration from 0.81 to 0.86.[3] Thus, the calibrated model is better able to represent day-to-day fluctuations, compared to the uncalibrated model, for most days.

7.2.3.1 Model Bias

For EXP6L2, the posterior predictions of the model bias, averaged across posterior predictions and the dwellings for each day, were plotted against the weather variables in Figure 7.10 to examine whether any systematic model deficiencies exist. To examine whether the model bias differed between the training and validation period, both periods were included in Figure 7.10.

A visual inspection of Figure 7.10 suggests that the behaviour of model bias is broadly consistent for the training and validation period. Hence, the rest of this analysis considers all data points together. The mean model bias ranges from -1.26 °C to 1.05 °C, suggesting that the simulator may both under- and over-predict the living room temperatures. However, with a median of -0.5 °C and 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 inputs have been calibrated.

[3] The R^2 of the uncalibrated model would also improve if this point was excluded in the validation procedure.

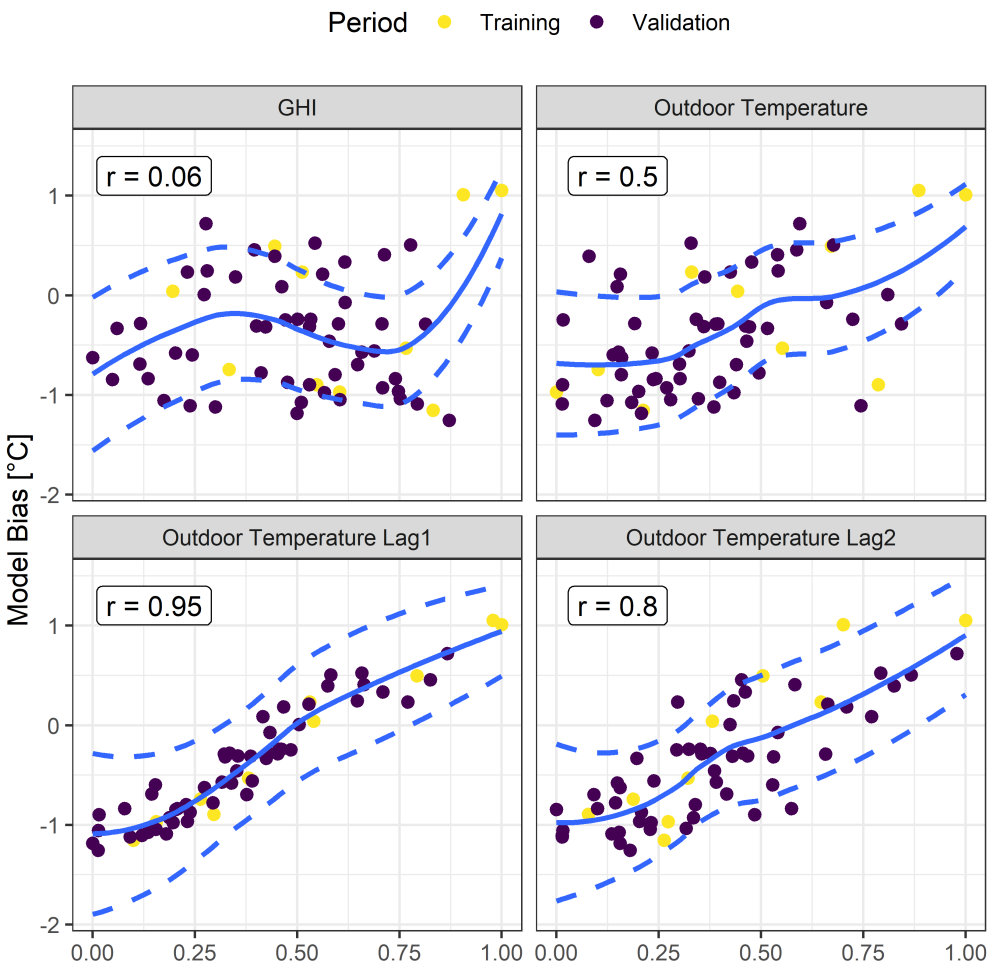


Figure 7.10: Scatterplots and lineplots of model bias, averaged across the dwellings of the homogeneous cluster, plotted against the weather variables for the training and validation period for EXP6L2. Individual points represent the model bias on different days, the central line represents the smoothed mean model bias, while the dashed lines represent an uncertainty of one standard deviation around the mean. r is the Pearson correlation coefficient.

The scatterplot of mean model bias against the global horizontal irradiance reveals no clear pattern, and the Pearson correlation coefficient is approximately zero. In addition, a zero-value in model bias is consistently within the uncertainty bound visualised by the dashed lines. On the contrary, clear patterns of association are observed between the model bias and all three outdoor temperature components. The association is strongest for the first lag component of the outdoor temperature, with an almost linear trend and a correlation coefficient of 0.95, followed by the second lag component with $r = 0.8$. An association between the Outdoor Temperature and

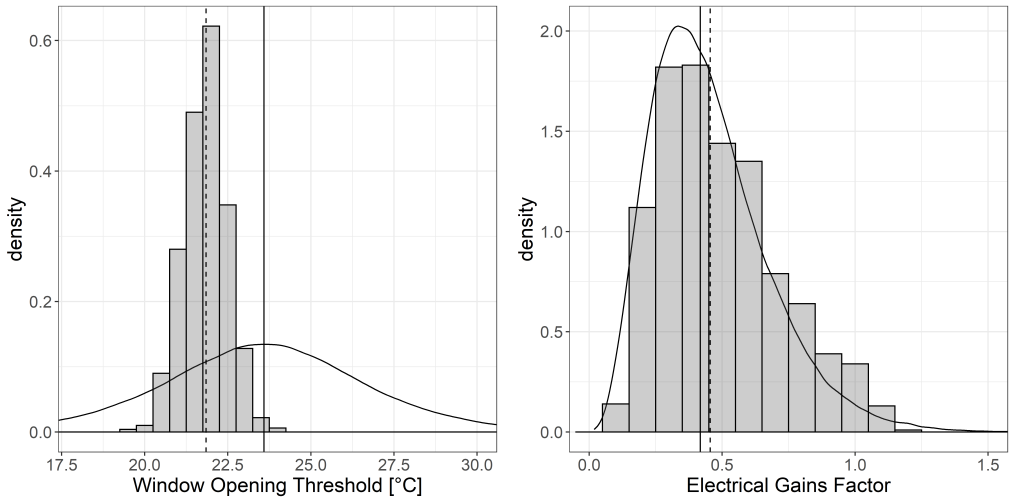
model bias may also be observed, albeit weaker with an $r = 0.5$. In addition and similarly to the GHI, a zero-value is consistently within the uncertainty bounds of this variable. Given that the outdoor temperature variables are themselves correlated ($r = 0.58$ in Figure 7.6), the association observed between model bias and outdoor temperature in Figure 7.10 might be an artefact of the relationship between model bias and the first lag component of outdoor temperature.

Based on the smoothed central line, the model bias tends to be negative when the mean outdoor temperature of the previous day is below 17.3 °C (0.50 on the standardised scale), and positive above, although uncertainties exist. The implication is that even if the model inputs were specified as the best possible values, there would be an under or over-prediction of indoor temperatures that seems to vary with the outdoor temperature on consecutive days and could impact UK-HSM's applications. Since the lag components were used to capture the effects of thermal mass, the relationship observed between lag components and model bias could suggest that the thermal mass modelling in UK-HSM may be limited. However, other possible explanations might exist and this deserves further investigations.

7.2.3.2 Parameter posteriors

Figure 7.11 compares the prior and posterior distributions of two calibration parameters, the Window Opening Threshold (WOT) and the Electrical Gains Factor (EGF). The prior is visualised as a density plot line, with its median indicated by a solid line. The posterior is represented by a histogram which consists of the MCMC draws for this parameter, and the dashed line represents the median value.

The spread in the posterior of the WOT is smaller than of its corresponding prior (Figure 7.11(a)). The posterior median is 21.8 °C, with a 90 % credible interval of 20.7–22.9 °C. A credible interval contains a specified amount of posterior probability, in this case the central 90 % probability. Therefore, the WOT value lies within 20.7–22.9 °C, with a probability of 0.9. In the case of EGF, the posterior distribution is similar to the prior, with a small shift in the median being observed (Figure 7.11(b)). This might be the result of identifiability problems, and more data could potentially resolve this. Nevertheless, the potential lack of identifiability for EGF has a relatively



(a) **(b)**

Figure 7.11: 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.

small impact on model performance, given its improved out of sample predictive ability. 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).

7.2.3.3 Hyperparameter posteriors

The prior and posterior distributions of the hyperparameters that define the error terms (see Appendix Sections 7.1.2 and E.1 for further information about these hyperparameters) are visualised in Figure 7.12. The posterior distributions are the result of the calibration experiment EXP6L2. Large shifts, in relation to their prior distribution (defined in Section 7.1.3) are observed for both terms.

The expectation value of the prior distribution of λ_{sim} was $\approx 10,000$, thus assuming that only 0.01 % of the simulation data variance would be associated with ε_{sd}. The expectation values of the posterior distribution for the same parameter is ≈ 57, resulting in approximately 1.75 % of the simulation data variance represented by ε_{sd}. A contributor to the simulation data variance is the sampling of EnergyPlus model inputs at the simulation stage that were subsequently not used for calibration(e.g. Orientation). Comparing to experiment EXP3L2, where the orientation was included as a calibration parameter, the expectation value of λ_{sim} was ≈ 303, thus resulting in

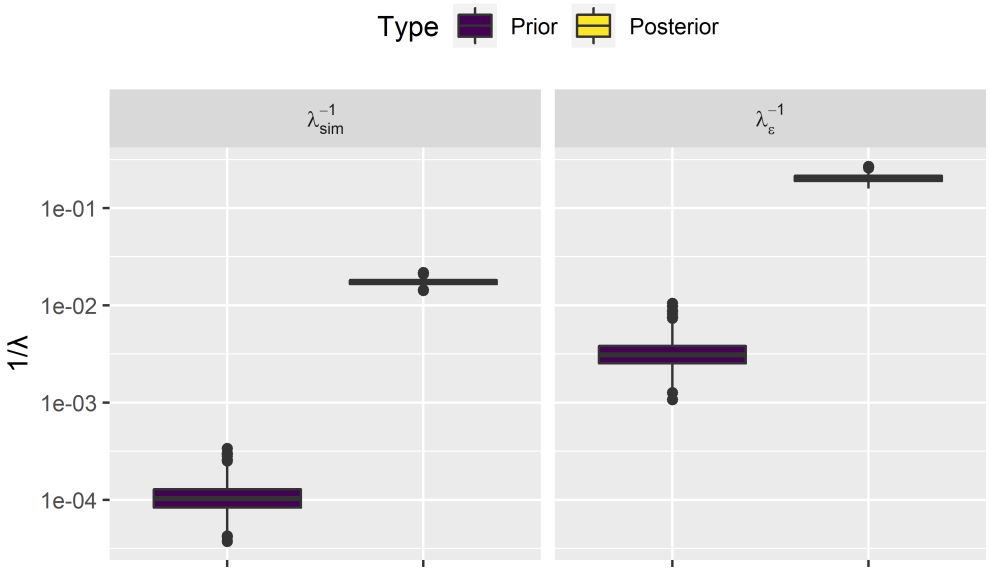


Figure 7.12: Box plots representing the prior and posterior distributions of the precision hyperparameters used to define the field and simulation data error terms.

a smaller simulation data variance (0.33 %) that could not be accounted for by the calibrated model or the model bias.

The expectation value of the prior distribution of λ_{ε} was ≈ 333, thus assuming that only 0.3 % of the monitored data variance would not be resolved by the calibrated model or the model bias. The expectation values of the posterior distribution for the same parameter is ≈ 5, resulting in approximately 20 % of the variance considered as unresolved. Factors expected to contribute to this unresolved variance include: the measurement observation errors, influential parameters not calibrated, the uncertainty of weather and explanatory variables, in addition to occupant variability and unresolved heterogeneity (Kristensen et al., 2017a).

7.2.4 Kronecker

A comparison of the out-of-sample predictive performance of the Kronecker product implementation of Bayesian calibration against the uncalibrated model and the traditional method is presented in Table 7.5. The Kronecker product method is an adaptation of the implementation proposed by Bayarri et al. (2009), while the traditional method is based on the commonly used approach described by Chong and Menberg (2018). The out-of-sample predictive performance following the Kronecker

calibration improved across all metrics, except for R^2. Specifically, while both the GOF and RMSE improved with pre- and post-calibration values of 11.36 % to 2.43 % and 2.53 °C to 0.74 °C, respectively, R^2 reduced from 0.79 to 0.64. Comparatively, the predictive performance of the traditional method is better than of the Kronecker product implementation, with a smaller GOF and RMSE, and a higher R^2.

Where the Kronecker method excels is the computational cost. Both calibration approaches converged within the same number of MCMC iterations (500). The traditional method took 1.67 hours, roughly 5.5 times longer than the Kronecker method that converged within 0.3 hours.

Table 7.5: Comparison of the out-of-sample predictive performance and computational cost of the traditional and Kronecker product method of calibration. The Kronecker product method is an adaptation of the implementation proposed by Bayarri et al. (2009), while the traditional method is summarised by Chong and Menberg (2018).

Method	CV(RMSE) [%]	NMBE [%]	GOF [%]	RMSE [°C]	R^2	Cost [hours]
Uncalib.	11.51	-11.20	11.36	2.53	0.79	-
Traditional	2.67	-0.22	1.89	0.59	0.77	1.67
Bayarri	3.38	-0.61	2.43	0.74	0.64	0.30

7.3 Discussion

For the preferred method of Bayesian calibration, a parametric experiment was run and the out-of-sample predictions of the archetype-mean daytime living room temperature were compared against those measured within the homogeneous cluster of dwellings for the summer period. The inclusion of 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 that parameter. For all calibrations that converged, there was a reduction in RMSE of the out-of-sample prediction, compared to pre-calibration, from 2.53 °C to 0.58–0.70 °C. The detailed investigation of one of the best performing models revealed

parameter identifiability problems for EGF. The improvement in out-of-sample prediction despite the lack of identifiability for a parameter is not surprising, and it has been noted in the past that predictive performance can be improved even if parameter identifiability problems exist (Arendt et al., 2012).

7.3.1 Comparison with Other Studies

Since this is the first example of Bayesian calibration on an archetype-based model of summer indoor temperature, it is not possible to compare its post-calibration predictive accuracy against similar work. However, useful insights can be generated by revisiting the outcomes of other calibration studies.

From the review of studies on Bayesian calibration of housing stock models of energy performance, presented in Section 2.4, three papers used CV(RMSE) and NMBE to quantify their model's performance and bias in the validation period (Hedegaard et al., 2019; Kristensen et al., 2017b; Kristensen et al., 2018). All three studies used the ISO:13790:2008 model, but the temporal resolution of observations varied from hourly (Hedegaard et al., 2019) to annual (Kristensen et al., 2017b), while the train/validation split ranged from 75 %/ 25 % (Kristensen et al., 2017b) to 25 %/ 75 % (Kristensen et al., 2018). In all three studies, the CV(RMSE) was higher (5.6 % to 26.5 %), and the lowest model bias was worse (NMBE = -1.39 %) than what was achieved in this study (CV(RMSE) = 2.67 %, NMBE = -0.22 %). Since the variability of errors (CV(RMSE), closer to 0 is better) and the extent to which the calibrated model under- or over-predicts (NMBE, closer to 0 is better) are smaller in this work than in the studies reviewed, the relative performance of the calibrated UK-HSM is better. However, this is does not suggest that the calibration procedure employed in this work is superior, since a different model, output type and data were used.

While the use of daily mean indoor temperature has limited the influence of stochastic occupant behaviour in this thesis (compared to hourly temperature, for example), the same could be argued for the calibration of annual energy use by Kristensen et al. (2017b). The classification process or the priors used may be partly responsible for differences in the post-calibration accuracy, as would the model

being calibrated. However, the most important factor is expected to be the modelled quantity, its dynamics and causal influences.

7.3.2 Is the model calibrated?

Calibration was defined in Section 2.3.5 as the process of learning the values of unknown model inputs using field observations of the model output (Kennedy and O'Hagan, 2001). While the predictive performance of UK-HSM following the calibration has substantially improved, it is not perfect, as indicated by the non-zero CV(RMSE) and NMBE. Given these results, a pertinent question to ask is whether the model input values have been learned, and if the model can be considered *calibrated*.

One could approach this by questioning whether the model's predictive performance could be further improved. In most cases, further improvements are possible, especially for models of real-world complex systems. In the case of UK-HSM, model refinement along with the use of more training data and lag components could potentially improve predictive performance further. However, validation errors will still exist as demonstrated by Calama-González et al. (2021): for the well-defined conditions of a test cell, Bayesian calibration of an indoor temperature model resulted in CV(RMSE) = 0.81 %, NMBE = -0.22 % and R^2 = 0.98.

A more practical way of evaluating the level of model calibration is by establishing whether the model's predictive performance is sufficient for its intended use. As Ruiz and Bandera (2017) discussed, guidelines that prescribe acceptable levels of discrepancy have been developed for models of energy use. One such example is ASHRAE Guideline 14, with thresholds for hourly values of CV(RMSE) = 30 % and NMBE = $\pm$10 % (ASHRAE, 2002). In the absence of equivalent guidelines for indoor temperatures, the same thresholds have been adopted (O' Donovan et al., 2019). However, as argued by Jain et al. (2020), the use of these acceptance limits is problematic since CV(RMSE) and NMBE are scale-dependent and, thus, allow for a high variation in temperatures that can exceed comfort bands. For example, assuming an average indoor temperature of 20 °C, the CV(RMSE) threshold of 30 % translates to an RMSE of 6 °C.

Developing a set of appropriate thresholds for models of summer indoor temperature is beyond the scope of this work, however it merits some discussion. Defining such thresholds would largely depend on the magnitude of deviations in the quantity of interest that would be acceptable, although factors such as measurement error might also need to be taken into consideration. For example, if the model's purpose is to estimate heat-related mortality, what level of under or over-estimation would be tolerable, and how does that relate to indoor temperatures? It is likely that this threshold will be different if the model's purpose was to quantify hourly thermal comfort.

Since there are no established thresholds on the calibration of indoor temperature models for heat mortality calculation purposes, a direct answer to this question is not provided. The reduction in RMSE from roughly 2.5 °C, to approximately 0.6 °C is promising. Such RMSE is smaller than the 0.94 - 1.73 °C RMSE reported in the empirical validation of the semi-detached UK-HSM archetype (Symonds et al., 2017).[4] Even more encouraging is the fact that NMBE for the calibrated models ranges between -0.22–0.22 %, similar to the NMBE obtained for the test cell calibration of Calama-González et al. (2021), suggesting that the calibrated model does not have a tendency to under- or over-predict. Therefore, if using the model to estimate heat mortality over several days, for example, differences on individual days would tend to cancel out. The use of more training data would likely further improve the calibrated model's predictive performance, however, the improvement already observed is significant. Note that for simplicity, throughout this chapter, the model that had undergone calibration was referred to as *calibrated*.

7.3.3 Posterior Analysis Interpretation

The main outcome of any Bayesian analysis is the posterior distribution(s). This remains true even when improved model prediction following Bayesian calibration is the research aim; it is through the estimation of posterior distributions for the calibration parameters and model hyperparameters that predictive performance may

[4] Although Symonds et al. (2017) used the daily maximum instead of the daily mean living room temperature.

improve. Thus, it is important to reflect on the generalisability and interpretability of the posterior distributions.

The posterior distributions resulting from this work are specific to the Bayesian analysis undertaken, one that combined archetype-level data and appropriate priors. The posterior distribution of calibration parameters represents the set of plausible values, and their associated probability, that may be used to model the archetype mean. By sampling from the posterior distributions, it is possible to estimate a distribution of archetype mean predictions for each daily value.[5] Using the derived posteriors in predicting the archetype-mean of daily indoor temperature for a group of dwellings that do not share the same characteristics as those used for the calibration (e.g. detached homes), may not result in accurate predictions and is not advisable. However, the use of these posteriors in the analysis of a different group of dwellings that shares the same characteristics as those of the homogeneous cluster used in the calibration, would be appropriate. If a calibration was to be performed for such a group of dwellings, the posteriors obtained in this PhD study could be used as the priors. The use of different weather conditions, but within the same range as those used for calibration, should result in predictions with similar deviation from reality as those presented in Figure 7.8.

In developing a model of a physical system, the aim is to try to represent reality as closely as possible. Since model inputs have a physical interpretation, it is worth considering whether the posterior distributions can inform modellers about the real, yet unknown, values of these quantities. For example, can something be inferred about real-life window-opening behaviour given the WOT posterior? With the introduction of this calibration framework, Kennedy and O'Hagan (2001) warned that it is dangerous to interpret the posteriors of the calibration parameters as direct estimates of physical values, especially when it is not expected that the model could provide a perfect fit to the observations even if the physical quantities were accurately known. UK-HSM tries to represent an inherently stochastic system in

[5]Note that the spread in predicted daily archetype mean temperatures is different to the (non-averaged) spread of indoor temperatures expected to be observed within the homogeneous group on any given day.

a deterministic fashion with multiple hard-coded assumptions and simplifications, including those relating to occupancy. With regard to window operation, UK-HSM assumes that windows can only open during specific times of the day if the indoor temperature exceeds a threshold and is higher than the outdoor temperature. The calibration of UK-HSM can inform the choice of the threshold, resulting in better predictions, but it is not possible to claim that something has been learned about the real-life window-opening behaviour of occupants; in reality the *probability* of window opening – not the certainty offered by the deterministic specification of UK-HSM – varies during the day, between occupants and is influenced by several factors such as indoor CO_2 concentration, solar radiation, noise levels, occupant preferences, activities and security concerns (Fabi et al., 2012; Mavrogianni et al., 2016). The inclusion of a model bias term aims to capture discrepancies between monitored data and model predictions when the "true" calibration parameter values are used, however, its effectiveness may depend on the application, and it has been demonstrated that its inclusion may itself be a reason for parameter identifiability issues (Arendt et al., 2012). Thus, it is unlikely that this term is able to completely capture uncertainties arising from limitations in the model's structure. A final point regarding the interpretability of posterior distribution regards the use of sensitivity analysis to select the calibration parameters. Even if it could be assumed that the model is capable of providing perfect predictions, the uncertainty of parameters that are not calibrated may be partly lumped onto the posteriors of the calibrated parameters (Booth et al., 2012). Previous work that investigated this effect has shown that its magnitude can be small (Heo et al., 2015), yet it is likely to differ between applications.

7.3.4 Kronecker Product

Implementing a Kronecker product specification for the Bayesian calibration process, inspired by Bayarri et al. (2009), was shown to have advantages and disadvantages. The predictive performance improved following the calibration according to most but not all metrics, and the improvement was smaller compared to the traditional method. A clear benefit of this alternative implementation is the computational cost,

as the traditional method took 5.5 times longer under the same settings.

Following from these results, choosing the traditional method over the Kronecker product implementation may depend on the application. If minimising computational cost is crucial, and minimising RMSE is more important than maximising the R^2, then the Kronecker product may be preferred. This decision will also depend on the size of the data. By considering the theoretical estimates of computational cost (see Chong et al. (2017) and Appendix Section F.2), doubling the data used for the traditional method would result in an eight-fold increase in computational cost (since it scales proportionally to the cubic power of the total number of observations used in the calibration (N_c), $\propto \mathcal{O}(N_c^3)$). In comparison, the increase in computational cost for the Kronecker product method will depend on which quantity is doubled: the number of homes (monitored (M) and simulated (S)) or the number of observations per home (D_c). If the contribution to the computational cost is equal for the two quantities (i.e. the number of homes is equal to the number of observations), doubling one of the two would theoretically result in a 4.5 increase in computational cost (since the cost for the two quantities may be approximated as $\mathcal{O}((M+S)^3 + D_c^3)$, based on Hung et al. (2015)). Therefore, in most cases, the relative benefit of using the Kronecker product method increases with an increase in the amount of data.

What remains unknown is how the performance of the Kronecker product method compares to the traditional approach under different scenarios. The differences between the two calibration approaches are likely due to the simplifying assumptions made for the model bias hyperparameters (see Appendix Section F.2). It is possible that the impact of these assumptions may be altered if more data were used or if a different model was being calibrated.

7.3.5 Limitations

This chapter has focused on the calibration of the free-floating, daytime, mean living room temperature of a single homogeneous cluster of semi-detached dwellings. It would be of interest to apply this approach to other clusters and for the nighttime bedroom temperature and quantify the improvement in predictive performance. Since the aim of this work was to demonstrate the application of the Bayesian calibration

framework, the use of a single cluster was considered sufficient.

Data were only available for a relatively small number of homes and for only part of the 2009 summer. The availability of data over a longer time period, and preferably over several summers, would have allowed the model to be calibrated and validated against a wide range of weather conditions. Nevertheless, the summer of 2009 included both a heatwave which was partly captured by the monitoring campaign and a few cool days with daily-mean ambient temperatures below 15 °C.

This calibration focused on free-floating indoor temperatures, since UK-HSM assumes no heating or air conditioning during the summer period. Air conditioning was not present in the homes monitored during the 2009 4M survey, and the penetration of such technologies remained low (2 %) in 2017 based on the most recent Energy Follow Up Survey (BEIS, 2021c). The prevalence of heating during the summer period is higher, with 18.4 % showing signs of heating between July-August according to Lomas and Kane (2013). Since the calibration did not consider heating or cooling, it is not possible to extend the prediction of the calibrated model to conditioned homes.

While informative, the parametric experiment did not exhaust all possible combinations of calibration variables and weather variables. As an example, the inclusion of a third outdoor temperature lag component could have resulted in better out-of-sample prediction. It would also be interesting to determine the effect of including the Floor Area Factor on the model's predictive ability. However, due to computational cost of the calibration, ranging from one to two hours per calibration and generally increasing with the number of variables considered, it was determined that limiting the parametric experiment to the permutations outlined in Table 7.3 would suffice for this work.

An important simplification within this work is the assumption that the values of explanatory and weather variables are known exactly (i.e. there is no uncertainty associated with these values or any uncertainty that exists has a negligible influence). This is a commonly used assumption (Booth et al., 2012; Chong and Menberg, 2018), yet the extent to which it is appropriate has not been investigated.

The hyperparameter priors were based on the recommendations of Chong and Menberg (2018) and Menberg et al. (2019). It is possible that other priors may have been more appropriate, especially for the variance hyperparameters, yet this was not investigated. While Chong and Menberg (2018) provides some explanation on the choice of these hyperparameters, the field of Bayesian calibration modelling would likely benefit from a more in-depth analysis of these choices and how they would differ for different applications.

7.4 Summary

In response to the third research objective, this chapter sought to quantify the improvement in predictive performance of UK-HSM following the application of the Bayesian calibration framework, and to reduce model input uncertainty for a homogeneous group of dwellings. The homogeneous group consisted of 26 semi-detached dwellings, monitored during the 4M project. A 10-day period was used for the calibration, and a 52-day unseen period was used for validation. The calibration was repeated 24 times for a different combination of weather and calibration parameters.

An improvement in out-of-sample predictive performance was observed for all calibration experiments. The Root Mean Square Error (RMSE) reduced from 2.53 °C to 0.58–0.70 °C depending on the choice of parameters. The inclusion of a second outdoor temperature lag component resulted in a marginally better performance for all models compared to those with just one. The detailed analysis of one calibration experiment revealed a strong positive relationship between model bias and the lag components of the outdoor temperature, indicating a potential model inadequacy. The posterior distribution of the Electrical Gains Factor was similar to the prior, suggesting a lack of identifiability. This was not the case for the Window Opening Threshold, with a posterior distribution centred around 21.8 °C and 90 % credible interval of 20.7–22.9 °C, compared to a prior centred around 23.6 °C. An alternative formulation for the Bayesian calibration, novel within the field of built environment research, was proposed and tested; it resulted in a reduction of 82 % in computational cost, albeit with a smaller improvement in predictive performance compared to the

traditional implementation.

The next chapter provides the overarching discussion of this doctoral study. It reflects on how the aim and research objectives set out in Section 1.3 have been achieved, discusses the overall limitations of this work, and suggests future work that may be motivated by this study.

Chapter 8

Discussion & Conclusions

Over the next two decades, decisive action must be taken to reduce global greenhouse gas (GHG) emissions and adapt to a changing climate; failure to do so may result in impacts that are multiple times higher than those already experienced (IPCC, 2022). Effective climate change mitigation and adaptation requires actions in several sectors, including the housing sector, both globally and within the UK (CCC, 2021b; HMG, 2021b; IEA, 2021; IPCC, 2022). While policies to future-proof the UK housing stock and reduce its GHG emissions have been introduced, clear gaps remain (CCC, 2022). One such gap relates to the adaptation of the existing housing stock to high temperatures in order to safeguard the occupants' health and wellbeing against indoor overheating (CCC, 2022). Clear and strong policies coupled with rapid and effective implementation are required to address this adaptation gap, in conjunction with policies that guide the reduction of the housing stock's carbon footprint. Building stock models for energy consumption are thought to be a key tool for "assisting with the efficient and rational implementation of policy" (Kavgic et al., 2010). Similarly, building stock models that evaluate the impact of climate change, mitigation and adaptation policies on indoor overheating can also provide crucial support to policymakers, and one such model is the archetype-based UK Housing Stock Model (UK-HSM) (Taylor et al., 2015). Yet, as with any modelling approach that aims to mathematically represent a complex natural system, uncertainties are unavoidable (Mulligan and Wainwright, 2013; Saltelli et al., 2008). To confidently use model predictions to guide policymakers, modelling uncertainties should be

quantified and reduced, something that can be achieved using Bayesian calibration (Booth et al., 2012).

Literature reviews by Hou et al. (2021) and Oraiopoulos and Howard (2022) revealed that model calibration based on Bayesian inference has found several uses within the field of building modelling, including several calibration examples of archetype-based building stock models of energy use. Fewer examples of Bayesian calibration for building models of indoor temperature exist, and no published work could be identified where such methods were applied to archetype-based models of summer indoor temperature (Hou et al., 2021). Undertaking Bayesian calibration for archetype-based housing stock models requires that the housing stock is classified into homogeneous groups of dwellings. This process is often not discussed in great detail in published examples of Bayesian calibration, with a clear definition of homogeneity not being provided, and the process described in published work is often disjointed from the calibration (see Section 2.4.2). Further, a clear and rigorous process of identifying appropriate prior probability distributions for archetype-based models could not be identified.

Motivated by these gaps in the published literature, and the importance of using models whose uncertainties have been quantified and reduced to support policymaking, this thesis introduced a new Bayesian calibration framework for archetype-based models of summer indoor temperature. The framework addresses the classification of dwellings into homogeneous groups, and the characterisation of model input distributions that may be used for forward uncertainty propagation or as priors in the calibration. Chapter 3 outlined the proposed framework, while Chapters 4–7 described and discussed the findings and limitations from its application to UK-HSM. This chapter evaluates the achievements, contributions, limitations and implications of this work. Section 8.1 provides a summary of the key conclusions of this work. Section 8.2 reflects on whether the research aim and objectives set out at the start of this thesis have been achieved. The novel contributions of this study – which include methodological advancements in building modelling and Bayesian calibration, and empirically-based advice regarding indoor overheating to industry

practitioners and policymakers – are detailed in Section 8.3, grouped into three categories based on the target actors: academia, industry and policy. In Section 8.4, a discussion regarding the limitations of this work's scope is provided (methodological limitations specific to each set of results can be found in Chapters 4–7). Finally, Section 8.5 outlines the future work that may follow on from this thesis.

8.1 Conclusions

Several key conclusions arise from this PhD thesis, and these are summarised below:

1. In Chapter 4, several dwelling and household variables were shown to have a statistically significant association with the summer indoor temperature monitored in approximately 800 dwellings during the 2011 English Housing Survey Energy Follow-Up Survey (EHS-EFUS) and standardised against regional weather conditions. These findings contribute to the existing body of knowledge, and can inform the design and adaptation of homes to reduce their propensity for overheating. However, the statistically significant correlation found to exist between explanatory variables (Section 4.2.4) highlights the difficulty in drawing causal conclusions from these results.

2. In Section 4.2.1, the indoor overheating risk was quantified based on the criteria defined in CIBSE's Technical Memorandum 59 (CIBSE, 2017). For the relatively cool summer of 2011, the prevalence of indoor overheating according to Criterion 1 was 2.5 %. However, when considering Criterion 2 almost 26 % of dwellings were found to overheat. These results were not in good agreement with the occupants' stated thermal discomfort, reinforcing concerns regarding the effective quantification of indoor overheating risk using these criteria.

3. The analysis carried out in Chapter 5 revealed that distributions informed by empirical data can be derived for several building characteristics of English homes modelled by UK-HSM, and a method for identifying an appropriate distribution for a given dataset was introduced. However, detailed empirical

data from English homes were not available for all UK-HSM inputs, including the building's solar absorptivity and window opening threshold. The lack of such data can compromise the accuracy of model predictions, as highlighted in Section 7.2.1.

4. The sensitivity analysis described in Chapter 6 revealed that Window Opening Threshold was the dominant UK-HSM model input, followed by the Glazing Fraction, Orientation and Electrical Gains Factor (this ranking depends on the range of values assessed for each model input, which were informed by the best available evidence). This result provides further evidence to the importance of window opening in determining summer indoor temperature. The sensitivity analysis also suggested that several model inputs may exhibit a non-linear relationship with summer indoor temperature or an interaction with other inputs.

5. The application of the Bayesian calibration framework introduced in Section 3.2 was demonstrated using UK-HSM, the 2011 EHS-EFUS, and the 2009 4M survey in Leicester. The calibration was successful in reducing the model's out of sample root-mean-square error from approximately 2.5 °C to roughly 0.6 °C, and in quantifying and reducing model input uncertainty (Sections 7.2.2-7.2.3).

6. The post-calibration improvement in predictive performance depended on the choice of calibration and explanatory variables, as was revealed by the parametric experiment described in Section 7.2.2. Of particular significance to this specific application was the use of lag components of outdoor temperature.

7. The use of an alternative formulation for Gaussian processes, that takes advantage of the regular structure in building simulation data and the properties of the Kronecker product, can substantially reduce computational cost (Section 7.2.4). However, a trade-off may exist between computational complexity and the calibrated model's predictive accuracy. While further investigation of this trade-off is required, the alternative formulation may enable the use

of Bayesian calibration using Gaussian processes in cases where it would otherwise be computationally prohibitive.

8.2 Research Aim and Objectives

The aim of this work, as set out in Section 1.3, was *to quantify and reduce uncertainties of archetype-based housing stock models of indoor temperature.* Following from this aim, three objectives were specified to guide this research.

8.2.1 Research Objective 1

The first research objective was:

1. To develop a Bayesian calibration framework for archetype-based housing stock models of summer indoor temperature

In response to this objective, a framework was successfully developed and introduced in Section 3.2. The framework consists of five steps and relies on a clear and practical definition of homogeneity. It covers the classification of the housing stock into homogeneous groups of dwellings (Steps 1–2 and 4), the stochastic characterisation of model inputs (Step 3) for each group, the identification of the most influential model inputs (Step 4) and the Bayesian calibration of each group (Step 5).

8.2.2 Research Objective 2

The second research objective was:

2. To quantify the uncertainty of the UK Housing Stock Model inputs with the greatest influence on summer indoor temperature for a single homogeneous group of dwellings.

This objective was successfully achieved via classification, stochastic characterisation and sensitivity analysis work as described in Steps 1–4 of the Bayesian calibration framework (Section 3.2). In Step 1, the association of nine household and eleven dwelling characteristics with the standardised summer indoor temperature (SIT), monitored as part of the 2011 EHS-EFUS (Hulme et al., 2013a), was analysed

using Kruskal Wallis and Pairwise Mann-Whitney U-tests (Section 4.2). Based on the variables identified to be statistically associated with SIT, Categorical Variable Classification (Step 2) was used to identify a single group of dwellings suspected to be homogeneous (Section 4.2.6.2). In Step 3, a probability distribution was identified for each continuous model input of UK-HSM (Section 5.2). Where empirical data were available, a novel technique for fitting probability distributions was used to identify evidence-based distributions. In the absence of empirical data, probability distributions were defined based on the expected distributional form of that model input. A two-stage sensitivity analysis (Step 4) revealed that all influential continuous model inputs of UK-HSM are described by unimodal distributions (Section 6.2). Thus, the cluster of dwellings identified in Step 2 can be considered homogeneous, and the uncertainty of the most influential model inputs (identified in Step 4) is described by the probability distributions defined in Step 3.

8.2.3 Research Objective 3

The third and final research objective was:

3. To quantify the level of improvement in the predictive ability of the UK Housing Stock Model following application of the Bayesian calibration framework and reduce model input uncertainty for a homogeneous group of dwellings.

The final research objective was also successfully achieved – it relied on Step 5 of the Bayesian calibration framework, but also drew on Steps 1–4; the group was identified based on Steps 1, 2 and 4 (as per the 2nd research objective), while the probability distributions identified in Step 3, and the outcomes of the sensitivity analysis in Step 4 fed into Step 5, the Bayesian Calibration. Empirical data of summer indoor temperature, collected during the 4M project (Lomas and Kane, 2013) and aggregated to a daily resolution, were used in the calibration. Data were split into a training set, a 10-day period used to calibrate UK-HSM, and a validation set, consisting of 52 days and used to quantify the out-of-sample predictive performance post-calibration (Section 7.1.5). The calibration procedure was applied on twenty-four combinations of calibration and weather variables, in a type of parametric

experiment (Section 7.1.4). All experiments were assessed using five metrics, and all showed an improvement compared to the uncalibrated model across all metrics except R^2 (Section 7.2.2). For the best performing experiment, the Root Mean Square Error (RMSE) for the unseen validation period reduced from approximately 2.5 °C to roughly 0.6 °C. The model input uncertainty for Window Opening Threshold, the most influential model input, was reduced following the calibration from 90 % credible interval of 10.9 °C to 2.2 °C (Section 7.2.3.2).

8.3 Novel Contributions

The contributions of this doctoral study are summarised in the following sections, grouped under three categories: academia, industry and policy.

8.3.1 Academia

With its focus on capturing and reducing uncertainties in housing stock models of summer indoor temperature, this thesis has made a number of contributions to built environment research, especially relevant for modellers working in the area of climate change adaptation.

The first contribution is a modular framework for the classification of a building stock into homogeneous groups of dwellings, and the Bayesian calibration of archetype-based building stock models of summer indoor temperature (Section 3.2). While developed with models of summer indoor temperature in mind, it is expected that the framework can also be used in the calibration of other types of archetype-based building stock models, such as those of winter indoor temperature, energy use, ventilation or indoor air quality. The framework is flexible in that the methods used at each step can be modified depending on the application, the data available and the modeller's preference.

A further contribution to academia 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. Calibration with a single

lag component resulted in substantial improvement in the model's performance compared to the use of no lag components, the addition of a second lag component further improved out-of-sample predictive performance (Section 7.2.2). Further, results suggest a strong correlation between model discrepancy and outdoor temperature lag components, possibly relating to the modelling of thermal mass (Section 7.2.3), which merits further investigation. Moreover, the posterior distributions identified through the calibration of a group of semi-detached dwellings could be used for the prediction of indoor temperature in a group of dwellings with similar characteristics, and as priors if further calibration of this UK-HSM archetype were carried out.

As part of Step 1 of the framework, the statistical association of summer indoor temperatures with dwelling and household characteristics was examined for approximately 800 dwellings monitored during the 2011 EHS-EFUS (Section 4.2.2–4.2.3). In addition, the indoor overheating risk was quantified for the same sample of homes according to the two indoor overheating risk criteria defined within TM59 (Section 4.2.1). The findings from this analysis, published as a journal paper, contributed to the pool of knowledge regarding summer overheating in the English housing stock (Petrou et al., 2019b).

An open-source method, novel within the field of building modelling, for identifying model input distributions based on empirical data was developed (Section 5.1.1). As demonstrated within this work, the method may be used to inform the prior distributions in archetype-based Bayesian model calibration. It may also be used to identify distributions of model inputs for individual buildings if repeated measurements may need to be taken, especially when the measurement method is associated with large uncertainties. In addition, the approach detailed in this thesis may be applicable to modelling parameters that vary over the simulation period in a single building, as might be the case for occupancy-related inputs. Certain model outputs may also be represented as distributions, and an appropriate distribution may be identified using the same procedure. A further contribution to knowledge within the building modelling field is the discussion surrounding the interpretation of fitted theoretical distributions in Section 5.3.

An alternative implementation to Bayesian calibration that relies on Gaussian Processes (GP) for surrogate modelling has been proposed (Section 7.1.6). The GP is defined using a Kronecker product formulation, taking advantage of the structure of the data. This approach has the potential to significantly reduce computational cost, one of the key obstacles to GP-based Bayesian calibration. Compared to the traditional method, the Kronecker product approach took a fraction of the time ($\approx$18 %) and resulted in similar, albeit slightly worse, performance (Section 7.2.4). It is expected that gains in computational efficiency will increase together with dataset size. Within academia, this method may allow calibration and emulation to be implemented using Gaussian Processes where previously the processing time would have been prohibitive, even for research purposes.

8.3.2 Industry

Outcomes from the empirical and modelling work of this thesis provide insights to industry practitioners, such as architects or building engineers, regarding factors associated with high summer indoor temperature. In turn, such knowledge encourages building design and retrofit practices that are less prone to indoor overheating.

The analysis of the 2011 EHS-EFUS of indoor temperature revealed that dwellings which are purpose-built flats, are relatively small, are located in the city, have communal heating or have little-to-no loft insulation are likely to experience higher summer indoor temperatures (see Section 4.2 for all such results). Further, outcomes from the modelling component of this work reinforce the importance of ventilation in minimising indoor overheating in the UK (Section 6.2). Window operation was shown to be far more dominant than any other building- or occupancy-related model input in regulating mean daily indoor temperature. Following from these results, homes with high levels of loft insulation, designed in a way that allows effective ventilation while limiting internal gains during the summer – that could in some cases arise from poorly installed hot water and heating systems – are expected to be less likely to overheat. It is likely that such factors are especially important in small, purpose-built flats located in urban areas.

Findings from this thesis may also contribute to the improvement of indoor

overheating assessment at the building design stage. A relevant methodology is described by Technical Memorandum 59 (TM59), released by the Chartered Institution of Building Services Engineers (CIBSE) (CIBSE, 2017). This method, which relies on the use of Building Performance Simulation (BPS) tools and the overheating criteria used in Chapter 4, has been adopted by Approved Document O as a means to demonstrate compliance with Part O of the Building Regulations (HMG, 2021a). A pertinent finding is the relatively large disagreement observed in Section 4.2.1 between the empirical assessment of indoor overheating, using the TM59 overheating criteria, and the stated thermal discomfort, especially for the bedroom overnight criterion. Such disagreement could be evidence of the partial inability of the overheating criteria to effectively detect thermal discomfort. A second finding of relevance to the current and future use of TM59 and ADO is the discrepancy observed between modelled and monitored indoor temperatures pre-calibration, and the relative increase in agreement post-calibration (Section 7.2.1–7.2.2). While this work is not the first to demonstrate that such differences exist (e.g. Roberts et al., 2019; Symonds et al., 2017), it has highlighted the importance of parametric uncertainty, model discrepancy and calibration for models of summer indoor temperatures. Given these findings, it is important that industry practitioners critically evaluate the outcomes of their overheating assessment, and where possible, collect empirical data post-construction that may be used to refine overheating criteria, calibrate their models and improve future iterations of TM59. While this may be a challenging undertaking for many practitioners, especially for smaller firms with limited experience in using BPS tools, it should be pursued where possible. To enable and encourage this, software developers should make it easy for industry practitioners to carry out uncertainty quantification, model optimisation, and calibration. Thus, it is promising to see that widely used commercial software, such as DesignBuilder (DesignBuilder, 2021), have incorporated such capabilities. The consideration of such factors within the wider modelling community could reduce the performance gap (Jain et al., 2020) and result in better performing buildings.

The methodological outcomes of this thesis, such as the distribution fitting

method or the alternative Bayesian calibration implementation, may also contribute to the advancement of existing BPS tools. Software developers can advance their existing uncertainty quantification method, if present, using the distribution fitting method, which can also be used to summarise model outputs. The Kronecker product implementation proposed within this work could also be implemented within BPS tools to enable Bayesian calibration at a more manageable computational cost. Finally, the distributions identified in Section 5.2.13, and potentially the posterior distributions in Section 7.2.3.2, could form the basis for a library of default distributions in such tools, with the caveat that they are updated as data becomes available.

8.3.3 Policy

As with industry professionals, findings from the statistical analysis in Chapter 4 can also inform policymakers on aspects relating to indoor overheating. Households with children, rented (privately, from local authority or registered social landlords), or with at least one occupant on means tested or certain disability related benefits were characteristics associated with a higher indoor temperature in the bedroom and/or living room (Section 4.2). Thus, policies that prioritise such groups, while also taking into account their vulnerability, will likely be the most effective in reducing thermal discomfort in homes and the adverse impacts associated with it. Moreover, the design of homes which, due to their physical characteristics, are more likely to overheat (e.g. small, top-floor flats with little loft insulation and limited ventilation capacity) should be discouraged, and the inclusion of overheating interventions should be strongly encouraged.

A further contribution of this doctoral work to policymakers is indirect, but may be substantial. Building stock models, can be used to inform policy (Oraiopoulos and Howard, 2022). Furthermore, building models are now considered a viable option to demonstrate compliance with the requirement for overheating assessment in the Building Regulations. As argued in Section 2.3.1 through Rosen's diagram, it is not possible to avoid modelling uncertainties, and efforts should instead concentrate on their quantification and reduction. Through the application of the Bayesian calibra-

tion framework, described in Chapter 7, the out-of-sample predictive performance of UK-HSM improved, its predictive uncertainty was quantified and possible reasons for model discrepancy were identified. Thus, more trust may be placed in the future use of UK-HSM in advising policy. Such practice should become commonplace in all models, and policymakers should expect and require modelling uncertainties to be quantified and reduced where possible. To enable this, emphasis must be placed on the large-scale and frequent collection of data, including detailed household and dwelling characteristics, thermal comfort surveys and high spatio-temporal resolution indoor temperatures. A good example is the Energy Follow-Up Survey, which may be improved if it takes place at regular and frequent intervals, covers a larger sub-sample of the English Housing Survey, and if the data becomes available to the wider research community as soon as possible to facilitate timely research. Making anonymised data Open Access, where possible, would likely accelerate research in the field.

8.4 Limitations

The previous sections in this chapter discussed how each of the research objectives in Section 1.3 have been met, and summarised the accomplishments and novel contributions of this work. Despite these achievements, limitations exist, of course, as the following paragraphs discuss.

The first limitation relates to the focus on the example of UK-HSM, despite the fact the aim is not bound to a specific archetype-based model. It would not have been possible to apply this framework on multiple such models within this thesis, nor it is thought to be necessary, since the process of quantifying and reducing uncertainties should apply to similar models. The development of a framework for the calibration of such models, and the insights generated through its application to UK-HSM have addressed the aim of this research. Yet, it should be acknowledged that application of this framework to other models might result in a different set of influential model inputs being calibrated, and a different level of post-calibration improvement – these will depend on factors such as the model's structure, choices

about temporal resolution and the empirical data available.

Another limitation relates to the choice of UK-HSM output used in the calibration: the mean of the daytime living room temperature during the summer period. The choice of a daily resolution was largely motivated by the recent and planned applications of UK-HSM in estimating the impact of home energy efficiency measures or overheating interventions on heat-related mortality (Taylor et al., 2018b; Taylor et al., 2021). This work did not explore the improvements in hourly performance that may be associated with a calibration at a daily resolution, nor did it attempt to calibrate UK-HSM at an hourly resolution. This is a limitation of this work, given the importance of hourly indoor temperature in indoor overheating assessment. Further, since the empirical data available for calibration did not include mechanically-cooled homes, and UK-HSM only models free-floating summer indoor temperatures, the calibration does not extend to homes with any kind of thermal conditioning in the summer. This modelling scenario was considered appropriate given the limited penetration of air-conditioning in the UK currently (BEIS, 2021b), yet there is merit in future work carrying out a calibration for air-conditioned homes.

The calibration carried out in this thesis was limited to the living room temperatures of a single group of semi-detached dwellings. This decision was mainly driven by the limited empirical data available for calibration, and the fact that the purpose of the second and third objective was to demonstrate the framework's application. It is important to determine the level of improvement that may be achieved for bedroom temperatures and for different group of dwellings. However, as with the case of applying this framework to a single archetype-model, the purpose was to validate the proposed framework, and it is expected that the calibration on other groups of dwellings and for other rooms would also result in an improvement in predictive performance.

A final limitation might be considered the choice of methods used for each step of the calibration framework. The framework is considered modular since each step is associated with multiple options, and a modeller could choose one option over another without impacting the rest of the workflow. In the calibration of UK-HSM,

while all options were informed by literature, only one of them was implemented in each step and the differences that might arise from the use of alternative methods in this case study were not investigated. Implementing several options for each step to compare and contrast their benefits and shortcomings was out of scope for this work.

8.5 Future Work

In future work being planned by the author, the framework will be used to calibrate the UK-HSM using the daily indoor living room and bedroom temperatures monitored during the 2011 EHS-EFUS (Hulme et al., 2013a), and the more recent EHS-EFUS 2017 (BEIS, 2021c). Through this approach, there are likely to be sufficient data for the calibration of several homogenous groups, and over a wide range of indoor temperature. In addition, any differences in posterior distributions for the same groups calibrated using the two EHS-EFUS datasets would be explored. Further future work will aim to calibrate UK-HSM using data of hourly resolution to enable its use for higher temporal resolution analysis than the daily frequency used in this work.

Furthermore, the author is planning to investigate the benefits and drawbacks of different data aggregation methods described in Section 2.4.5.3. In particular, the author would like to examine the implementation of a hierarchical approach, as suggested by Kristensen et al. (2018), but with the addition of a model bias component. Such an approach is thought to result in a calibrated model that is less biased by extreme values, yet its benefits compared to the method used in this work have not been quantified.

The statistical analysis presented in Chapter 4 revealed several dwelling and household characteristics that had a statistically significant association with summer indoor temperatures. However, it was not possible to confidently determine whether some of these associations were an artefact of correlations between variables. In addition, due to the lack of data on factors known to influence indoor temperatures, such as shading or window operation, confounding factors are likely to exist. Future work should aim to disentangle such correlations and quantify the causal effect

that dwelling and household variables have on summer indoor temperatures. This will likely require the use of causal inference techniques designed to extract such effects, and account for confounders, in observational studies. Of great benefit to the understanding of indoor overheating would be the systematic and large-scale collection of detailed empirical data of indoor temperature, dwelling and household characteristics, occupant actions and expressed thermal comfort. This would not be a small or inexpensive undertaking, but it would facilitate research that could make significant advancements in the prediction and mitigation of indoor overheating.

In addition, despite the few examples of using regression to standardise indoor temperature against ambient conditions (Hamilton et al., 2017; Oreszczyn et al., 2006; Wilkinson et al., 2001), a clear set of guidelines for how to carry out the standardisation procedure and determine its efficacy does not currently exist. Future work could try and develop such a set of guidelines with the use of detailed empirical or synthetic data, as described in Section 4.2.5.

In Section 6.2, the two-stage sensitivity analysis informed the use of the Floor Area Factor as an explanatory variable in the calibration. However, as discussed in Section 6.3, this approach did not examine its importance in conjunction with that of the weather variables. Future work could evaluate the use of other methods of sensitivity analysis that could assess the influence of model inputs and weather variables concurrently.

An important component in the calibration described in Chapter 7 was the use of lag components of outdoor temperature. The use of a single component resulted in improved predictive performance compared to the absence of a lag component and the uncalibrated model; further improvement was observed with the addition of a second lag component. It is possible that the inclusion of further lag components, or other variables derived from hourly outdoor temperature (e.g. daily maximum outdoor temperature) could result in further improvements. However, as the number of predictors increases, so does the number of hyperparameters that need to be tuned and the computational complexity of the calibration process. Thus, the inclusion of more variables may require the use of more data, inevitably leading to a larger

computational cost. Future work could investigate the effects of including further weather variables, but also replacing some of them with ones carrying a similar level of information. This may be achieved by the use of weighted weather variables, for example an exponentially weighted outdoor temperature variable (CIBSE, 2013). Given the moderate level of correlation observed between lag components in this work, an alternative could be variables derived by Principal Component Analysis (PCA) (Reimann et al., 2008).

An assumption made during the Bayesian calibration was that explanatory and weather variables have a negligible level of uncertainty. This is a significant assumption, yet it is far from uncommon (all papers reviewed in Section 2.4 made the same assumption). Future work should strive to quantify the impact that uncertainty in explanatory and weather variables has on the calibration of models of indoor temperature and energy use. This could be taken a step further if the calibration procedure is adapted to account for the fact that some predictors may be uncertain (Huard and Mailhot, 2006).

This work has revealed some shortcomings of UK-HSM that should be further investigated and addressed. One limitation identified in Chapter 5 was that the collection of occupant profiles should be further expanded. In addition, in Chapter 7 the median and mean values of the model bias were found to be non-zero, and a strong trend was identified between the model bias and the first lag component of the outdoor temperature. It is currently unclear whether this is a shortcoming of the intrinsic assumptions of EnergyPlus, or some of the hard-coded specifications of UK-HSM, and further work is needed to understand and correct this.

Finally, the application of the Kronecker product formulation is promising. However, further work is required to explore the advantages and limitations of this implementation compared to the traditional approaches for different modelling applications and dataset sizes.

Bibliography

ACE (2015). *The Cold Man of Europe*. Association for the Conservation of Energy.

AECOM (2019). *Research into Overheating in New Homes. Phase 1 Report.* Ministry of Housing, Communities and Local Government.

Anderson, M. et al. (2013). 'Defining Indoor Heat Thresholds for Health in the UK'. *Perspectives in Public Health* 133.3, pp. 158–164. DOI: 10.1177/1757913912453411.

Arendt, P. D., Apley, D. W. and Chen, W. (2012). 'Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability'. *Journal of Mechanical Design* 134.10. DOI: 10.1115/1.4007390.

Armstrong, B. et al. (2010). 'Association of Mortality with High Temperatures in a Temperate Climate: England and Wales'. *Journal of Epidemiology and Community Health*. DOI: 10.1136/jech.2009.093161.

Armstrong, B. et al. (2018). 'The Impact of Home Energy Efficiency Interventions and Winter Fuel Payments on Winter- and Cold-Related Mortality and Morbidity in England: A Natural Equipment Mixed-Methods Study'. *Public Health Research* 6.11, pp. 1–110. DOI: 10.3310/phr06110.

ASHRAE (2002). *ASHRAE Guideline 14-2002. Measurement of Energy and Demand Savings*.

Bastos, L. S. and O'Hagan, A. (2009). 'Diagnostics for Gaussian Process Emulators'. *Technometrics* 51.4, pp. 425–438. DOI: 10.1198/TECH.2009.08019.

Basu, R. and Samet, J. M. (2002). 'Relation between Elevated Ambient Temperature and Mortality: A Review of the Epidemiologic Evidence'. *Epidemiologic Reviews* 24.2, p. 190. DOI: 10.1093/epirev/mxf007.

Bayarri, M. J. et al. (2007). 'A Framework for Validation of Computer Models'. *Technometrics* 49.2, pp. 138–154. DOI: 10.1198/004017007000000092.

Bayarri, M. J. et al. (2009). 'Predicting Vehicle Crashworthiness: Validation of Computer Models for Functional and Hierarchical Data'. *Journal of the American Statistical Association* 104.487, pp. 929–943. DOI: 10.1198/jasa.2009.ap06623.

Behboodian, J. (1970). 'On a Mixture of Normal Distributions'. *Biometrika* 57.1, pp. 215–217. DOI: 10.1093/biomet/57.1.215.

BEIS (2017). *The Clean Growth Strategy. Leading the Way to a Low Carbon Future.* P. 167.

BEIS (2021a). *2020 UK Greenhouse Gas Emissions, Provisional Figures*. Department for Business, Energy & Industrial Strategy, p. 22.

BEIS (2021b). *Cooling in the UK*. Department for Business, Energy & Industrial Strategy.

BEIS (2021c). *Energy Follow Up Survey: Thermal Comfort, Damp and Ventilation*. Department for Business, Energy & Industrial Strategy.

BEIS (2021d). *Household Energy Efficiency Detailed Release: Great Britain Data to December 2020*. Department for Business, Energy & Industrial Strategy.

Beizaee, A., Lomas, K. J. and Firth, S. K. (2013). 'National Survey of Summertime Temperatures and Overheating Risk in English Homes'. *Building and Environment* 65, pp. 1–17. DOI: 10.1016/j.buildenv.2013.03.011.

Bilionis, I. et al. (2013). 'Multi-Output Separable Gaussian Process: Towards an Efficient, Fully Bayesian Paradigm for Uncertainty Quantification'. *Journal of Computational Physics* 241, pp. 212–239. DOI: 10.1016/j.jcp.2013.01.011.

BIPM et al. (2008). *Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement. Joint Committee for Guides in Metrology, JCGM 100:2008.* First edition.

Bolstad, W. M. and Curran, J. M. (2017). *Introduction to Bayesian Statistics*. Third edition. Hoboken, New Jersey: Wiley.

Booth, A. T., Choudhary, R. and Spiegelhalter, D. J. (2012). 'Handling Uncertainty in Housing Stock Models'. *Building and Environment* 48 (Supplement C), pp. 35–47. DOI: 10.1016/j.buildenv.2011.08.016.

Bouchama, A. and Knochel, J. P. (2002). 'Heat Stroke'. *New England Journal of Medicine* 346.25, pp. 1978–1988. DOI: 10.1056/NEJMra011089. pmid: 12075060.

Braulio-Gonzalo, M. et al. (2016). 'Modelling Energy Efficiency Performance of Residential Building Stocks Based on Bayesian Statistical Inference'. *Environmental Modelling & Software* 83, pp. 198–211. DOI: 10.1016/j.envsoft.2016.05.018.

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 (2014). *The Government's Standard Assessment Procedure for Energy Rating of Dwellings (SAP 2012)*. Watford, England: Department of Energy and Climate Change (DECC).

BRE (2016). *Consultation Paper: CONSP:16. Review of Default U-values for Existing Buildings in SAP*. Building Research Establishment.

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

Brotas, L. and Nicol, F. (2017). 'Estimating Overheating in European Dwellings'. *Architectural Science Review* 60.3, pp. 180–191. DOI: 10.1080/00038628.2017.1300762.

BSI (2007). *BS EN 15251: Indoor Environmental Input Parameters for Design and Assessment of Energy Performance of Buildings Addressing Indoor Air Quality, Thermal Environment, Lighting and Acoustics*. London: British Standards Institution.

BSI (2019). *BS EN 16798-1:2019 Energy Performance of Buildings - Ventilation for Buildings. Indoor Environmental Input Parameters for Design and Assessment of Energy Performance of Buildings Addressing Indoor Air Quality, Thermal Environment, Lighting and Acoustics - Module M1-6*. The British Standards Institution.

Building Research Establishment and Department of Energy and Climate Change (2016a). 'Energy Follow Up Survey, 2011: Lookup File: Secure Access. [Data Collection].'

Building Research Establishment and Department of Energy and Climate Change (2016b). 'Energy Follow Up Survey, 2011: Lookup File: Secure Access. [Data Collection].'

Burnham, K. P. and Anderson, D. R. (2004). 'Multimodel Inference: Understanding AIC and BIC in Model Selection'. *Sociological Methods & Research* 33.2, pp. 261–304. DOI: 10.1177/0049124104268644.

Calama-González, C. M. et al. (2021). 'Bayesian Calibration of Building Energy Models for Uncertainty Analysis through Test Cells Monitoring'. *Applied Energy* 282, p. 116118. DOI: 10.1016/j.apenergy.2020.116118.

Campolongo, F., Cariboni, J. and Saltelli, A. (2007). 'An Effective Screening Design for Sensitivity Analysis of Large Models'. *Environmental Modelling & Software*. Modelling, Computer-Assisted Simulations, and Mapping of Dangerous Phenomena for Hazard Assessment 22.10, pp. 1509–1518. DOI: 10.1016/j.envsoft.2006.10.004.

CCC (2019). *UK Housing: Fit for the Future?* Committee on Climate Change.

CCC (2020). *The Sixth Carbon Budget: The UK's Path to Net Zero*. Climate Change Committee.

CCC (2021a). *COP26: Key Outcomes and next Steps for the UK*. Climate Change Committee.

CCC (2021b). *Independent Assessment of UK Climate Risk: Advice to Government For The UK's Third Climate Change Risk Assessment (CCRA3)*. Climate Change Committee.

CCC (2022). *Independent Assessment: The UK's Heat and Buildings Strategy*. Climate Change Committee.

Cerezo, C. et al. (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, pp. 321–334. DOI: 10.1016/j.enbuild.2017.08.029.

Chappells, H. and Shove, E. (2005). 'Debating the Future of Comfort: Environmental Sustainability, Energy Consumption and the Indoor Environment'. *Building Research & Information* 33.1, pp. 32–40. DOI: 10.1080/0961321042000322762.

Chong, A. and Menberg, K. (2018). 'Guidelines for the Bayesian Calibration of Building Energy Models'. *Energy and Buildings* 174, pp. 527–547. DOI: 10.1016/j.enbuild.2018.06.028.

Chong, A. et al. (2017). 'Bayesian Calibration of Building Energy Models with Large Datasets'. *Energy and Buildings* 154, pp. 343–355. DOI: 10.1016/j.enbuild.2017.08.069.

CIBSE (2006). *Guide A, Environmental Design*. London: Chartered Institute of Building Services Engineers.

CIBSE (2013). *The Limits of Thermal Comfort: Avoiding Overheating in European Buildings,TM52: 2013*. London: Chartered Institution of Building Services Engineers.

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

CIBSE (2017). *Design Methodology for the Assessment of Overheating Risk in Homes, TM59: 2017*. London: Chartered Institution of Building Services Engineers.

CIBSE (2020). *Health and Wellbeing in Building Services, TM40:2020*. S.l.: Chartered Institution of Building Services Engineers.

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

Coughlin, S. S. (1990). 'Recall Bias in Epidemiologic Studies'. *Journal of Clinical Epidemiology* 43.1, pp. 87–91. DOI: 10.1016/0895-4356(90)90060-3.

Crawley, D. B. et al. (2001). 'EnergyPlus: Creating a New-Generation Building Energy Simulation Program'. *Energy and Buildings*. Special Issue: BUILDING SIMULATION'99 33.4, pp. 319–331. DOI: 10.1016/S0378-7788(00)00114-6.

Crawley, D. B. et al. (2008). 'Contrasting the Capabilities of Building Energy Performance Simulation Programs'. *Building and Environment* 43.4, pp. 661–673.

De Dear, R. J. et al. (2013). 'Progress in Thermal Comfort Research over the Last Twenty Years'. *Indoor Air* 23.6, pp. 442–461. DOI: 10.1111/ina.12046.

De Gusmão, F. R. S., Ortega, E. M. M. and Cordeiro, G. M. (2011). 'The Generalized Inverse Weibull Distribution'. *Statistical Papers* 52.3, pp. 591–619. DOI: 10.1007/s00362-009-0271-3.

Delignette-Muller, M. L. and Dutang, C. (2015). 'Fitdistrplus: An R Package for Fitting Distributions'. *Journal of Statistical Software* 64.1 (1), pp. 1–34. DOI: 10.18637/jss.v064.i04.

Department for Communities and Local Government (2017). 'English Housing Survey, 2008-2014: Secure Access. [Data Collection]'.

DesignBuilder (2021). *DesignBuilder Software Ltd - Optimisation*. URL: https://designbuilder.co.uk/optimisation (visited on 27/12/2021).

Divine, G. W. et al. (2018). 'The Wilcoxon–Mann–Whitney Procedure Fails as a Test of Medians'. *The American Statistician* 72.3, pp. 278–286. DOI: 10.1080/00031305.2017.1305291.

DLUHC (2021). *English Housing Survey*. English Housing Survey. URL: https://www.gov.uk/government/collections/english-housing-survey (visited on 26/02/2022).

DOE (2016). *Engineering Reference Guide - EnergyPlus Version 8.6 Documentation*. U.S. Department of Energy.

DOE (2020). *EnergyPlus Version 9.4.0 Documentation: Tips and Tricks for Using EnergyPlus*.

DOE (2022). *EnergyPlus*. EnergyPlus. URL: https://energyplus.net/ (visited on 21/06/2022).

Fabi, V. et al. (2012). 'Occupants' Window Opening Behaviour: A Literature Review of Factors Influencing Occupant Behaviour and Models'. *Building and Environment* 58, pp. 188–198. DOI: 10.1016/j.buildenv.2012.07.009.

Famuyibo, A. A., Duffy, A. and Strachan, P. (2012). 'Developing Archetypes for Domestic Dwellings—An Irish Case Study'. *Energy and Buildings* 50, pp. 150–157. DOI: 10.1016/j.enbuild.2012.03.033.

Filogamo, L. et al. (2014). 'On the Classification of Large Residential Buildings Stocks by Sample Typologies for Energy Planning Purposes'. *Applied Energy* 135, pp. 825–835. DOI: 10.1016/j.apenergy.2014.04.002.

Firth, S. K. and Wright, A. J. (2008). 'Investigating the Thermal Characteristics of English Dwellings: Summer Temperatures'. *Windsor Conference*. Air Conditioning and the Low Carbon Cooling Challenge. Windsor, UK, p. 15.

Flaxman, S. et al. (2015). 'Fast Hierarchical Gaussian Processes', p. 18.

Fosas, D. et al. (2018). 'Mitigation versus Adaptation: Does Insulating Dwellings Increase Overheating Risk? - ScienceDirect'. *Building and Environment* 143, pp. 740–759. DOI: 10.1016/j.buildenv.2018.07.033.

Fouillet, A. et al. (2006). 'Excess Mortality Related to the August 2003 Heat Wave in France'. *International Archives of Occupational and Environmental Health* 80.1, pp. 16–24.

Garcia Sanchez, D. et al. (2014). 'Application of Sensitivity Analysis in Building Energy Simulations: Combining First- and Second-Order Elementary Effects Methods'. *Energy and Buildings* 68, pp. 741–750. DOI: 10.1016/j.enbuild.2012.08.048.

Gasparrini, A. et al. (2015). 'Temporal Variation in Heat–Mortality Associations: A Multicountry Study'. *Environmental Health Perspectives* 123.11, pp. 1200–1207. DOI: 10.1289/ehp.1409070.

Gelman, A. (2014). *Bayesian Data Analysis*. Third edition. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton: CRC Press. 661 pp.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge ; New York: Cambridge University Press. 625 pp.

Greenland, S. et al. (2016). 'Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations'. *European Journal of Epidemiology* 31.4, pp. 337–350. DOI: 10.1007/s10654-016-0149-3.

Ground Truth Definition and Meaning — Collins English Dictionary (2023). URL: https://www.collinsdictionary.com/dictionary/english/ground-truth (visited on 08/01/2023).

Hacker, J. N. et al. (2008). 'Embodied and Operational Carbon Dioxide Emissions from Housing: A Case Study on the Effects of Thermal Mass and Climate Change'. *Energy and Buildings* 40.3, pp. 375–384.

Hajat, S. et al. (2006). 'Impact of High Temperatures on Mortality: Is There an Added Heat Wave Effect?' *Epidemiology (Cambridge, Mass.)* 17.6, pp. 632–638. DOI: 10.1097/01.ede.0000239688.70829.63. pmid: 17003686.

Hajat, S. et al. (2014). 'Climate Change Effects on Human Health: Projections of Temperature-Related Mortality for the UK during the 2020s, 2050s and 2080s'. *J Epidemiol Community Health* 68.7, pp. 641–648. DOI: 10.1136/jech-2013-202449. pmid: 24493740.

Hamilton, I. G. et al. (2017). 'Old and Cold? Findings on the Determinants of Indoor Temperatures in English Dwellings during Cold Conditions'. *Energy and Buildings* 141 (Supplement C), pp. 142–157. DOI: 10.1016/j.enbuild.2017.02.014.

Heaviside, C., Macintyre, H. and Vardoulakis, S. (2017). 'The Urban Heat Island: Implications for Health in a Changing Environment'. *Current Environmental Health Reports* 4.3, pp. 296–305. DOI: 10.1007/s40572-017-0150-3.

Hedegaard, R. E. et al. (2019). 'Bottom-up Modelling Methodology for Urban-Scale Analysis of Residential Space Heating Demand Response'. *Applied Energy* 242, pp. 181–204. DOI: 10.1016/j.apenergy.2019.03.063.

Heo, Y., Choudhary, R. and Augenbroe, G. A. (2012). 'Calibration of Building Energy Models for Retrofit Analysis under Uncertainty'. *Energy and Buildings* 47, pp. 550–560. DOI: 10.1016/j.enbuild.2011.12.029.

Heo, Y. and Zavala, V. M. (2012). 'Gaussian Process Modeling for Measurement and Verification of Building Energy Savings'. *Energy and Buildings* 53, pp. 7–18. DOI: 10.1016/j.enbuild.2012.06.024.

Heo, Y. et al. (2015). 'Evaluation of Calibration Efficacy under Different Levels of Uncertainty'. *Journal of Building Performance Simulation* 8.3, pp. 135–144. DOI: 10.1080/19401493.2014.896947.

Herman, J. and Usher, W. (2017). 'SALib: An Open-Source Python Library for Sensitivity Analysis'. *The Journal of Open Source Software* 2.9. DOI: 10.21105/joss.00097.

Hesterberg, T. (2011). 'Bootstrap'. *WIREs Computational Statistics* 3.6, pp. 497–526. DOI: 10.1002/wics.182.

Higdon, D. et al. (2004). 'Combining Field Data and Computer Simulations for Calibration and Prediction'. *SIAM Journal on Scientific Computing* 26.2, pp. 448–466. DOI: 10.1137/S1064827503426693.

Higdon, D. et al. (2008). 'Computer Model Calibration Using High-Dimensional Output'. *Journal of the American Statistical Association* 103.482, pp. 570–583. DOI: 10.1198/016214507000000888.

Hinson, S. and Bolton, P. (2022). *Fuel Poverty*.

HMG (2002). *Approved Document L1: Conservation of Fuel and Power in Dwellings*.

HMG (2006). *Approved Document L1: Conversation of Fuel and Power in New Dwellings*. Office of the Deputy Prime Miniser.

HMG (2016). *Approved Document L1A. Conservation of Fuel and Power in New Dwellings. 2013 Edition Incorporating 2016 Amendments.* London: NBS.

HMG (2021a). *Approved Document O. Requirement O1: Overheating Mitigation. Regulations: 40B*. HM Government.

HMG (2021b). *Heat and Buildings Strategy*. HM Government.

HMG (2021c). *Net Zero Strategy: Build Back Greener*. HM Government.

HMSO (1995). *Approved Document L1: Conservation of Fuel and Power*.

Hong, S. H. et al. (–29th Aug. 2004). 'THE IMPACT OF ENERGY EFFICIENT REFURBISHMENT ON THE AIRTIGHTNESS IN ENGLISH DWELLINGS'. *25th AVIC Conference*. Prague, Czech Republic, p. 7.

Hou, D., Hassan, I. G. and Wang, L. (2021). 'Review on Building Energy Model Calibration by Bayesian Inference'. *Renewable and Sustainable Energy Reviews* 143, p. 110930. DOI: 10.1016/j.rser.2021.110930.

Huard, D. and Mailhot, A. (2006). 'A Bayesian Perspective on Input Uncertainty in Model Calibration: Application to Hydrological Model "Abc"'. *Water Resources Research* 42.7. DOI: 10.1029/2005WR004661.

Hughes, I. and Hase, T. (2014). *Measurements and Their Uncertainties: A Practical Guide to Modern Error Analysis.* Oxford: OUP Oxford.

Hulme, J., Beaunmont, A. and Summers, C. (2013a). *Energy Follow-Up Survey 2011, Report 11: Methodology*. Building Research Establishment on behalf of the Department of Energy and Climate Change.

Hulme, J., Beaunmont, A. and Summers, C. (2013b). *Energy Follow-Up Survey 2011, Report 7: Thermal Comfort & Overheating*. Watford, UK: Building Research Establishment.

Hulme, J. and Doran, S. (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).

Hung, Y., Joseph, V. R. and Melkote, S. N. (2015). 'Analysis of Computer Experiments With Functional Response'. *Technometrics* 57.1, pp. 35–44. DOI: 10.1080/00401706.2013.869263.

IEA (2021). *Net Zero by 2050 - A Roadmap for the Global Energy Sector*. International Energy Agency, p. 224.

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

IPCC (2018). *Global Warming of 1.5°C. An IPCC Special Report on the Impacts of Global Warming of 1.5°C above Pre-Industrial Levels and Related Global*

Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty. Intergovernmental Panel on Climate Change.

IPCC (2021). 'Summary for Policymakers'. *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.

IPCC (2022). 'Summary for Policymakers'. *Climate Change 2022: Impacts, Adaptation and Vulnerability.* Intergovernmental Panel on Climate Change.

Iwanaga, T., Usher, W. and Herman, J. (2022). 'Toward SALib 2.0: Advancing the Accessibility and Interpretability of Global Sensitivity Analyses'. *Socio-Environmental Systems Modelling* 4, p. 18155. DOI: 10.18174/sesmo.18155.

Jain, N. et al. (2020). 'Cross-Sectoral Assessment of the Performance Gap Using Calibrated Building Energy Performance Simulation'. *Energy and Buildings* 224, p. 110271. DOI: 10.1016/j.enbuild.2020.110271.

Johnson, H. et al. (2005). 'The Impact of the 2003 Heat Wave on Daily Mortality in England and Wales and the Use of Rapid Weekly Mortality Estimates'. *Euro Surveill* 10.7, pp. 168–171. pmid: 16088043.

Joshi, S. S. et al. (2016). 'The Importance of Temperature and Thermoregulation for Optimal Human Sleep'. *Energy and Buildings* 131, pp. 153–157. DOI: http://dx.doi.org/10.1016/j.enbuild.2016.09.020.

Kavgic, M. et al. (2010). 'A Review of Bottom-up Building Stock Models for Energy Consumption in the Residential Sector'. *Building and Environment* 45.7, pp. 1683–1697. DOI: 10.1016/j.buildenv.2010.01.021.

Kendon, M. et al. (2020). 'State of the UK Climate 2019'. *International Journal of Climatology* 40.S1, pp. 1–69. DOI: 10.1002/joc.6726.

Kendon, M. et al. (2021). 'State of the UK Climate 2020'. *International Journal of Climatology* 41.S2, pp. 1–76. DOI: 10.1002/joc.7285.

Kennedy, M. and O'Hagan, A. (2001). 'Bayesian Calibration of Computer Models'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3, pp. 425–464. DOI: 10.1111/1467-9868.00294.

Kiureghian, A. D. and Ditlevsen, O. (2009). 'Aleatory or Epistemic? Does It Matter?' *Structural Safety* 31.2, pp. 105–112. DOI: 10.1016/j.strusafe.2008.06.020.

Kneifel, J. (2012). *Annual Whole Building Energy Simulation of the NIST Net Zero Energy Residential Test Facility Design*.

Kougionis, T. (2018). 'Overheating in Residential Properties — Define, Identify and Prevent : An Overview'. *Journal of Building Survey, Appraisal & Valuation* 6.4, pp. 363–371.

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

Kristensen, M. H., Choudhary, R. and Petersen, S. (2017a). 'Bayesian Calibration of Building Energy Models: Comparison of Predictive Accuracy Using Metered Utility Data of Different Temporal Resolution'. *Energy Procedia*. CISBAT 2017 International ConferenceFuture Buildings & Districts – Energy Efficiency from Nano to Urban Scale 122 (Supplement C), pp. 277–282. DOI: 10.1016/j.egypro.2017.07.322.

Kristensen, M. H., Hedegaard, R. E. and Petersen, S. (2018). 'Hierarchical Calibration of Archetypes for Urban Building Energy Modeling'. *Energy and Buildings* 175, pp. 219–234. DOI: 10.1016/j.enbuild.2018.07.030.

Kristensen, M. H. and Petersen, S. (2016). 'Choosing the Appropriate Sensitivity Analysis Method for Building Energy Model-Based Investigations'. *Energy and Buildings* 130, pp. 166–176. DOI: 10.1016/j.enbuild.2016.08.038.

Kristensen, M. H. et al. (2017b). 'Bayesian Calibration Of Residential Building Clusters Using A Single Geometric Building Representation', p. 10.

Lan, L. et al. (2011). 'Effects of Thermal Discomfort in an Office on Perceived Air Quality, SBS Symptoms, Physiological Responses, and Human Performance'. *Indoor Air* 21.5, pp. 376–390. DOI: 10.1111/j.1600-0668.2011.00714.x.

Lan, L. et al. (2017). 'Thermal Environment and Sleep Quality: A Review'. *Energy and Buildings* 149, pp. 101–113. DOI: 10.1016/j.enbuild.2017.05.043.

Lee, K. and Lee, D. (2015). 'The Relationship Between Indoor and Outdoor Temperature in Two Types Of Residence'. *Energy Procedia*. 6th International Building Physics Conference, IBPC 2015 78, pp. 2851–2856. DOI: 10.1016/j.egypro.2015.11.647.

Li, Q., Augenbroe, G. and Brown, J. (2016). 'Assessment of Linear Emulators in Lightweight Bayesian Calibration of Dynamic Building Energy Models for Parameter Estimation and Performance Prediction'. *Energy and Buildings* 124, pp. 194–202. DOI: 10.1016/j.enbuild.2016.04.025.

Lim, H. and Zhai, Z. J. (2017a). 'Comprehensive Evaluation of the Influence of Meta-Models on Bayesian Calibration'. *Energy and Buildings* 155, pp. 66–75. DOI: 10.1016/j.enbuild.2017.09.009.

Lim, H. and Zhai, Z. J. (2017b). 'Review on Stochastic Modeling Methods for Building Stock Energy Prediction'. *Building Simulation* 10.5, pp. 607–624. DOI: 10.1007/s12273-017-0383-y.

Liu, F., Bayarri, M. J. and Berger, J. O. (2009). 'Modularization in Bayesian Analysis, with Emphasis on Analysis of Computer Models'. *Bayesian Analysis* 4.1, pp. 119–150. DOI: 10.1214/09-BA404.

Loga, T., Stein, B. and Diefenbach, N. (2016). 'TABULA Building Typologies in 20 European Countries—Making Energy-Related Features of Residential Building Stocks Comparable'. *Energy and Buildings*. Towards an Energy Efficient European Housing Stock: Monitoring, Mapping and Modelling Retrofitting Processes 132, pp. 4–12. DOI: 10.1016/j.enbuild.2016.06.094.

Lomas, K. J. and Kane, T. (2013). 'Summertime Temperatures and Thermal Comfort in UK Homes'. *Building Research & Information* 41.3, pp. 259–280. DOI: 10.1080/09613218.2013.757886.

Lomas, K. J. and Porritt, S. M. (2017). 'Overheating in Buildings: Lessons from Research'. *Building Research & Information* 45.1-2, pp. 1–18. DOI: 10.1080/09613218.2017.1256136.

Lomas, K. J. et al. (2021). 'Dwelling and Household Characteristics' Influence on Reported and Measured Summertime Overheating: A Glimpse of a Mild Climate in the 2050's'. *Building and Environment* 201, p. 107986. DOI: 10.1016/j.buildenv.2021.107986.

Lowe, J. A. et al. (2018). *UKCP18 Science Overview Report*, p. 73.

Mangiafico, S. S. (2016). *Summary and Analysis of Extension Program Evaluation in R*.

Mantesi, E. et al. (2018). 'The Modelling Gap: Quantifying the Discrepancy in the Representation of Thermal Mass in Building Simulation'. *Building and Environment* 131, pp. 74–98. DOI: 10.1016/j.buildenv.2017.12.017.

Mata, É., Sasic Kalagasidis, A. and Johnsson, F. (2014). 'Building-Stock Aggregation through Archetype Buildings: France, Germany, Spain and the UK'. *Building and Environment* 81, pp. 270–282. DOI: 10.1016/j.buildenv.2014.06.013.

Mavrogianni, A. et al. (2009). 'Space Heating Demand and Heatwave Vulnerability: London Domestic Stock'. *Building Research & Information* 37.5-6, pp. 583–597. DOI: 10.1080/09613210903162597.

Mavrogianni, A. et al. (2010). 'LONDON HOUSING AND CLIMATE CHANGE: Impact on Comfort and Health - Preliminary Results of a Summer Overheating Study'. *Open House International* 35.2, pp. 49–59.

Mavrogianni, A. et al. (2012). 'Building Characteristics as Determinants of Propensity to High Indoor Summer Temperatures in London Dwellings'. *Building and Environment*. Implications of a Changing Climate for Buildings 55, pp. 117–130. DOI: 10.1016/j.buildenv.2011.12.003.

Mavrogianni, A. et al. (2014). 'The Impact of Occupancy Patterns, Occupant-Controlled Ventilation and Shading on Indoor Overheating Risk in Domestic Environments'. *Building and Environment* 78, pp. 183–198.

Mavrogianni, A. et al. (2016). 'Inhabitant Actions and Summer Overheating Risk in London Dwellings'. *Building Research & Information* 45.1-2, pp. 119–142.

McCartney, K. J. and Nicol, J. F. (2002). 'Developing an Adaptive Control Algorithm for Europe'. *Energy and Buildings*. Special Issue on Thermal Comfort Standards 34.6, pp. 623–635. DOI: 10.1016/S0378-7788(02)00013-0.

McClarren, R. G. (2018). *Uncertainty Quantification and Predictive Computational Science: A Foundation for Physical Scientists and Engineers*. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-99525-0.

McDonald, J. H. (2014). *Handbook of Biological Statistics*. 3rd. Baltimore, Maryland: Sparky House Publishing.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor and Francis, CRC Press.

McLeod, R. S., Hopfe, C. J. and Kwan, A. (2013). 'An Investigation into Future Performance and Overheating Risks in Passivhaus Dwellings'. *Building and Environment* 70, pp. 189–209. DOI: 10.1016/j.buildenv.2013.08.024.

McLeod, R. S. and Swainson, M. (2017). 'Chronic Overheating in Low Carbon Urban Developments in a Temperate Climate'. *Renewable and Sustainable Energy Reviews* 74, pp. 201–220. DOI: 10.1016/j.rser.2016.09.106.

Meinke, A. et al. (2017). 'Comfort-Related Feedforward Information: Occupants' Choice of Cooling Strategy and Perceived Comfort'. *Building Research & Information* 45.1-2, pp. 222–238. DOI: 10.1080/09613218.2017.1233774.

Menberg, K., Heo, Y. and Choudhary, R. (2019). 'Influence of Error Terms in Bayesian Calibration of Energy System Models'. *Journal of Building Performance Simulation* 12.1, pp. 82–96. DOI: 10.1080/19401493.2018.1475506.

Met Office (2018a). *Land Projection Maps: Probabilistic Projections*. Met Office. URL: https://www.metoffice.gov.uk/research/approach/collaboration/ukcp/land-projection-maps (visited on 22/03/2022).

Met Office (2018b). *Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (1853-Current)*. URL: http://catalogue.

ceda . ac . uk / uuid / 220a65615218d5c9cc9e4785a3234bd0 (visited on 15/01/2019).

Met Office (2021). *UK Climate Projections: Headline Findings*.

Met Office (2022a). *Record High Temperatures Verified*. Met Office. URL: https://www.metoffice.gov.uk/about-us/press-office/news/weather-and-climate/2022/record-high-temperatures-verified (visited on 11/09/2022).

Met Office (2022b). *What Is a Heatwave?* Met Office. URL: https://www.metoffice.gov.uk/weather/learn-about/weather/types-of-weather/temperature/heatwave (visited on 14/05/2022).

MHCLG (2019). *English Housing Survey: Private Rented Sector, 2017-18*. Ministry of Housing, Communities and Local Government.

Morris, M. D. (1991). 'Factorial Sampling Plans for Preliminary Computational Experiments'. *Technometrics* 33.2, pp. 161–174. DOI: 10.2307/1269043. JSTOR: 1269043.

Mulligan, M. and Wainwright, J. (2013). 'Modelling and Model Building'. *Environmental Modelling: Finding Simplicity in Complexity: Second Edition*. John Wiley and Sons, pp. 7–26.

Mun, J. (2012). 'Understanding and Choosing the Right Probability Distributions'. *Advanced Analytical Models*. John Wiley & Sons, Ltd, pp. 899–917. DOI: 10.1002/9781119197096.app03.

Nguyen, J. L., Schwartz, J. and Dockery, D. W. (2014). 'The Relationship between Indoor and Outdoor Temperature, Apparent Temperature, Relative Humidity, and Absolute Humidity'. *Indoor air* 24.1, pp. 103–112. DOI: 10.1111/ina.12052. pmid: 23710826.

NHBC Foundation (2019). *House Building: A Century of Innovation. Technical Advances in Conventional Construction.* NHBC Foundation, p. 24.

Nicol, F. and Humphreys, M. (2002). 'Adaptive Thermal Comfort and Sustainable Thermal Standards for Buildings'. *Energy and Buildings*. Special Issue

on Thermal Comfort Standards 34.6, pp. 563–572. DOI: 10.1016/S0378-7788(02)00006-3.

Nicol, F. and Humphreys, M. (2010). 'Derivation of the Adaptive Equations for Thermal Comfort in Free-Running Buildings in European Standard EN15251'. *Building and Environment*. International Symposium on the Interaction between Human and Building Environment Special Issue Section 45.1, pp. 11–17. DOI: 10.1016/j.buildenv.2008.12.013.

Nicol, F. and Humphreys, M. (2018). 'Room Temperature during Sleep'. *10th Windsor Conference - Rethinking Comfort*. Windsor Conference, p. 6.

O' Donovan, A., O' Sullivan, P. D. and Murphy, M. D. (2019). 'Predicting Air Temperatures in a Naturally Ventilated Nearly Zero Energy Building: Calibration, Validation, Analysis and Approaches'. *Applied Energy* 250, pp. 991–1010. DOI: 10.1016/j.apenergy.2019.04.082.

O'Hagan, A. (2006). 'Bayesian Analysis of Computer Code Outputs: A Tutorial'. *Reliability Engineering & System Safety*. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004) 91.10, pp. 1290–1300. DOI: 10.1016/j.ress.2005.11.025.

Oikonomou, E. et al. (2012). 'Modelling the Relative Importance of the Urban Heat Island and the Thermal Quality of Dwellings for Overheating in London'. *Building and Environment* 57, pp. 223–238.

Oikonomou, E. et al. (2018). *English Archetypes*. URL: https://www.ucl.ac.uk/energy-models/models/english-archetypes (visited on 22/06/2022).

Okamoto-Mizuno, K. and Mizuno, K. (2012). 'Effects of Thermal Environment on Sleep and Circadian Rhythm'. *Journal of Physiological Anthropology* 31.1, p. 14. DOI: 10.1186/1880-6805-31-14. pmid: 22738673.

Oraiopoulos, A. and Howard, B. (2022). 'On the Accuracy of Urban Building Energy Modelling'. *Renewable and Sustainable Energy Reviews* 158, p. 111976. DOI: 10.1016/j.rser.2021.111976.

Oreszczyn, T. et al. (2006). 'Determinants of Winter Indoor Temperatures in Low Income Households in England'. *Energy and Buildings* 38.3, pp. 245–252. DOI: 10.1016/j.enbuild.2005.06.006.

Oseland, N. A. (1995). 'Predicted and Reported Thermal Sensation in Climate Chambers, Offices and Homes'. *Energy and Buildings* 23.2, pp. 105–115. DOI: 10.1016/0378-7788(95)00934-5.

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, pp. 2387–2399. DOI: 10.1016/j.buildenv.2010.04.011.

Pathan, A. et al. (2017). 'Monitoring Summer Indoor Overheating in the London Housing Stock'. *Energy and Buildings* 141, pp. 361–378. DOI: 10.1016/j.enbuild.2017.02.049.

Peacock, A. D., Jenkins, D. P. and Kane, D. (2010). 'Investigating the Potential of Overheating in UK Dwellings as a Consequence of Extant Climate Change'. *Energy Policy*. Large-Scale Wind Power in Electricity Markets with Regular Papers 38.7, pp. 3277–3288. DOI: 10.1016/j.enpol.2010.01.021.

Perera, E. and Parkins, L. (1992). 'Airtightness of UK Buildings: Status and Future Possibilities'. 2.2.

Petersen, S., Kristensen, M. H. and Knudsen, M. D. (2019). 'Prerequisites for Reliable Sensitivity Analysis of a High Fidelity Building Energy Model'. *Energy and Buildings* 183, pp. 1–16. DOI: 10.1016/j.enbuild.2018.10.035.

Petrou, G. (2021). *Distrmultifit*. https://github.com/giorgospetrou/distrmultifit.

Petrou, G. et al. (2019a). 'Can the Choice of Building Performance Simulation Software Significantly Alter the Level of Predicted Indoor Overheating Risk in London Flats?' *Building Services Engineering Research and Technology* 40.1, pp. 30–46. DOI: 10.1177/0143624418792340.

Petrou, G. et al. (2019b). '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, pp. 492–511. DOI: 10.1177/0143624419847621.

Petrou, G. et al. (2021a). 'A Case Study on the Impact of Fixed Input Parameter Values in the Modelling of Indoor Overheating'. *Journal of Physics: Conference Series* 2069.1, p. 012137. DOI: 10.1088/1742-6596/2069/1/012137.

Petrou, G. et al. (2021b). 'Beyond Normal: Guidelines on How to Identify Suitable Model Input Distributions for Building Performance Analysis'. Building Simulation 2021. Bruges, Belgium.

Pisello, A. L. et al. (2018). 'Facing the Urban Overheating: Recent Developments. Mitigation Potential and Sensitivity of the Main Technologies'. *WIREs Energy and Environment* 7.4, e294. DOI: 10.1002/wene.294.

Pollock, D. (2013). 'On Kronecker Products, Tensor Products and Matrix Differential Calculus'. *International Journal of Computer Mathematics* 90.11, pp. 2462–2476. DOI: 10.1080/00207160.2013.783696.

Pompe, E., Holmes, C. and Łatuszyński, K. (2020). 'A Framework for Adaptive MCMC Targeting Multimodal Distributions'. *The Annals of Statistics* 48.5, pp. 2930–2952. DOI: 10.1214/19-AOS1916.

Portet, S. (2020). 'A Primer on Model Selection Using the Akaike Information Criterion'. *Infectious Disease Modelling* 5, pp. 111–128. DOI: 10.1016/j.idm.2019.12.010. pmid: 31956740.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. manual. R Foundation for Statistical Computing. Vienna, Austria.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press. 248 pp.

Reimann, C. et al. (2008). *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. Chichester, England ; Hoboken, NJ: John Wiley & Sons. 343 pp.

Reinhart, C. F. and Cerezo Davila, C. (2016). 'Urban Building Energy Modeling – A Review of a Nascent Field'. *Building and Environment* 97 (Supplement C), pp. 196–202. DOI: 10.1016/j.buildenv.2015.12.001.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis*. Third Edition. Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley. 758 pp.

Rijal, H. B. et al. (2007). 'Using Results from Field Surveys to Predict the Effect of Open Windows on Thermal Comfort and Energy Use in Buildings'. *Energy and Buildings*. Comfort and Energy Use in Buildings - Getting Them Right 39.7, pp. 823–836. DOI: 10.1016/j.enbuild.2007.02.003.

Roberts, B. M. et al. (2019). 'Predictions of Summertime Overheating: Comparison of Dynamic Thermal Models and Measurements in Synthetically Occupied Test Houses'. *Building Services Engineering Research and Technology* 40.4, pp. 512–552. DOI: 10.1177/0143624419847349.

Robine, J.-M. et al. (2008). 'Death Toll Exceeded 70,000 in Europe during the Summer of 2003'. *Comptes Rendus Biologies*. Dossier : Nouveautés En Cancérogenèse / New Developments in Carcinogenesis 331.2, pp. 171–178. DOI: 10.1016/j.crvi.2007.12.001.

Robinson, D. et al. (–30th July 2009). 'CITYSIM: COMPREHENSIVE MICRO-SIMULATION OF RESOURCE FLOWS FOR SUSTAINABLE URBAN PLANNING'. *Building Simulation 2009: Eleventh International IBPSA Conference*. Glasgow, Scotland, p. 8.

Ruiz, G. R. and Bandera, C. F. (2017). 'Validation of Calibrated Energy Models: Common Errors'. *Energies* 10.10, p. 1587. DOI: 10.3390/en10101587.

Rupp, R. F., Vásquez, N. G. and Lamberts, R. (2015). 'A Review of Human Thermal Comfort in the Built Environment'. *Energy and Buildings* 105, pp. 178–205. DOI: 10.1016/j.enbuild.2015.07.047.

Saltelli, A., ed. (2004). *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Hoboken, NJ: Wiley. 219 pp.

Saltelli, A. et al., eds. (2008). *Global Sensitivity Analysis: The Primer*. Chichester, England ; Hoboken, NJ: John Wiley. 292 pp.

Schweizer, C. et al. (2007). 'Indoor Time–Microenvironment–Activity Patterns in Seven Regions of Europe'. *Journal of Exposure Science & Environmental Epidemiology* 17.2 (2), pp. 170–181. DOI: 10.1038/sj.jes.7500490.

Shrubsole, C. et al. (2014). '100 Unintended Consequences of Policies to Improve the Energy Efficiency of the UK Housing Stock'. *Indoor and Built Environment* 23.3, pp. 340–352. DOI: 10.1177/1420326X14524586.

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Chapman and Hall/CRC.

Smith, R. C. (2013). *Uncertainty Quantification: Theory, Implementation, and Applications*. Computational Science and Engineering Series. Philadelphia: Society for Industrial and Applied Mathematics. 382 pp.

Snell, C., Bevan, M. and Thomson, H. (2015). 'Justice, Fuel Poverty and Disabled People in England'. *Energy Research & Social Science* 10, pp. 123–132. DOI: 10.1016/j.erss.2015.07.012.

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

Stephen, R. K. (1998). *Airtightness in UK Dwellings: BRE's Test Results and Their Significance*. BRE Reports BR : 359. London: CRC. 38 pp.

Stephen, R. (2000). *Airtightness in UK Dwellings*. Bracknell, UK: BRE press.

Symonds, P. et al. (2016). 'Development of an England-wide Indoor Overheating and Air Pollution Model Using Artificial Neural Networks'. *Journal of Building Performance Simulation* 9.6, pp. 606–619. DOI: 10.1080/19401493.2016.1166265.

Symonds, P. et al. (2017). 'Overheating in English Dwellings: Comparing Modelled and Monitored Large-Scale Datasets'. *Building Research & Information* 45.1-2, pp. 195–208.

TABULA Project Team (2012). *Typology Approach for Building Stock Energy Assessment. Main Results of the TABULA Project*.

TABULA Project Team (2017). *TABULA WebTool*. URL: https://webtool.building-typology.eu/#bm (visited on 30/05/2022).

Tardioli, G. et al. (2018). 'Identification of Representative Buildings and Building Groups in Urban Datasets Using a Novel Pre-Processing, Classification, Clustering and Predictive Modelling Approach'. *Building and Environment* 140, pp. 90–106. DOI: 10.1016/j.buildenv.2018.05.035.

Tardioli, G. et al. (2020). 'A Methodology for Calibration of Building Energy Models at District Scale Using Clustering and Surrogate Techniques'. *Energy and Buildings* 226, p. 110309. DOI: 10.1016/j.enbuild.2020.110309.

Taylor, J. et al. (2014). 'The Relative Importance of Input Weather Data for Indoor Overheating Risk Assessment in Dwellings'. *Building and Environment* 76, pp. 81–91.

Taylor, J. et al. (2015). 'Mapping the Effects of Urban Heat Island, Housing, and Age on Excess Heat-Related Mortality in London'. *Urban Climate* 14, pp. 517–528. DOI: 10.1016/j.uclim.2015.08.001.

Taylor, J. et al. (2016). 'Mapping Indoor Overheating and Air Pollution Risk Modification across Great Britain: A Modelling Study'. *Building and Environment* 99, pp. 1–12. DOI: 10.1016/j.buildenv.2016.01.010.

Taylor, J. et al. (2018a). 'Comparison of Built Environment Adaptations to Heat Exposure and Mortality during Hot Weather, West Midlands Region, UK'. *Environment International* 111, pp. 287–294. DOI: 10.1016/j.envint.2017.11.005.

Taylor, J. et al. (2018b). 'Estimating the Influence of Housing Energy Efficiency and Overheating Adaptations on Heat-Related Mortality in the West Midlands, UK'. *Atmosphere* 9.

Taylor, J. et al. (2019). 'Application of an Indoor Air Pollution Metamodel to a Spatially-Distributed Housing Stock'. *Science of The Total Environment* 667, pp. 390–399. DOI: 10.1016/j.scitotenv.2019.02.341.

Taylor, J. et al. (2021). 'Projecting the Impacts of Housing on Temperature-Related Mortality in London during Typical Future Years'. *Energy and Buildings* 249, p. 111233. DOI: 10.1016/j.enbuild.2021.111233.

Tempcon Instrumentation Ltd (2022). *HOBO UA-001-08 Pendant Waterproof Temperature Logger (8K Byte Memory)*. URL: https://www.tempcon.co.uk/hobo-ua-001-08-8k-pendant-temp-logger-ua-001-08 (visited on 26/08/2022).

The Building Regulations Etc. (Amendment) (England) Regulations 2021 (2022).

The Climate Change Act 2008 (2050 Target Amendment) Order 2019 (n.d.).

Tian, W. (2013). 'A Review of Sensitivity Analysis Methods in Building Energy Analysis'. *Renewable and Sustainable Energy Reviews* 20, pp. 411–419.

Tian, W. and Choudhary, R. (2012). 'A Probabilistic Energy Model for Non-Domestic Building Sectors Applied to Analysis of School Buildings in Greater London'. *Energy and Buildings* 54, pp. 1–11. DOI: 10.1016/j.enbuild.2012.06.031.

Tian, W. et al. (2018). 'A Review of Uncertainty Analysis in Building Energy Assessment'. *Renewable and Sustainable Energy Reviews* 93, pp. 285–301. DOI: 10.1016/j.rser.2018.05.029.

UCL IEDE (2016). *Health Protection Research Unit (HPRU) in Environmental Change and Health*. UCL Institute for Environmental Design and Engineering. URL: https://www.ucl.ac.uk/bartlett/environmental-design/research-projects/2020/nov/health-protection-research-unit-hpru-environmental-change-and-health (visited on 03/06/2022).

UCL IEDE (2017). *AWESOME*. UCL Institute for Environmental Design and Engineering. URL: https://www.ucl.ac.uk/bartlett/environmental-design/research-projects/2020/nov/awesome (visited on 03/06/2022).

UNFCCC (2015). *Adoption of the Paris Agreement*. FCCC/CP/2015/L.9/Rev.1. 21st Conference of the Parties, Paris: United Nations.

Vellei, M. et al. (2017). 'Overheating in Vulnerable and Non-Vulnerable Households'. *Building Research & Information* 45.1-2, pp. 102–118. DOI: 10.1080/09613218.2016.1222190.

Vicedo-Cabrera, A. M. et al. (2021). 'The Burden of Heat-Related Mortality Attributable to Recent Human-Induced Climate Change'. *Nature Climate Change* 11.6 (6), pp. 492–500. DOI: 10.1038/s41558-021-01058-x.

Wang, C.-K. et al. (2020). 'Bayesian Calibration at the Urban Scale: A Case Study on a Large Residential Heating Demand Application in Amsterdam'. *Journal of Building Performance Simulation* 13.3, pp. 347–361. DOI: 10.1080/19401493.2020.1729862.

WHO (2009). *Improving Public Health Responses to Extreme Weather/Heat-Waves - EuroHEAT*. World Health Organisation (WHO).

WHO (2018). *Housing and Health Guidelines*. Copenhagen: World Health Organization. Regional Office for Europe.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wild, C. (2006). 'The Concept of Distribution'. *Statistics Education Research Journal* 5.2, p. 17.

Wilkinson, P. et al. (2001). *Cold Comfort: The Social and Environmental Determinants of Excess Winter Death in England, 1986–1996*. Joseph Rowntree Foundation.

Williams, B. et al. (2006). 'Combining Experimental Data and Computer Simulations, with an Application to Flyer Plate Experiments'. *Bayesian Analysis* 1.4, pp. 765–792. DOI: 10.1214/06-BA125.

Yun, J. et al. (2020). 'Stochastic Approximation Hamiltonian Monte Carlo'. *Journal of Statistical Computation and Simulation* 90.17, pp. 3135–3156. DOI: 10.1080/00949655.2020.1797031.

ZCH (2015). *Overheating in Homes, The Big Picture, Full Report*. Zero Carbon Housing.

Appendix A

Theory of Gaussian Processes

A *Gaussian Process* (GP) may be formally defined as *a collection of random variables, any finite number of which have a joint Gaussian distribution* and may be written as (Rasmussen and Williams, 2006):

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{A.1}$$

From A.1, a GP is fully defined by its mean ($m(\mathbf{x})$) and covariance function ($k(\mathbf{x}, \mathbf{x}')$). The random variable corresponds to the value of function $f(\mathbf{x})$ at location $\mathbf{x}$. $\mathbf{x}'$ is another realisation of $\mathbf{x}$.

The mean function may take any form that is appropriate for the problem, often that of a linear function, or even more commonly that of a function that returns a vector of constants, such as zeros (Higdon et al., 2004; O'Hagan, 2006; Higdon et al., 2008). Similarly, the covariance function should also be selected based on the characteristics of the problem being studied and several options exist. A frequently used option, is the *squared exponential* (may also be referred to as *Gaussian correlation function* or *Radial Basis Function*) (Rasmussen and Williams, 2006):

$$cov(f(x_p), f(x_q)) = k(x_p, x_q) = \sigma_f^2 \exp(-\frac{1}{2l^2}|x_p - x_q|^2), \tag{A.2}$$

where σ_f^2 is the signal variance and l is the length-scale. Parameters that are free to be varied and are used to define the Gaussian Process, such as σ_f^2 and l, may be collectively referred to as *hyperparameters* (Rasmussen and Williams, 2006). It

should be highlighted that the covariance between outputs is written as function of the inputs.

The most commonly used covariance function within the field of Bayesian calibration of computer models is the following (Arendt et al., 2012; Chong and Menberg, 2018):

$$cov(f(x_p), f(x_q)) = k(x_p, x_q) = \frac{1}{\lambda} \exp(-\beta |x_p - x_q|^{\alpha}), \tag{A.3}$$

where $\frac{1}{\lambda}$ is the precision hyperparameter, β is the correlation hyperparameter and α is the smoothness hyperparameter. By fixing $\alpha = 2$, it is assumed that the function modelled by the GP is smooth and infinitely differentiable. Doing so is common practice in Bayesian calibration applications (Chong and Menberg, 2018; Kristensen et al., 2017b; Menberg et al., 2019), and this results in a covariance function that is equivalent to Equation A.2.

A.1 Prediction with Noise-free Observations

Assuming a set of noise-free observations $\{(x_i, f_i) | i = 1, \ldots, n\}$, it is possible to make predictions about the joint distribution of the training data $\mathbf{f}$ and test outputs $\mathbf{f}_*$ according to the following GP prior (Rasmussen and Williams, 2006):

$$\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{x}, \mathbf{x}) & \mathbf{K}(\mathbf{x}, \mathbf{x}_*) \\ \mathbf{K}(\mathbf{x}_*, \mathbf{x}) & \mathbf{K}(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right) \tag{A.4}$$

For n training points ($\mathbf{x}$) and n_* test points ($\mathbf{x}_*$), $\mathbf{K}(\mathbf{x}, \mathbf{x}_*)$ is the $n \times n_*$ matrix of covariances for all pairs of training and testing. By *conditioning* the joint GP on the observations, the posterior distribution contains only functions which agree with the observations. This results in posterior prediction provided by the following equation (Rasmussen and Williams, 2006):

$$\begin{aligned} f(\mathbf{x}_*) | \mathbf{x}_*, \mathbf{x}, f(\mathbf{x}) \sim \mathcal{N}(&\mathbf{K}(\mathbf{x}_*, \mathbf{x}) \mathbf{K}(\mathbf{x}, \mathbf{x})^{-1} f(\mathbf{x}), \\ &\mathbf{K}(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{K}(\mathbf{x}_*, \mathbf{x}) \mathbf{K}(\mathbf{x}, \mathbf{x})^{-1} \mathbf{K}(\mathbf{x}, \mathbf{x}_*)) \end{aligned} \tag{A.5}$$

A.2 Prediction with Noisy Observations

It is more realistic to assume that observations in real-world application are noisy:

$$\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\varepsilon}. \tag{A.6}$$

By assuming independent and identically distributed Gaussian noise $\boldsymbol{\varepsilon}$ with variance σ_n^2, the matrix of covariances for all noisy observations becomes (Rasmussen and Williams, 2006):

$$\text{cov}(\mathbf{y}) = \mathbf{K}(\mathbf{x},\mathbf{x}) + \sigma_n^2 I \tag{A.7}$$

Because of the independence assumption, the noise is only added to the diagonal of the matrix. For notational simplification, let's assume that $\mathbf{K}(\mathbf{x},\mathbf{x}) + \sigma_n^2 I = \mathbf{K}_y$, $\mathbf{K}(\mathbf{x},\mathbf{x}_*) = \mathbf{K}_*$, $\mathbf{K}(\mathbf{x}_*,\mathbf{x}) = \mathbf{K}_*^T$ and $\mathbf{K}(\mathbf{x}_*,\mathbf{x}_*) = \mathbf{K}_{**}$. The joint distribution is written as:

$$\begin{bmatrix} \mathbf{y} \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_y & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right) \tag{A.8}$$

To posterior predictive distribution now becomes (Rasmussen and Williams, 2006):

$$f(\mathbf{x}_*)|\mathbf{x},\mathbf{y},\mathbf{x}_* \sim \mathcal{N}(\bar{f}(\mathbf{x}_*), \text{cov}(f(\mathbf{x}_*))) \tag{A.9}$$

where

$$\bar{f}(\mathbf{x}_*) = \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{y} \tag{A.10}$$

$$\text{cov}(f(\mathbf{x}_*)) = \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{K}_* \tag{A.11}$$

$\bar{f}(\mathbf{x}_*)$ is the mean of $f(\mathbf{x}_*)$.

Appendix B

UK Housing Stock Model

Table B.1: Summary of the number of rooms (inc. hallways), bedrooms, ground floor area and total volume (excl. roof) per typology used in UK-HSM. Adapted from the supplementary materials of Symonds et al. (2016).

	Rooms	Bedrooms	Ground Floor Area (m^2)	Total Volume (m^3)
End Terrace	8	3	45.0	216.0
Mid Terrace	8	3	45.0	216.0
Semi Detached	8	3	51.6	268.3
Detached	14	4	71.0	340.8
Bungalow	7	2	70.0	168.0
Converted Flat	6	2	72.3	187.9
Low-Rise Flat	5	1	51.8	134.7
High-Rise Flat	6	2	65.3	156.5

Table B.2: Algorithms assumed in UK-HSM. V signifies a UK-HSM model input that can be varied.

Algorithm/Setting	Value
Surface Convection Algorithm: Outside	TARP
Surface Convection Algorithm: Outside	TARP
Heat Balance Algorithm	ConductionTransferFunction
Zone Air Heat Balance Algorithm	ThirdOrderBackwardDifference
Timestep	6
Terrain	V
Solar Distribution	Full Exterior

Figure B.1: Floor plans of end terrace, mid-terrace, semi-detached and detached typologies specified within UK-HSM. Adapted from the supplementary materials of Symonds et al. (2016).

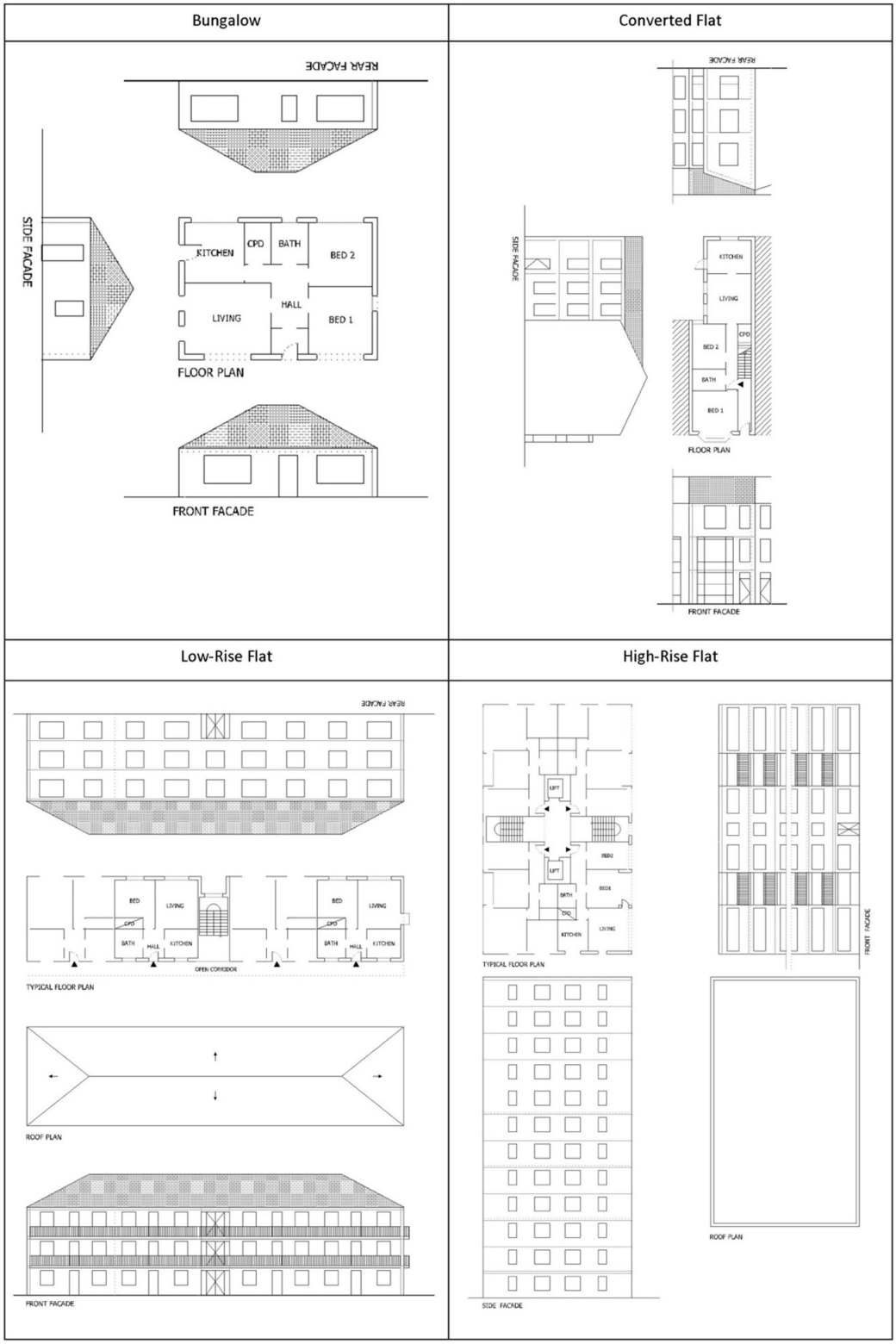


Figure B.2: Floor plans of bungalow, converted flat, low-rise flat and high-rise flat typologies specified within UK-HSM. Adapted from the supplementary materials of Symonds et al. (2016).

Table B.3: Double glazing construction details assumed in the model. V signifies a UK-HSM model input that can be varied.

Construction		Thickness	Solar Trans. at Normal Inc.	Visible Trans. at Normal Inc.	Conductivity
Double Glazing	Glass	0.01	0.775	0.881	0.9
	Gas (Air)	V	-	-	0.02485
	Glass	0.01	0.775	0.881	0.9

Table B.4: Construction and material characteristics used to model a semi-detached dwelling with filled cavity walls and double glazing. V signifies a UK-HSM model input that can be varied. For each construction, the materials are ordered from external to internal.

Construction	Material	Roughness	Thickness (m)	Conductivity (W/(mK))	Density (kg/m^3)	Specific Heat (J/(kgK))	Thermal Abs	Solar Abs	Visible Abs
Filled Cavity Wall	Brick	Rough	0.11	0.6	1712	850	V	V	V
	Cellulose	Rough	V	0.04	55	1880	0.9	0.7	0.7
	Brick	Rough	0.11	0.6	1712	850	V	V	V
	Gypsum	Rough	0.01	0.2	850	850	0.9	0.6	0.7
Internal Wall	Gypsum	Rough	0.01	0.2	850	850	0.9	0.6	0.7
	Brick	Rough	0.1	0.6	1712	850	V	V	V
	Gypsum	Rough	0.01	0.2	850	850	0.9	0.6	0.7
Floor	Concrete	Rough	0.5	1.6	2300	850	V	V	V
	Fiberglass	Rough	V	0.04	80	840	0.9	0.7	0.7
	Gypsum	Rough	0.1	0.2	850	850	0.9	0.7	0.7
Roof	Concrete	Rough	0.04	1.6	2300	850	V	V	V
	Gypsum	Rough	0.01	0.2	850	850	0.9	0.6	0.7
Suspended Wooden Floor	Gypsum	Rough	0.01	0.2	850	850	0.9	0.6	0.7
	Air gap	Rough	0.1	0.2485	1.204	1004	0.9	0.7	0.7
	Spruce	Rough	0.05	0.09	455	1500	0.9	0.7	0.7
Loft	Concrete	Rough	0.01	1.6	2300	850	V	V	V
	Spruce	Rough	0.0125	0.09	455	1500	0.9	0.7	0.7
	Air gap	Rough	V	0.02485	1.204	1004	0.9	0.7	0.7
	Gypsum	Rough	0.01	0.2	850	850	0.9	0.6	0.7
Door	Spruce	Rough	0.04	0.09	455	1500	0.9	0.7	0.7

Table B.5: Internal gain schedule for *pensioners* occupancy in UK-HSM. Adapted from the supplementary materials of Symonds et al. (2016).

						Hourly Multiplier																						
Zone	Internal Gain	Load (W)	Fraction Useful	Fraction Latent	Source	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Liv.	People	100			1	0	0	0	0	0	0	0	0	1	2	2	2	0	2	2	2	2	2	2	0	2	2	0	0
	Lighting	100			1	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0
	TV	62.2	0.734	0.16	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	0	0
	TV Standby	20	1	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	0	0	1	1
	Laptop	36.9	0.734	0.16	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
	Wireless Router	24	0.734	0.16	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Kit.	People	100			1	0	0	0	0	0	0	0	1	1	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0
	Lighting	100			1	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	Hob	3600	0.4	0.3	2	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	Oven	5100	0.4	0.3	2	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	Fridge	46	1	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Washing Machine	500	0.8	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
	Dishwasher	1090	0.6	0.15	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Bed.	People	100			1	2	2	2	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
	Lighting	100			1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	Mobile Phones	17.72	0.734	0.16	2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

1 CIBSE (2006)
2 Kneifel (2012)

Appendix C

Standardisation of Indoor Temperature: Supplementary Material

To standardise the summer Mean Daytime Living Room Temperature (MDLRT) and the Mean Nighttime Bedroom Temperature (MNBT), 12 regression models were evaluated whose performance was compared using the adjusted R^2. In addition, 12 regression models for a different temporal resolution of summer living room and bedroom temperature were also evaluated for the purpose of comparison.

The six combinations of explanatory variables tested were:

1. $T_{in} = \beta_0 + \beta_1 T_{ext}$
2. $T_{in} = \beta_0 + \beta_1 T_{ext} + \beta_2 T_{ext}^2$
3. $T_{in} = \beta_0 + \beta_1 T_{ext} + \beta_2 GHI$
4. $T_{in} = \beta_0 + \beta_1 T_{ext} + \beta_2 GHI + \beta_3 T_{ext}^2$
5. $T_{in} = \beta_0 + \beta_1 T_{ext} + \beta_2 GHI + \beta_3 T_{ext}^2 + \beta_4 GHI^2$
6. $T_{in} = \beta_0 + \beta_1 T_{ext} + \beta_2 GHI + \beta_3 T_{ext} * GHI + \beta_5 T_{ext}^2 + \beta_6 GHI^2$

where T_{in} is the indoor temperature, T_{ext} is the external temperature and GHI is the global horizontal irradiance. β_{0-6} are the regression coefficients. For each of the above models, four temporal resolutions were assessed:

a Hourly mean T_{in}, T_{ext}, GHI

b Daily mean T_{in}, T_{ext}, GHI

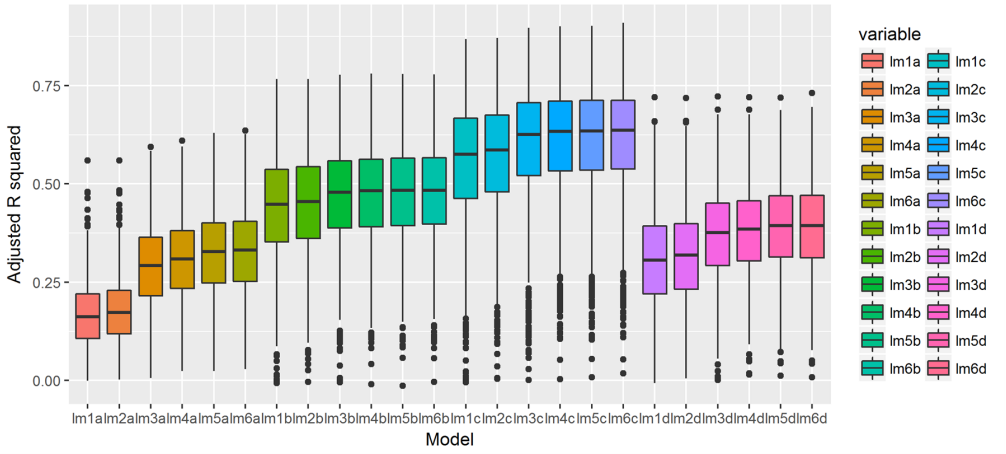


Figure C.1: Boxplots of adjusted R^2 for the 24 regression models fitted to the monitored 2011 Energy Follow-Up Survey bedroom temperature.

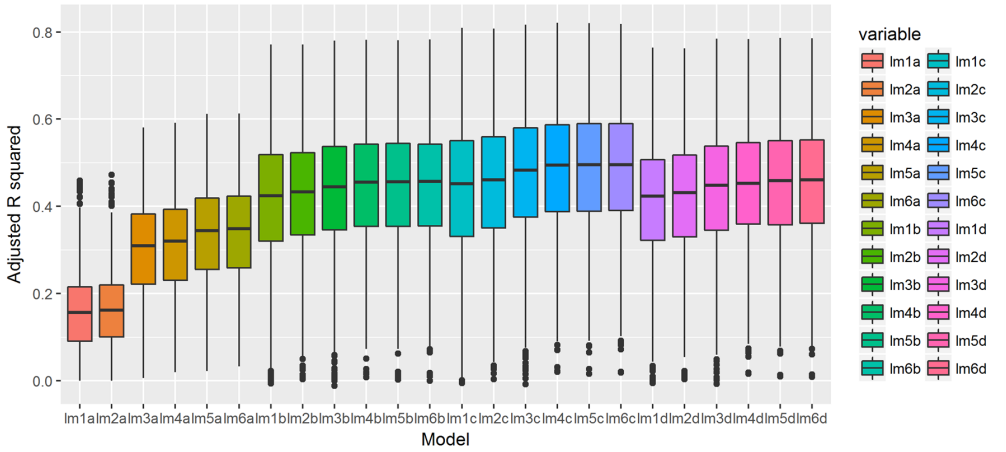


Figure C.2: Boxplots of adjusted R^2 for the 24 regression models fitted to the monitored 2011 Energy Follow-Up Survey living room temperature.

c MDLRT & MNBT with daily mean T_{ext}, GHI

d MDLRT & MNBT with daytime/nighttime mean T_{ext}, GHI

where daytime is 08:00-22:00 and nighttime is 22:00-08:00. Daytime T_{in} only refers to the living room, while nighttime T_{in} refers to the bedroom. The boxplot of adjusted R^2 for each model is plot in Figure C.1 for the bedroom and Figure C.2 for the living room.

Appendix D

Bayesian Inference & Calibration

D.1 Bayesian Inference

To capture the relationship between any observable variables $\mathbf{y}$ and the unobservable parameters $\boldsymbol{\theta}$, we start by defining a *joint probability distribution* for $\boldsymbol{\theta}$ and $\mathbf{y}$ (Gelman, 2014):

$$p(\boldsymbol{\theta},\mathbf{y}) = p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})\,, \tag{D.1}$$

where $p(\boldsymbol{\theta})$ is the *prior distribution* and $p(\mathbf{y}|\boldsymbol{\theta})$ is the *sampling (or data) distribution*. The prior distribution, $p(\boldsymbol{\theta})$, captures any prior knowledge we might have regarding the unobservable parameters $\boldsymbol{\theta}$, while the sampling distribution, $p(\mathbf{y}|\boldsymbol{\theta})$, describes the probability of observing $\mathbf{y}$ for a given value of $\boldsymbol{\theta}$. If observations of $\mathbf{y}$ have been made, $p(\mathbf{y}|\boldsymbol{\theta})$ can be regarded as a function of $\boldsymbol{\theta}$, for fixed $\mathbf{y}$, and is referred to as the *likelihood* function.

Through the conditional probability known as Bayes' rule (or Bayes' Theorem), the observations $\mathbf{y}$, may inform our knowledge of the parameters $\boldsymbol{\theta}$ (Gelman, 2014):

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta},\mathbf{y})}{p(\mathbf{y})} = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}\,, \tag{D.2}$$

where $p(\boldsymbol{\theta}|\mathbf{y})$ is the *posterior density*, which represents the relative weights of belief for each parameter value after considering the prior and likelihood function (Bolstad and Curran, 2017).

D.2 Theory of Bayesian Calibration

Note that in the equations that will follow, superscripts in brackets are used as qualifiers. Subscripts may be used as qualifiers or to index variables, and their purpose will be made clear in the text.

D.2.1 Computationally cheap computer model

Following the Bayesian perspective to statistical calibration and given eq. 2.2, the aim is to identify the appropriate likelihood function and parameter prior distributions to compute the posterior distribution, as per eq. D.2. The likelihood function will depend on the assumption made about the distribution of error terms (ε_i). A commonly used assumption is that ε_i is an independently and identically distributed (*iid*) variable (Higdon et al., 2004; Smith, 2013; Gelman, 2014), following a normal distribution with a mean of 0 and variance of σ_ε^2. Based on the *iid* assumption, the likelihood function, $\mathcal{L}(\mathbf{y}|\boldsymbol{\eta})$, takes the form (Higdon et al., 2004; Smith, 2013):

$$\mathcal{L}(\mathbf{y}|\boldsymbol{\eta}) \propto \exp\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\eta})^T \mathbf{K}_y^{-1}(\mathbf{y}-\boldsymbol{\eta})\}, \tag{D.3}$$

where $\mathbf{y} = (y(\boldsymbol{w}_1),\ldots,y(\boldsymbol{w}_n))^T$ and $\boldsymbol{\eta} = (\eta(\boldsymbol{w}_1,\theta),\ldots,\eta(\boldsymbol{w}_n,\theta))^T$. $\mathbf{K}_y = \sigma_\varepsilon^2 \cdot \mathbf{I}_{n\times n}$ is the observation covariance matrix, with $\mathbf{I}_{n\times n}$ being the identity matrix of size $n \times n$. Assuming *a priori* that $\boldsymbol{\theta}$ follows a distribution $p(\boldsymbol{\theta})$, the *unnormalised* posterior density is then defined as:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \mathcal{L}(\mathbf{y}|\boldsymbol{\eta})) \times p(\boldsymbol{\theta}). \tag{D.4}$$

If the computer model is non-linear, which is often the case for building simulation software, it is not possible to analytically estimate the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$, as often the numerator of Equation. D.2 may be hard or impossible to compute (Higdon et al., 2004; McClarren, 2018). An approach commonly used to overcome this problem is based on Markov Chain Monte Carlo (MCMC) (Higdon et al., 2004). MCMC is described in more detail in Section D.3.

D.2.2 Computationally expensive computer model

Often, the computational demands of the simulator prevent it from being used directly in the Bayesian calibration process. In such cases, a surrogate model may be used. The computer model is only run for a fixed number of simulation runs (m), upon which the surrogate model is trained on. For pairs of $(\mathbf{w},\mathbf{t})$ that the computer model was not evaluated at, the $\eta(\mathbf{w},\mathbf{t})$ output is treated as unknown. A commonly used surrogate model for Bayesian calibration is the Gaussian Processes (GP) (Kennedy and O'Hagan, 2001; Higdon et al., 2004; Bayarri et al., 2007). A GP can be fully defined by its mean, $\mu(\cdot)$, and covariance function, $(\text{cov}(\cdot,\cdot))$. It has often been used as a surrogate model since it can capture strong non-linearities and multivariable interactions (Heo and Zavala, 2012). It is common practice to define the mean function as to return zero, in which case the covariance function completely defines the GP model (Chong and Menberg, 2018; McClarren, 2018). Multiple options for covariance functions exist (Rasmussen and Williams, 2006), with the most commonly used option in calibration applications being the squared exponential (Kennedy and O'Hagan, 2001; Higdon et al., 2004):

$$k_{\eta,ij} = \frac{1}{\lambda_\eta}\exp\left\{-\sum_{h=1}^{p}\beta_h^{(\eta)}|w_{ih}-w_{jh}|^\alpha - \sum_{h'=1}^{l}\beta_{p+h'}^{(\eta)}|t_{ih'}-t_{jh'}|^\alpha\right\}, \tag{D.5}$$

where λ_η is the variance hyperparameter that controls the reciprocal of the marginal variance of $\eta(\cdot,\cdot)$, $\beta^{(\eta)}$ are the correlation hyperparameters and α is the smoothness hyperparameter of $\eta(\cdot,\cdot)$. p and l refer to the number of explanatory and calibration parameters, respectively. Equation D.5 may often be supplemented with a white noise component to ensure numerical stability and to account for small numerical fluctuations in the simulation (Higdon et al., 2004).

To define the likelihood function required for the calibration process, a joint $n+m$ vector $\mathbf{z}=(\mathbf{y}^T,\boldsymbol{\eta}^T)$ is defined with the first n components relating to the observations $\mathbf{y}=(y(w_1),\ldots,y(w_n))^T$ and the final m components relating to the simulation outputs $\boldsymbol{\eta}=(\eta(w_1^*,t_1^*),...,\eta(w_m^*,t_m^*))^T$. The associated training points for the surrogate model are $(w_1,\theta),...,(w_n,\theta)$ for the first n components and $(w_1^*,t_1^*),...,(w_m^*,t_m^*)$

for the final m components. A likelihood function can now be defined as follows:

$$\mathcal{L}(\mathbf{z}|\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\beta}^{(\eta)},\lambda_\eta,\lambda_\varepsilon) \propto |\mathbf{K}_z|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\mathbf{K}_z^{-1}(\mathbf{z}-\boldsymbol{\mu})\right\} \tag{D.6}$$

where:

$$\mathbf{K}_z = \mathbf{K}_\eta + \begin{pmatrix} \mathbf{K}_y & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{K}_y = I_n/\lambda_\varepsilon. \tag{D.7}$$

$\mathbf{K}_\eta$ is the result of applying D.5 to the $n+m$ group of training points and $\mathbf{K}_y$ is an $n \times n$ covariance matrix accounting for observation errors. In the next step, prior distributions should be defined for $\boldsymbol{\theta}$,$\boldsymbol{\mu}$, λ_η and $\beta^{(\eta)}$. By assuming independence and based on Bayes theorem the posterior distribution is (Higdon et al., 2004):

$$p(\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\beta}^{(\eta)},\lambda_\eta,\lambda_\varepsilon|\mathbf{z}) \propto \mathcal{L}(\mathbf{z}|\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\beta}^{(\eta)},\lambda_\eta,\lambda_\varepsilon)p(\boldsymbol{\theta})p(\boldsymbol{\mu})p(\boldsymbol{\beta}^{(\eta)})p(\lambda_\eta)p(\lambda_\varepsilon) \tag{D.8}$$

To explore the posterior distribution, MCMC may once again be used. In the approach described, the surrogate model is trained at the same time as the model calibration is performed. An alternative would be to train the surrogate model first only on the m simulation points and then perform the calibration – an approach referred to as modularisation (Bayarri et al., 2007; Liu et al., 2009).

The discrepancy in eq. 2.3 may be modelled with another GP model for $\delta(\mathbf{w})$, with a mean function of 0 and covariance function specified as (Higdon et al., 2004):

$$k_{\delta,ij} = \frac{1}{\lambda_\delta}\exp\{-\sum_{h=1}^{p}\beta_h^{(\delta)}|w_{ih}-w_{jh}|^{\alpha_\delta}\} \tag{D.9}$$

The likelihood function defined by eq D.6 remains unchanged with the exception of the definition of $\mathbf{K}_z$ which is now defined as:

$$\mathbf{K}_z = \mathbf{K}_\eta + \begin{pmatrix} \mathbf{K}_y + \mathbf{K}_\delta & 0 \\ 0 & 0 \end{pmatrix} \tag{D.10}$$

where $\mathbf{K}_\delta$ is an $n \times n$ matrix obtained by applying D.9 to each pair of the n input settings w_i that correspond to the monitored data $\mathbf{y}$.

D.3 MCMC Algorithm and Convergence.

MCMC involves the sequential sampling of $\boldsymbol{\theta}$ from approximate distributions and then correcting these draws to obtain a better approximation of the target (posterior) distribution. This process requires the estimation of the ratios of posterior density at different values of $\boldsymbol{\theta}$, and since the intractable term $p(y)$ is a constant, it cancels out (Higdon et al., 2004; McClarren, 2018). The draws form a *Markov Chain*, defined as a sequence of random variables $(\theta_1, \theta_2, \dots)$ for which at index i, the distribution of θ^i depends only on the value of θ_{i-1}. The effectiveness of the MCMC method depends on the fact that the approximate distribution is improved at every iteration by converging to the target distribution, and Gelman (2014) provides a simple proof as to why that is the case. MCMC methods can handle a large dimensional $\boldsymbol{\theta}$, deal with many nuisance parameters and multivariate output (Higdon et al., 2004).

MCMC convergence is assessed separately for each parameter by evaluating the level *mixing* and *stationarity* for each markov chain, but all must have converged before the distribution posteriors can be used for inference (Gelman, 2014). A chain may be considered well-mixed if it explores the full-probability region and is stationary if it remains within the same probability region. At least two chains initiated at different starting points are required to assess the mixing and stationarity of the distribution. Once a stationary distribution has been reached, the early iterations often discarded as they might be heavily influenced by the starting points. While the convergence can be assessed qualitatively by plotting the generated chains, a useful measure of convergence is the *potential scale reduction*, defined as (Gelman, 2014):

$$\hat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}(\psi|y)}{W}} \tag{D.11}$$

where $\widehat{\text{var}}^{+}(\psi|y)$ is the marginal posterior variance of the estimated quantity and W is the within-sequence variance. If $\hat{R}$ is high, then continuing the simulation is expected to improve the parameter inference. Generally, a value of $\hat{R} = 1 \pm 0.1$ indicates convergence.

Appendix E

Archetype-based Bayesian calibration

Section E.1 outlines the statistical formulation employed for the Bayesian calibration of archetype-based building stock models of summer indoor temperature. In section E.2, the prior probability distributions assumed for the hyperparameters are defined. Note that in the equations that will follow, superscripts in brackets are used as qualifiers. They are used, for example, to differentiate variables associated with monitored (M) and simulated (S) homes, where this was deemed necessary. They are, in some cases, omitted to ease the mathematical notation. Subscripts may be used as qualifiers or to index variables, and their purpose will be made clear in the text.

E.1 Statistical Formulation

The calibration approach employed within this work relies on the statistical formulation introduced by Kennedy and O'Hagan (2001). Given the large computational cost of UK-HSM (EnergyPlus) simulations, a Gaussian Process (GP) is used as a surrogate model ($\eta(\cdot)$) that is trained on EnergyPlus simulations ($\mathbf{y}_c^{(S)}$) and monitored data ($\mathbf{y}_c^{(M)}$), as suggested by Higdon et al. (2004). A GP is also used to represent the discrepancy term ($\delta(\cdot)$), as has often been the case in previous calibration work (Menberg et al., 2019). To clarify the use of explanatory and weather variables, Equation 2.3 is re-written as follows:

$$y_{md}^{(M)} = y(\mathbf{x}_m^{(M)}, \mathbf{w}_d) = \eta(\mathbf{x}_m^{(M)}, \mathbf{w}_d, \boldsymbol{\theta}) + \delta(\mathbf{x}_m^{(M)}, \mathbf{w}_d) + \varepsilon_{md}^{(M)}, \tag{E.1}$$

where $\mathbf{w}_d$ are the weather-related variables for day d, $\mathbf{x}_m^{(M)}$ are all other explanatory variables for house m and $\varepsilon_{md}^{(M)}$ is an error term associated with house m and day d. The set of calibration variables whose values are unknown and are the same between homes are denoted by $\boldsymbol{\theta}$. Both $\mathbf{x}_m^{(M)}$ and $\mathbf{w}_d$ are assumed to be known accurately (i.e. measurement error is negligible). For the computer simulations, the following relationship is established:

$$y_{sd}^{(S)} = y(\mathbf{x}_s^{(S)}, \mathbf{w}_d, \mathbf{t}_s) = \eta(\mathbf{x}_s^{(S)}, \mathbf{w}_d, \mathbf{t}_s) + \varepsilon_{sd}^{(S)}\,. \tag{E.2}$$

As per Higdon et al. (2004), a single combined vector of monitored and simulation data was constructed $\mathbf{z} = [\mathbf{y}_c^{(M)}, \mathbf{y}_c^{(S)}]$:

$$\mathbf{z} = \begin{bmatrix} y_{m=1,d=1}^{(M)} \\ \vdots \\ y_{m=1,d=D_c}^{(M)} \\ y_{m=2,d=1}^{(M)} \\ \vdots \\ y_{m=2,d=D_c}^{(M)} \\ y_{m=M,d=1}^{(M)} \\ \vdots \\ y_{m=M,d=D_c}^{(M)} \\ y_{s=1,d=1}^{(S)} \\ \vdots \\ y_{s=1,d=D_c}^{(S)} \\ y_{s=2,d=1}^{(S)} \\ \vdots \\ y_{s=2,d=D_c}^{(S)} \\ y_{s=S,d=1}^{(S)} \\ \vdots \\ y_{s=S,d=D_c}^{(S)} \end{bmatrix} = \begin{bmatrix} \eta(\mathbf{x}_1^{(M)}, \mathbf{w}_1, \boldsymbol{\theta}) + \delta(\mathbf{x}_1^{(M)}, \mathbf{w}_1) + \varepsilon_{11}^{(M)} \\ \vdots \\ \eta(\mathbf{x}_1^{(M)}, \mathbf{w}_{D_c}, \boldsymbol{\theta}) + \delta(\mathbf{x}_1^{(M)}, \mathbf{w}_{D_c}) + \varepsilon_{1D_c}^{(M)} \\ \eta(\mathbf{x}_2^{(M)}, \mathbf{w}_1, \boldsymbol{\theta}) + \delta(\mathbf{x}_2^{(M)}, \mathbf{w}_1) + \varepsilon_{21}^{(M)} \\ \vdots \\ \eta(\mathbf{x}_2^{(M)}, \mathbf{w}_{D_c}, \boldsymbol{\theta}) + \delta(\mathbf{x}_2^{(M)}, \mathbf{w}_{D_c}) + \varepsilon_{2D_c}^{(M)} \\ \eta(\mathbf{x}_M^{(M)}, \mathbf{w}_1, \boldsymbol{\theta}) + \delta(\mathbf{x}_M^{(M)}, \mathbf{w}_1) + \varepsilon_{M1}^{(M)} \\ \vdots \\ \eta(\mathbf{x}_M^{(M)}, \mathbf{w}_{D_c}, \boldsymbol{\theta}) + \delta(\mathbf{x}_M^{(M)}, \mathbf{w}_{D_c}) + \varepsilon_{MD_c}^{(M)} \\ \eta(\mathbf{x}_1^{(S)}, \mathbf{w}_1, \mathbf{t}_1) + \varepsilon_{11}^{(S)} \\ \vdots \\ \eta(\mathbf{x}_1^{(S)}, \mathbf{w}_{D_c}, \mathbf{t}_1) + \varepsilon_{1D_c}^{(S)} \\ \eta(\mathbf{x}_2^{(S)}, \mathbf{w}_1, \mathbf{t}_2) + \varepsilon_{21}^{(S)} \\ \vdots \\ \eta(\mathbf{x}_2^{(S)}, \mathbf{w}_{D_c}, \mathbf{t}_2) + \varepsilon_{2D_c}^{(S)} \\ \eta(\mathbf{x}_S^{(S)}, \mathbf{w}_1, \mathbf{t}_S) + \varepsilon_{S1}^{(S)} \\ \vdots \\ \eta(\mathbf{x}_S^{(S)}, \mathbf{w}_{D_c}, \mathbf{t}_S) + \varepsilon_{SD_c}^{(S)} \end{bmatrix}. \tag{E.3}$$

The surrogate model was defined as a zero-mean GP, with the squared exponential kernel used for its covariance function:

$$k_{\eta,ij} = \frac{1}{\lambda_\eta} \exp\left\{ -\sum_{h=1}^{p} \beta_h^{(\eta)} |x_{ih} - x_{jh}|^2 - \sum_{h'=1}^{q} \beta_{p+h'}^{(\eta)} |w_{ih'} - w_{jh'}|^2 - \sum_{h''=1}^{r} \beta_{p+q+h''}^{(\eta)} |t_{ih''} - t_{jh''}|^2 \right\}, \tag{E.4}$$

The discrepancy term was also specified as a zero-mean GP with a squared exponential kernel:

$$k_{\delta,ij} = \frac{1}{\lambda_\delta} \exp\left\{ -\sum_{h=1}^{p} \beta_h^{(\delta)} |x_{ih} - x'_{jh}|^2 - \sum_{h'=1}^{q} \beta_{p+h'}^{(\delta)} |w_{ih'} - w_{jh'}|^2 \right\}, \tag{E.5}$$

The use of a zero-mean discrepancy term translates to a prior belief that there is no systematic model bias.

Following from the assumption of independently and identically distributed errors, the resulting likelihood function is defined as:

$$\mathcal{L}(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\xi}) \propto |\mathbf{K}_z|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^T \mathbf{K}_z^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right\} \tag{E.6}$$

where $\boldsymbol{\mu}$ is a vector of zeros, $\boldsymbol{\xi} = [\boldsymbol{\beta}^{(\eta)}, \boldsymbol{\beta}^{(\delta)}, \lambda_\eta, \lambda_\delta, \lambda_\varepsilon, \lambda_{sim}]$ and:

$$\mathbf{K}_z = \mathbf{K}_\eta + \begin{pmatrix} \mathbf{K}_y + \mathbf{K}_\delta & 0 \\ 0 & \mathbf{K}_{sim} \end{pmatrix}, \quad \mathbf{K}_y = \mathbf{I}^{(M)} / \lambda_\varepsilon, \quad \mathbf{K}_{sim} = \mathbf{I}^{(S)} / \lambda_{sim}. \tag{E.7}$$

$\mathbf{K}_\eta$ is the result of applying E.4 to the N_c group of training points, $\mathbf{K}_\delta$ results from applying E.5 to the $N_c^{(M)}$ monitored data, $\mathbf{K}_y$ is an $N_c^{(M)} \times N_c^{(M)}$ covariance matrix of accounting for observation errors and $\mathbf{K}_{sim}$ is an $N_c^{(S)} \times N_c^{(S)}$ covariance matrix of accounting for numerical errors. $\mathbf{I}^{(M)}$ is an $N_c^{(M)} \times N_c^{(M)}$ identity matrix, $\mathbf{I}^{(S)}$ is an $N_c^{(S)} \times N_c^{(S)}$ identity matrix.

E.2 Hyperparameter priors

As with any Bayesian analysis, it is crucial to identify and list all priors used. The hyperparameter priors were defined based on the suggestions of Chong and Menberg (2018) and Menberg et al. (2019). In summary, the priors used were:

- $\rho_1^{(\eta)}, \ldots, \rho_{p+q}^{(\eta)} \sim Beta(shape1 = 1, shape2 = 0.3)$: Reparametrisations of the correlation hyperparameters $\beta_1^{(\eta)}, \ldots, \beta_{p+q}^{(\eta)}$ (see Equation D.5) used to define the emulator's GP, where $\rho_i^{(\eta)} = \exp(\beta_i^{(\eta)}/4), i, \ldots, (p+q)$.

- $\rho_1^{(\delta)}, \ldots, \rho_{p+q}^{(\delta)} \sim Beta(shape1 = 1, shape2 = 0.3)$ Reparametrisations of the correlation hyperparameters used to define the discrepancy term's GP.

- $\lambda_\eta \sim Gamma(shape = 10, rate = 10)$: Following the standardisation of the model output to have unit variance, this precision parameter is expected to have a value close to one – this prior specification result in an expectation value of 1.

- $\lambda_\delta \sim Gamma(shape = 10, rate = 0.3)$: This prior captures the assumption that a small model bias is expected, and is also used to avoid confounding between the calibrating the model parameters and tuning the discrepancy term.

Appendix F

Kronecker Product

F.1 Kronecker Product Definition

Assume two 2×2 matrices $\mathbf{A}$ and $\mathbf{B}$:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \tag{F.1}$$

The Kronecker product of these matrices is (Pollock, 2013):

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} & a_{12} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ a_{21} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} & a_{22} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \end{bmatrix} \tag{F.2}$$

$$= \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{bmatrix} \tag{F.3}$$

More generally, the Kronecker product of an $m \times n$ matrix $\mathbf{A}$ and a $p \times q$ matrix $\mathbf{B}$ is

defined as:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \cdots & \cdots & \ddots & \cdots \\ a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \end{bmatrix} \tag{F.4}$$

$$= \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & \cdots & a_{11}b_{1q} & \cdots & \cdots & a_{1n}b_{11} & a_{1n}b_{12} & \cdots & a_{1n}b_{1q} \\ a_{11}b_{21} & a_{11}b_{22} & \cdots & a_{11}b_{2q} & \cdots & \cdots & a_{1n}b_{21} & a_{1n}b_{22} & \cdots & a_{1n}b_{2q} \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots \\ a_{11}b_{p1} & a_{11}b_{p2} & \cdots & a_{11}b_{pq} & \cdots & \cdots & a_{1n}b_{p1} & a_{1n}b_{p2} & \cdots & a_{1n}b_{pq} \\ \vdots & \vdots & & \vdots & \ddots & & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & & \ddots & \vdots & \vdots & & \vdots \\ a_{m1}b_{11} & a_{m1}b_{12} & \cdots & a_{m1}b_{1q} & \cdots & \cdots & a_{mn}b_{11} & a_{mn}b_{12} & \cdots & a_{mn}b_{1q} \\ a_{m1}b_{21} & a_{m1}b_{22} & \cdots & a_{m1}b_{2q} & \cdots & \cdots & a_{mn}b_{21} & a_{mn}b_{22} & \cdots & a_{mn}b_{2q} \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{p1} & a_{m1}b_{p2} & \cdots & a_{m1}b_{pq} & \cdots & \cdots & a_{mn}b_{p1} & a_{mn}b_{p2} & \cdots & a_{mn}b_{pq} \end{bmatrix} \tag{F.5}$$

Assuming that $\mathbf{A}$ and $\mathbf{B}$ are invertible, the following is true (Pollock, 2013):

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = (\mathbf{A}^{-1} \otimes \mathbf{B}^{-1}) \tag{F.6}$$

F.2 Kronecker Product Implementation

Focusing only on the emulation component of the Bayesian calibration using the combined monitored and simulated data, if the vector of data points is of length $N_c = (M+S) \times D_c$, the covariance matrix $\mathbf{K}_\eta$ will be of size $N_c \times N_c$. The regular structure of the data discussed in Section 7.1.6 allows for the covariance matrix to be reformulated as follows:

$$\mathbf{K}_\eta = \mathbf{K}_\eta^{(M+S)} \otimes \mathbf{K}_\eta^{(D_c)}, \tag{F.7}$$

where $\mathbf{K}_{\eta}^{(M+S)}$ is a covariance matrix estimated using only the $(\mathbf{x}_m, \mathbf{x}_s, \mathbf{t}_s)$, $\mathbf{K}_{\eta}^{(D_c)}$ a covariance matrix estimated using only the $(\mathbf{w}_d)$, and $\otimes$ is the Kronecker product. Since it is true that $\mathbf{K}_{\eta}^{-1} = (\mathbf{K}_{\eta}^{(M+S)})^{-1} \otimes (\mathbf{K}_{\eta}^{(D_c)})^{-1}$ (Equation F.6), the reduction in computational cost is achieved by the fact that inverting $(M+S)$ and (D_c) - dimensional matrices and estimating their Kronecker product is faster than inverting a single $(M+S) \times D_c$ - dimensional matrix; the computational cost is reduced from $\mathcal{O}(N_c^3)$ to approximately $\mathcal{O}((M+S)^3 + D_c^3)$ (Hung et al., 2015).

Implementing the above formulation for emulation purposes can be relatively straightforward. However, a greater challenge exists when the entire calibration framework is considered and in particular when trying to account for model discrepancy. One potential approach of implementing the Kronecker product for Bayesian calibration was offered by Bayarri et al. (2009), who investigated vehicle crashworthiness and where the data structure was the result of the model's time-dependent functional output. It requires the key assumption that the GP correlation parameters of the weather variables are the same for the emulator, discrepancy and measurement error terms (i.e. $\boldsymbol{\beta}_w^{(\eta)} = \boldsymbol{\beta}_w^{(\delta)} = \boldsymbol{\beta}_w^{(\varepsilon)}$). This does not assume that the variations of the three functions with regard to the weather are the same, only that they have the same correlation structure (Bayarri et al., 2009). If this is a reasonable assumption to make, a Kronecker product implementation for the entire calibration problem may be implemented by defining the covariance matrix for the non-weather variables as follows (adapted for this application from Bayarri et al. (2009)):

$$\mathbf{K}^{(M,S)} = \begin{pmatrix} \mathbf{K}_{11}^{(M,S)} & \mathbf{K}_{12}^{(M,S)} \\ \mathbf{K}_{21}^{(M,S)} & \mathbf{K}_{22}^{(M,S)} \end{pmatrix}, \tag{F.8}$$

$$\mathbf{K}_{11}^{(M,S)} = \frac{1}{\lambda_\eta} C^{(\eta)}(\mathbf{D}^{(M)}, \mathbf{D}^{(M)}) + \frac{1}{\lambda_\delta} C^{(\delta)}(\mathbf{D}^{(M)}, \mathbf{D}^{(M)}) + \frac{1}{\lambda_\varepsilon} \mathbf{I}_{M \times M}, \tag{F.9}$$

$$\mathbf{K}_{12}^{(M,S)} = \frac{1}{\lambda_\eta} C^{(\eta)}(\mathbf{D}^{(M)}, \mathbf{D}^{(S)}), \tag{F.10}$$

$$\mathbf{K}_{21}^{(M,S)} = \frac{1}{\lambda_\eta} C^{(\eta)}(\mathbf{D}^{(S)}, \mathbf{D}^{(M)}), \tag{F.11}$$

$$\mathbf{K}_{22}^{(M,S)} = \frac{1}{\lambda_\eta} C^{(\eta)}(\mathbf{D}^{(S)}, \mathbf{D}^{(S)}) + \frac{1}{\lambda_{sim}} \mathbf{I}_{S \times S}, \tag{F.12}$$

where $\mathbf{D}^{(M)}$ is a data matrix of length M associated with the monitored homes and $\mathbf{D}^{(S)}$ is a data of matrix of length S associated with the simulated homes. $C^{(\eta)}(\mathbf{D}^{(M)},\mathbf{D}^{(M)})$, signifies the application of the squared exponential covariance function on data $\mathbf{D}^{(M)}$, with an equivalent meaning for other $C(\cdot,\cdot)$ terms. $\mathbf{I}_{M\times M}$ and $\mathbf{I}_{S\times S}$ are identity matrices of size $M\times M$ and $S\times S$, respectively. The covariance matrix for the weather variables is defined as:

$$\mathbf{K}^{(W)} = C^{(W)}(\mathbf{W}_c,\mathbf{W}_c) \tag{F.13}$$

where $\mathbf{W}_c$ is the weather data matrix used for the calibration of length D_c. The overall covariance matrix for the calibration problem becomes:

$$\mathbf{K}^{(z,kron)} = \mathbf{K}^{(M,S)} \otimes \mathbf{K}^{(W)}. \tag{F.14}$$

Appendix G

The Morris Method

To define what an elementary effect is, a model (Y) is assumed with k independent inputs $\boldsymbol{x} = x_1, x_2, ...x_k$ that may take a value between 0 and 1 depending on the chosen number of levels (p) (Saltelli et al., 2008). For a given starting point, the elementary effect for the ith input factor is defined as (Morris, 1991; Saltelli et al., 2008):

$$EE_i = \frac{Y(x_1, x_2, ..., x_i + \Delta, ..., x_k) - Y(x_1, x_2, ..., x_k)}{\Delta}, \tag{G.1}$$

where Δ is a predefined jump on the p-level grid Ω. For each trajectory, one elementary effect can be computed for each parameter i resulting in $k+1$ simulations. Repeating this process for r trajectories results in a distribution of elementary effects for the computational cost of $r(k+1)$ simulations. A visualisation of the Morris method for two parameters is shown in Figure G.1. From the distribution of elementary effects, the mean (μ) and the standard deviation (σ) are often used to describe the influence of the input parameter's uncertainty on the model output and are defined as (Saltelli et al., 2008):

$$\mu_i = \frac{1}{r} \sum_{t=1}^{r} EE_{it}, \tag{G.2}$$

$$\sigma_i = \sqrt{\frac{1}{r-1} \sum_{t=1}^{r} (EE_{it} - \mu_i)^2}, \tag{G.3}$$

where t is the index of each trajectory. The mean assesses the overall influence of

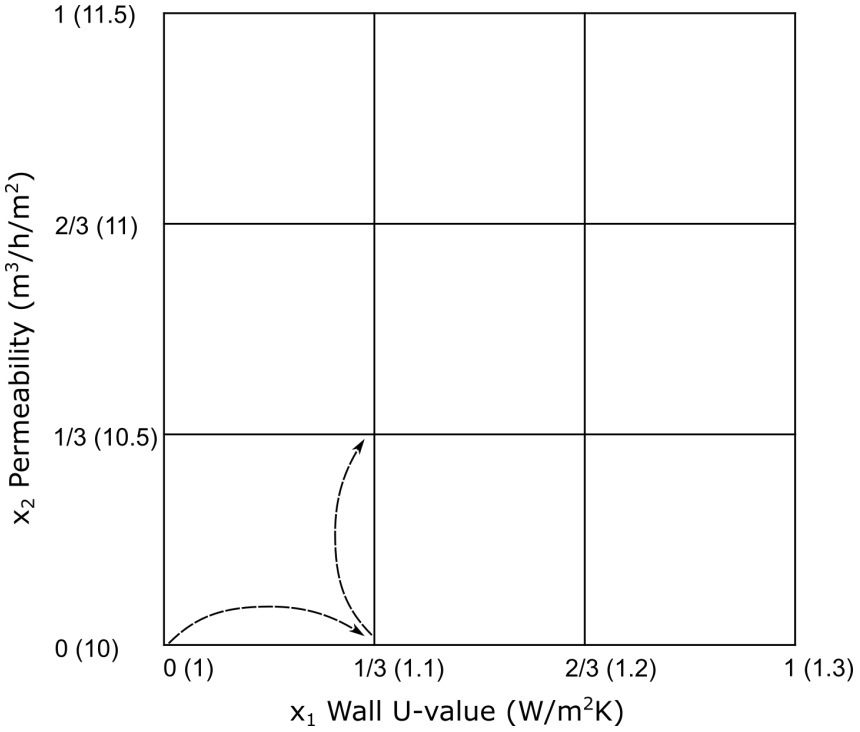


Figure G.1: A visualisation of the Elementary Effects method for two variables, Wall U-value (x_1) and Permeability (x_2). The true values of each variable's parameter range, shown in the brackets, have been scaled to the range of 0 to 1. In the first trajectory, from a starting point of $(x_1, x_2) = (0, 0)$, the value of x_1 changes by $\Delta = 1/3$. At the next step, x_1 remains constant while x_2 changes by $1/3$. For any number of new trajectories, a new starting point would be selected and each parameter would change by Δ, one parameter at a time.

the factor on the output while the standard deviation is a measure of the spread of elementary effects and indicates the level of dependence and interaction of the *i*th factor on other factors, along with any non-linearities.

As an improvement, Campolongo et al. (2007) suggested the use of the mean of absolute elementary effects (μ^*), defined below (Saltelli et al., 2008):

$$\mu_i^* = \frac{1}{r} \sum_{t=1}^{r} |EE_{it}| \tag{G.4}$$

It is considered to be a good proxy of the overall effect of a factor on the model output and should be preferred to the initially proposed arithmetic mean (μ) which is vulnerable to type II errors (i.e. failure to identify factors with a significant influence on the output) (Saltelli et al., 2008). Based on the tenets of normality, Garcia Sanchez et al. (2014) used the σ/μ^* ratio to identify non-linear and higher-order effect based

on the following categories: (i) Almost linear effects are identified in the region of $\sigma/\mu^* < 0.1$, (ii) monotonic effects at $0.1 < \sigma/\mu^* < 0.5$, (iii) almost monotonic at $0.5 < \sigma/\mu^* < 1$ and (iv) non-monotonic or with interactions at $\sigma/\mu^* > 1$. Further study of the parameter interactions was also proposed at a computational cost of kr^2.

Appendix H

Additional Results from Stochastic Characterisation

H.1 Wall U-value

The corrected arithmetic mean of the measured wall U-values per wall type are visualised in Figure H.1. The solid vertical lines represent the empirical median based on the data collected while the dashed lines are the theoretical values based on Appendix S (RdSAP) of SAP 2012 (BRE, 2014)

The histogram of filled cavity walls was discussed in Section 5.2.1. For non-standard solid walls, there is large spread from $0.4\,\mathrm{W/(m^2K)}$ to $2.3\,\mathrm{W/(m^2K)}$. The small sample size (33) makes it difficult to identify outliers by simply inspecting the histogram. The wide range of wall types that might fall under this construction might explain the wide variability (Hulme and Doran, 2014). For standard solid walls, 95 % of measured values are between $1.1\,\mathrm{W/(m^2K)}$ and $2.2\,\mathrm{W/(m^2K)}$. Values concentrated around $0.6\,\mathrm{W/(m^2K)}$ were partly attributed to an older wall construction (pre 1850s) which however appeared to be standard solid wall upon inspection. Although these older wall constructions are uncommon, they were kept in the data. Based on the density plot, the distribution of unfilled cavity walls is likely the result of two clusters. This is supported by Hulme and Doran (2014), with the two clusters assumed to be a brick only cavity construction and a brick-block construction.

Based on the analysis of the Cullen and Frey graph presented in Section 5.2.1,

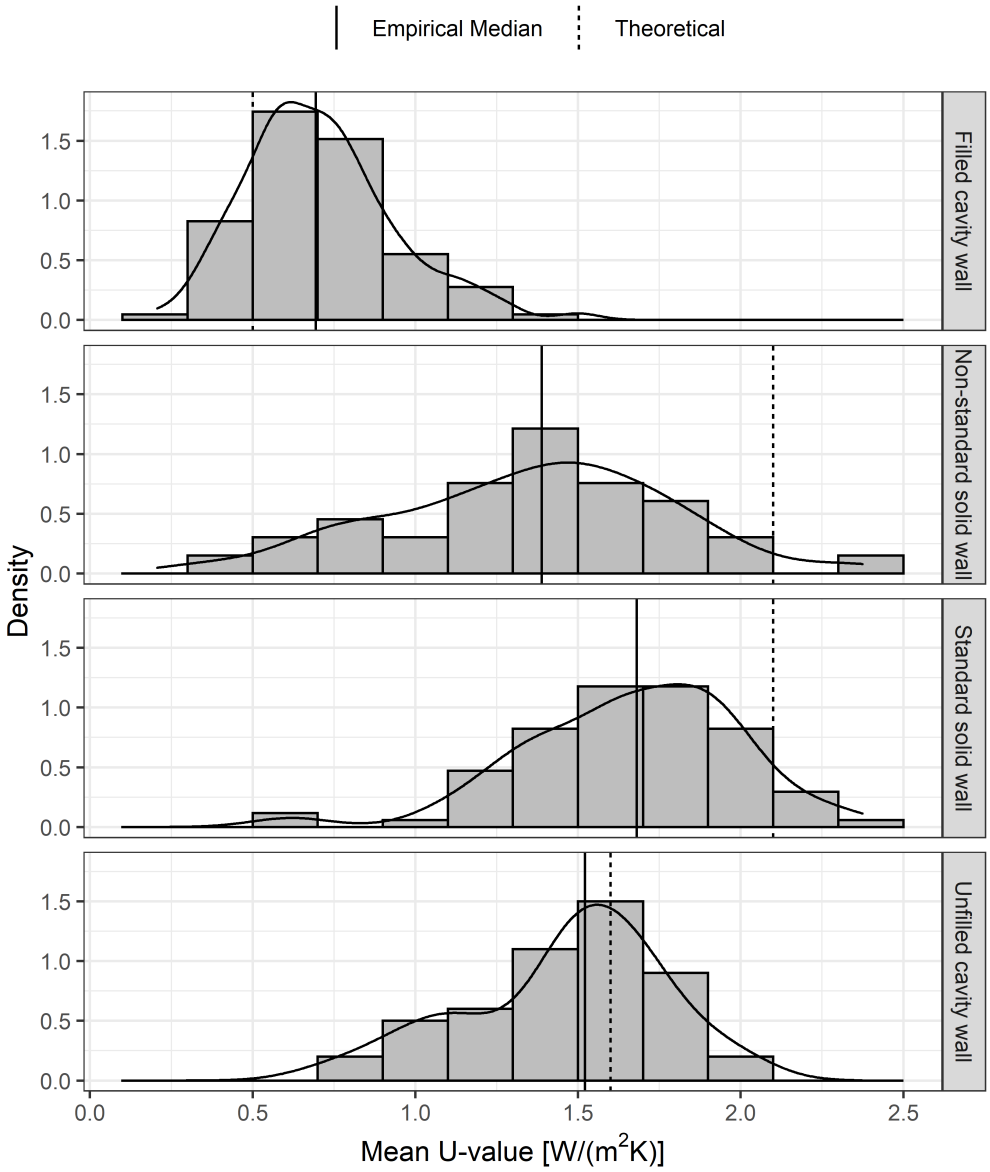


Figure H.1: Histograms and density lines of the measured wall U-value, following a 6 % correction. Data from Hulme and Doran (2014). The theoretical line is based on RdSAP of SAP 2012 (BRE, 2014).

four distributions were fitted to the filled cavity wall U-values: normal, Weibull, lognormal and gamma. Due to the small computational burden of fitting multiple distributions, all four candidate distributions were also fitted for the unfilled cavity wall and standard solid wall construction. Due to the small number of non-standard solid walls, a distribution for this wall type was not fitted. The results for all three wall types are presented in Table H.1. As discussed in Section 5.2.1, the gamma

Table H.1: Distributions for each wall type ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j). Corrected AIC was used for unfilled cavity wall. P1 and P2 represent the parameters of the fitted distribution, stated to two significant figures.

Wall Type	Distr.	AIC	Δ_j	w_j	P1	P2
Filled cavity	gamma	-16.07	0.00	0.75	shape = 9.5	rate = 13
	lnorm	-13.72	2.34	0.23	meanlog = -0.4	sdlog = 0.33
	norm	-8.04	8.03	0.01	mean = 0.71	sd = 0.23
	weibull	-7.59	8.47	0.01	shape = 3.2	scale = 0.79
Unfilled cavity	weibull	23.41	0.00	0.82	shape = 5.8	scale = 1.6
	norm	26.60	3.19	0.17	mean = 1.5	sd = 0.3
	gamma	31.64	8.24	0.01	shape = 20	rate = 14
	lnorm	35.25	11.84	0.00	meanlog = 0.35	sdlog = 0.23
Standard solid	weibull	51.61	0.00	0.87	shape = 6.0	scale = 1.8
	norm	55.44	3.83	0.13	mean = 1.7	sd = 0.33
	gamma	68.27	16.66	0.00	shape = 21	rate = 13
	lnorm	77.94	26.33	0.00	meanlog = 0.48	sdlog = 0.23

distribution best described the filled cavity wall U-values, which also provided a good fit based on the goodness-of-fit plot.

For the unfilled cavity wall, the Weibull distribution had the lowest AIC (23.41) and 0.82 probability of being the best fit amongst the candidate distributions. The normal provides the second-best fit (AIC = 26.60), followed by the gamma (AIC = 31.64) and lognormal (AIC = 35.25). The same order of fit is seen for the standard solid wall, with the 0.87 probability suggesting that the Weibull provides the best description of the data.

A good fit is observed for the Weibull distribution fitted to the standard solid wall U-value as seen in Figure H.2. All data points follow the P-P plot diagonal wall. The same is true for the Q-Q plot with the exception of two points from the left tail of the U-value distribution. These are the same data points identified earlier as lying away from the empirical distribution whose construction was also questioned by Hulme and Doran, 2014. The fitted distribution will be unable to represent these comparatively low U-values well, however the overall effect of this when modelling the stock will likely be small since they are infrequent.

The goodness of fit plots for the cavity wall construction when fitted with a

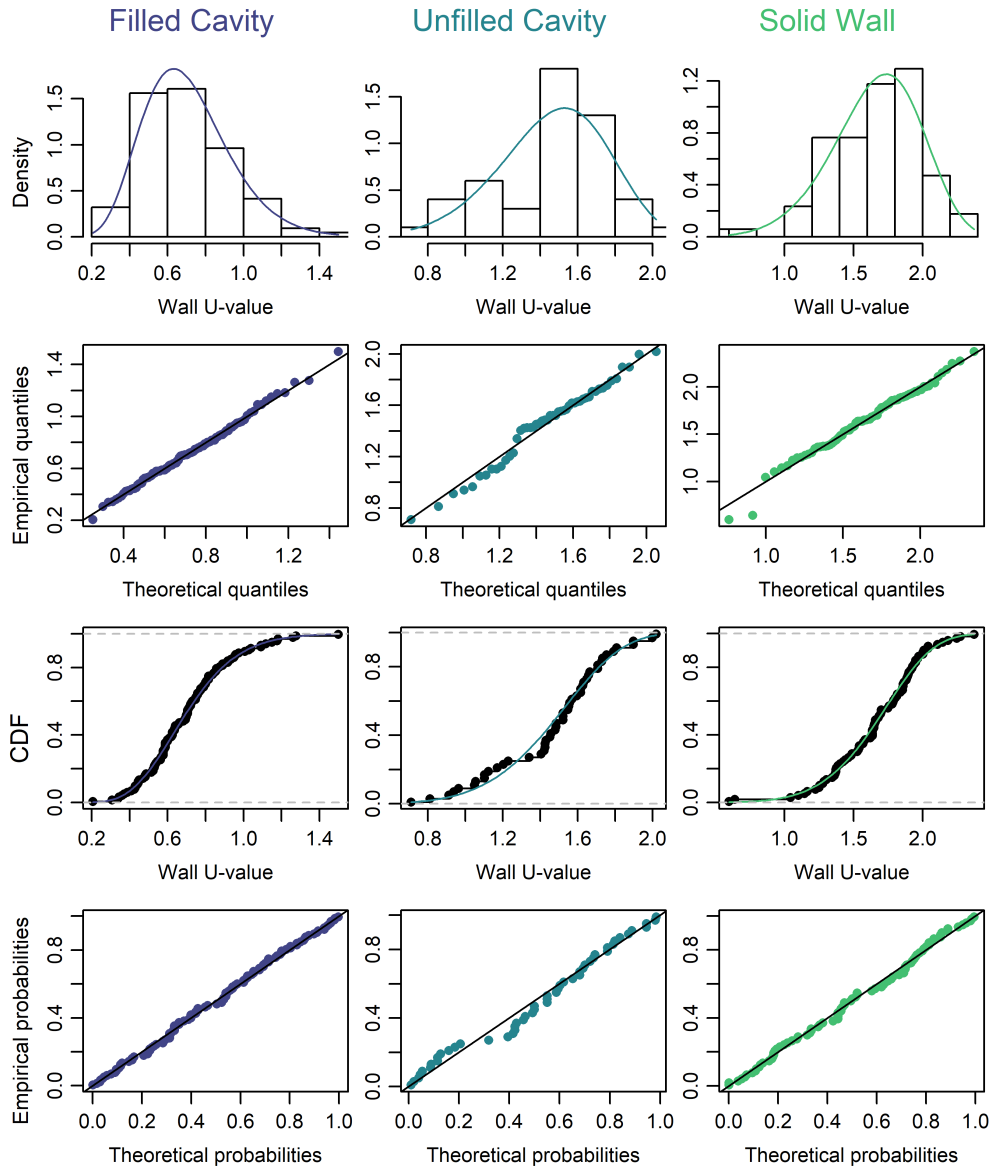


Figure H.2: Goodness of fit plots for the BRE dataset of wall U-values. A Gamma(9.5, 13) was assumed for the filled cavity, Weibull(5.8, 1.6) for the unfilled cavity and a Weibull(6.0, 1.8) for the solid wall.

Weibull distribution is also visualised in Figure H.2. Based on the Q-Q plot the extremes are represented fairly well, while the P-P plots suggests the probability of obtaining values around 1.3–1.4 $\mathrm{W}/(\mathrm{m}^2\mathrm{K})$ is greater than what the empirical evidence indicate. Since all the distributions fitted here are unimodal, the possible bi-modality due the presence of two different types of cavity wall construction is not captured. With the raw data of cavity walls not being separated into different

types, fitting a multimodal distribution (most commonly done as the combination of two unimodal distributions) would require artificially grouping data points into categories based on their values – an error-prone method since the values that each group should take (or what the groups are) is not accurately known.

H.2 Floor U-value

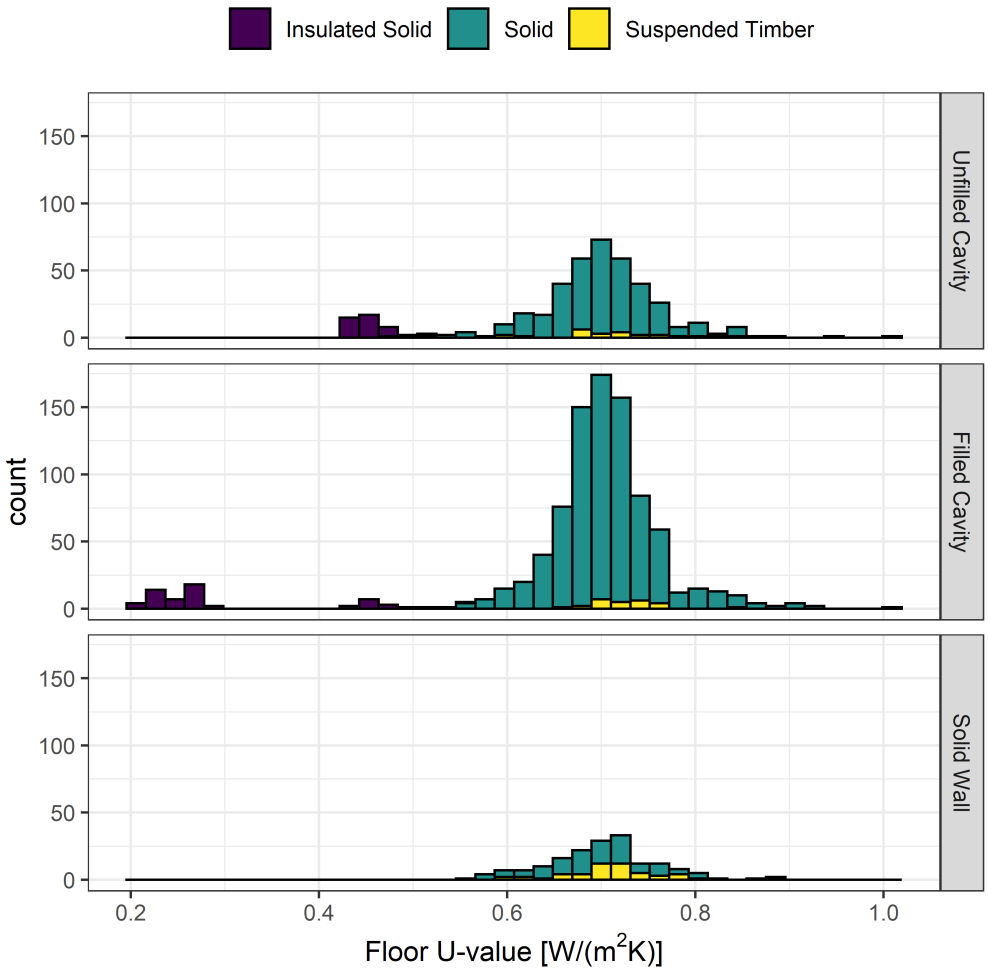


Figure H.3: Floor U-values estimated using the RdSAP S5.5 guidance for semi-detached dwellings in the 2012 English Housing Survey.

Solid floor is the most common floor type for all three wall types (Figure H.3). Suspended timber floor is uncommon for cavity wall dwellings, but fairly typical for solid wall dwellings. The distributions of U-values for uninsulated solid floors and suspended timber floors have similar central values with their median ranging

within 0.69–0.72 $W/(m^2K)$ depending on the wall type. These medians differ from those of insulated solid floor U-values with median values of 0.26 $W/(m^2K)$ and 0.45 $W/(m^2K)$. A single unimodal distribution might be able to describe the floor U-values of the solid wall group, but not the filled or unfilled cavity groups due to the clear multi-modality seen in Figure H.3. Multiple distributions would therefore be required to describe these datasets, or individual distributions that describe well the majority but not all data points.

H.3 Fabric Air Permeability

To explore the factors influencing airtightness, Stephen (2000) compared the mean leakage rate (ACH at 50 Pa) of different sub-groups of the dataset. Cavity masonry was more leaky ($\approx$ 15 ACH at 50 Pa, n = 205) than solid masonry ($\approx$ 11.5 ACH at 50 Pa, n = 108), while suspended timber floors were more leaky ($\approx$ 16 ACH at 50 Pa, n = 202) than solid concrete floors ($\approx$ 11.5 ACH at 50 Pa, n = 189). Other factors were studied through reductive sealing of air leakage paths but were found to have a smaller effect than wall or floor type (Stephen, 2000). For example, windows and doors were found to be responsible for 16 % of air leakage, contrary to 71 % attributed to the "myriad crack and openings" on the floor and walls (Stephen, 2000).

While informative, the analysis by Stephen (2000) could be expanded further. For example Stephen (1998) suggested that the choice of wall construction "does not guarantee a particular level of air tightness because there is still a range of airtightness for each wall type and the ranges greatly overlap". Yet, the comparison presented to readers omits the shape and range of air tightness and focuses only on the mean values. In addition, while comparing the different possible values of a factor is useful (e.g. cavity vs solid wall construction), a study of the interaction of different factors (e.g. cavity wall with suspended timber floors) would also be informative. This would be especially useful when looking at the association of construction age with other factors. Furthermore, Stephen (1998) discussed in their literature review that insulation may have an effect on airtightness, yet there was no description of the prevalence of insulation within the dataset analysed. Finally, it

is unclear how representative the dataset is of the English housing stock since the sampling of homes was not discussed.

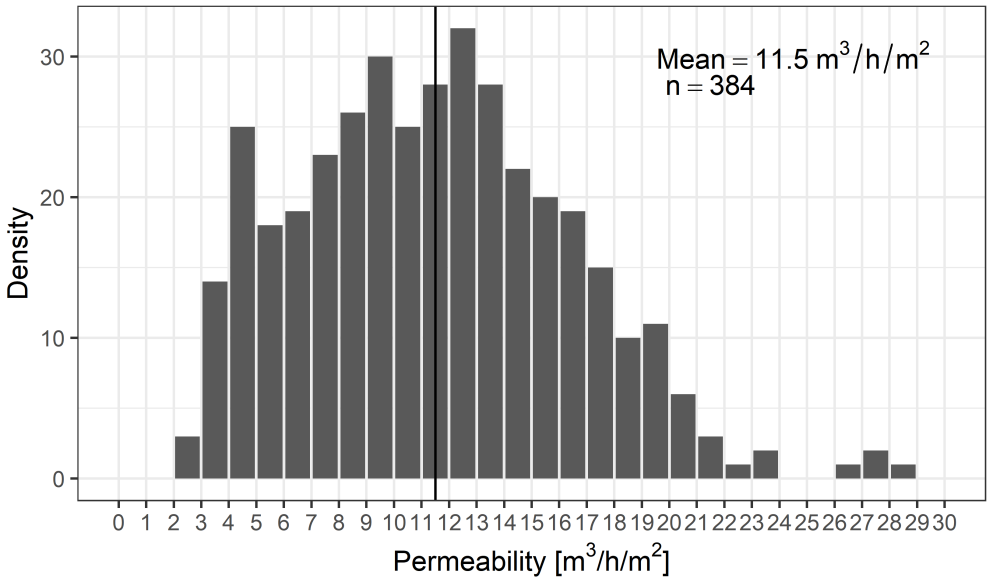


Figure H.4: Permeability measurements of pre-1995 from the BRE dataset, reproduced from Stephen (2000).

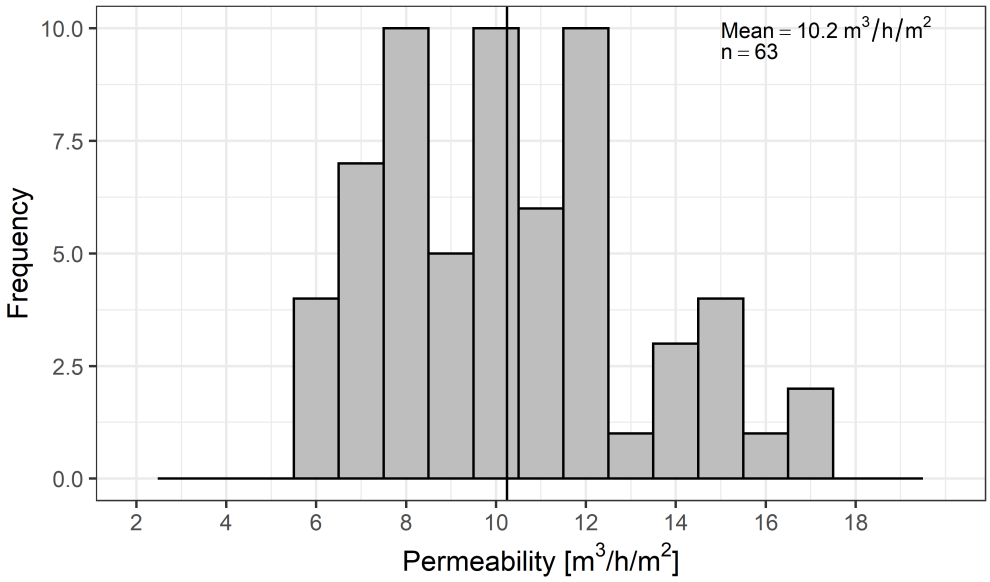


Figure H.5: Permeability measurements of 2002–2006 houses based on the dataset from BRE (2004).

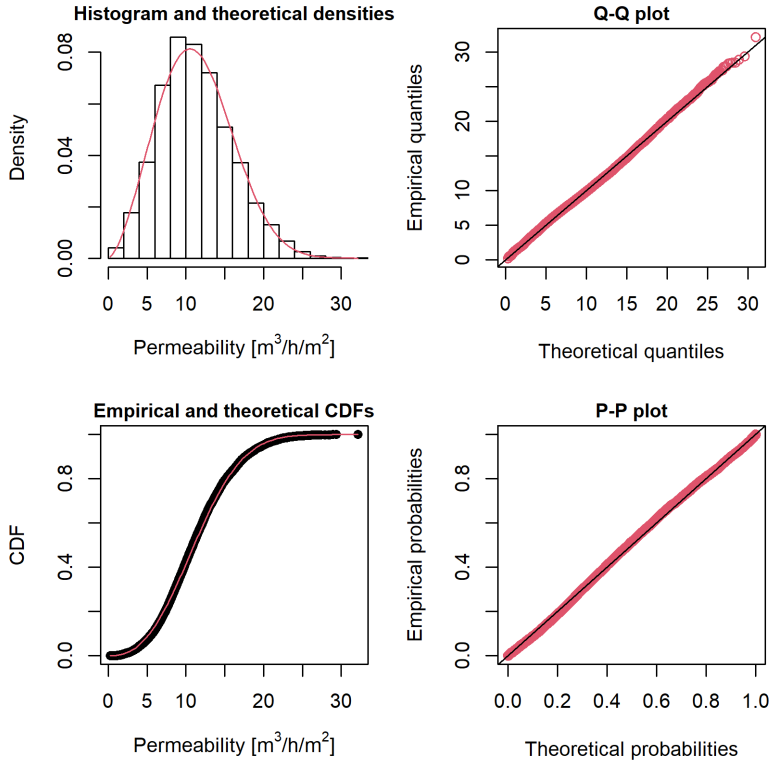


Figure H.6: Goodness of fit plots for the distribution of air permeability weighted by cluster dwelling age and assuming a Weibull distribution.

H.3.1 Mixture of Normal Distributions

The mean ($\mu^{(m)}$) of a mixture of normal distributions is defined as (Behboodian, 1970):

$$\mu^{(m)} = \sum_{j=1}^{k} p_j \mu_j, \tag{H.1}$$

where p_j is the weight for distribution j, and k is the total number of normal distributions. The variance ($\sigma^{2(m)}$) of a mixture of normal distributions is defined as (Behboodian, 1970):

$$\sigma^{2(m)} = \sum_{j=1}^{k} p_j \left(\sigma_j^2 + \mu_j^2\right) - \left(\sum_{j=1}^{k} p_j \mu_j\right)^2, \tag{H.2}$$

where σ_j^2 is the variance for distribution j.

H.4 Glazing Fraction

For all three wall types, the distribution that best describes the glazing fraction is a gamma although the shape and rate differ. The goodness of fit plots in Figure H.8 suggest that the gamma distribution provides a satisfactory fit.

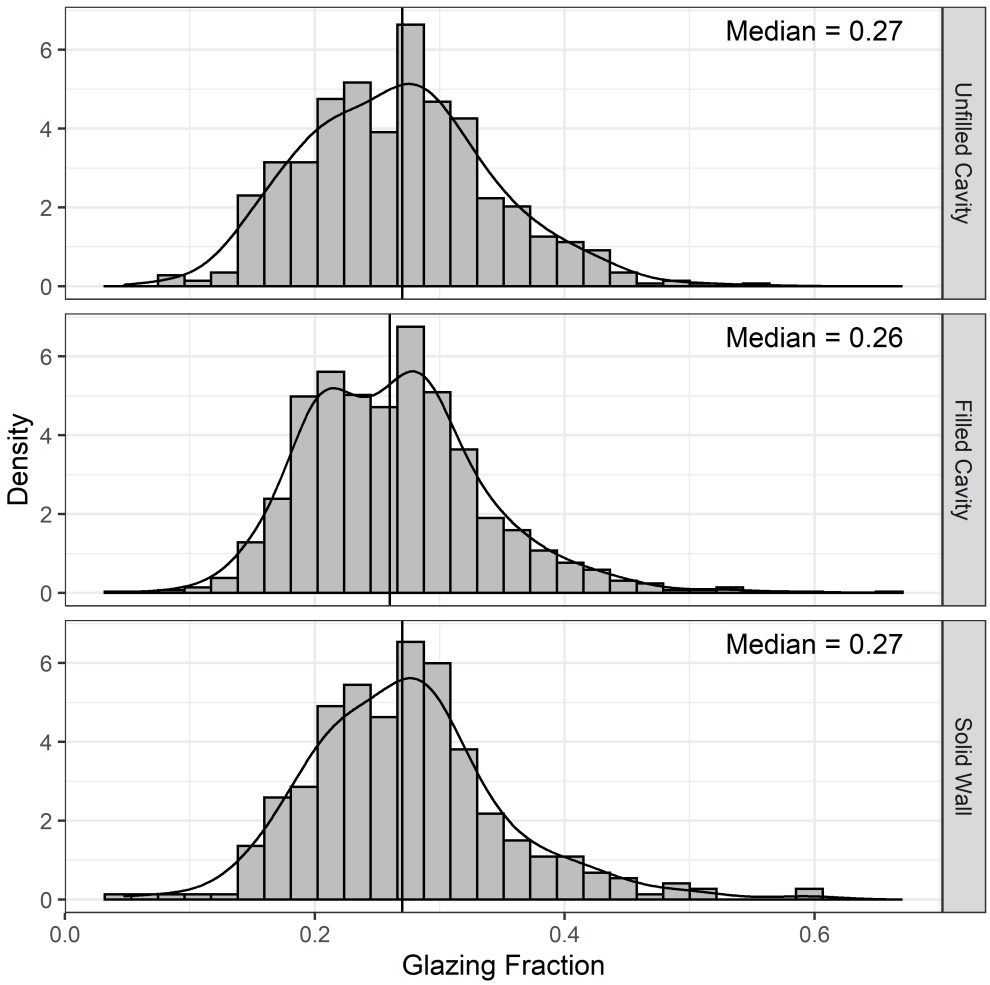


Figure H.7: Histograms of glazing fraction estimates of semi-detached dwellings in the English Housing Survey. The solid vertical line indicates the median.

Table H.2: Distributions fitted to the glazing fraction of semi-detached dwellings in the English Housing Survey, grouped by wall type. They are ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j). Wall types are: FCW = Filled Cavity Wall, UCW = Unfilled Cavity Wall and SW = Solid Wall.

Wall	Dist.	AIC	Δ_j	w_j	P1	P2
FCW	gamma	-3401	0	1.00	shape = 14	rate = 53
	lnorm	-3388	13	0.00	meanlog = -1.4	sdlog = 0.27
	norm	-3303	98	0.00	mean = 0.26	sd = 0.072
	weibull	-3220	181	0.00	shape = 3.7	scale = 0.29
UCW	gamma	-1604	0	1.00	shape = 13	rate = 47
	norm	-1591	12	0.00	mean = 0.27	sd = 0.074
	lnorm	-1581	22	0.00	meanlog = -1.4	sdlog = 0.29
	weibull	-1577	26	0.00	shape = 3.8	scale = 0.29
SW	gamma	-798	0	1.00	shape = 12	rate = 45
	norm	-783	15	0.00	mean = 0.27	sd = 0.077
	lnorm	-783	15	0.00	meanlog = -1.3	sdlog = 0.3
	weibull	-765	33	0.00	shape = 3.6	scale = 0.3

H.5 Floor-to-ceiling height

Figure H.9 provides histograms with density plots of the floor-to-ceiling height, grouped by wall construction and following the removal dwellings with values greater than 3.5 m. Both types of cavity wall construction are characterised by the same median value of 2.62 m, which differs by 0.1 m from the solid wall median of 2.52 m.

A lognormal describes the floor-to-ceiling height of dwellings with solid wall construction best, with a meanlog = 0.96 and sdlog = 0.054. Based on the Q-Q plot in Figure H.10, the theoretical and empirical quantiles are in good agreement for values up to 3.0 m, or 98 % of the dataset. The Fréchet provided the best description of the data originating from homes with unfilled cavity walls, with a shape = 27 and a scale = 2.5. This goodness of fit is supported by the GOF plots in Figure H.10, with close agreement between theoretical and empirical quantiles and probability throughout. A lognormal provided the second-best description for this subset of homes.

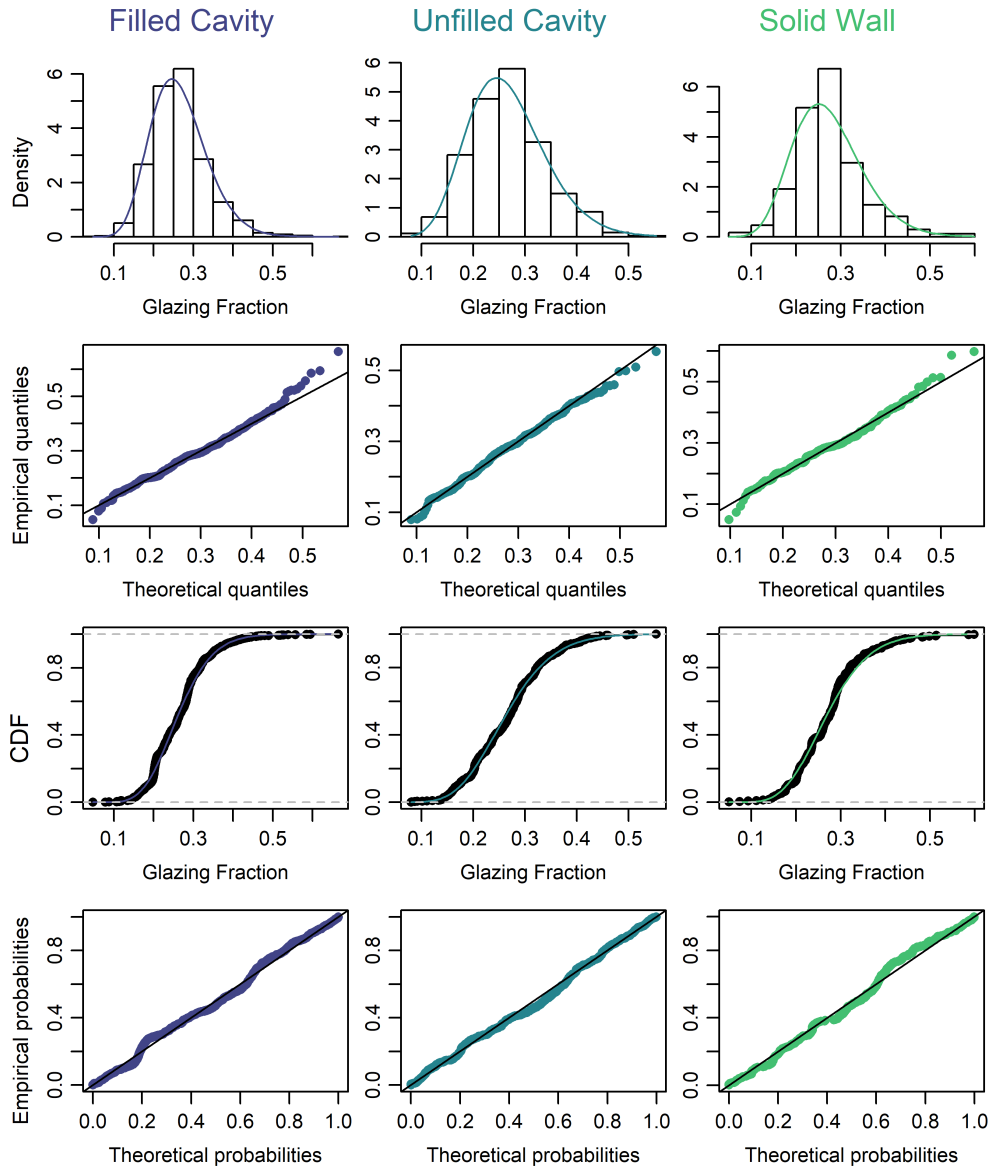


Figure H.8: Goodness of fit plots when a gamma distribution is fitted to the glazing fraction of semi-detached homes in the English Housing Survey.

H.6 Floor area factor

Five candidate distributions were fitted for the filled cavity and the solid wall constructions; distributions were not fitted to unfilled cavity wall data due to the small sample size (13 homes) and possible bimodality (Figure H.11).

The best fitting distribution for either construction type is the inverse Weibull. The goodness of fit plots for the fitted distribution of the filled cavity and solid wall are shown in Figure H.12. In the case of the filled cavity wall, the fitted distribution

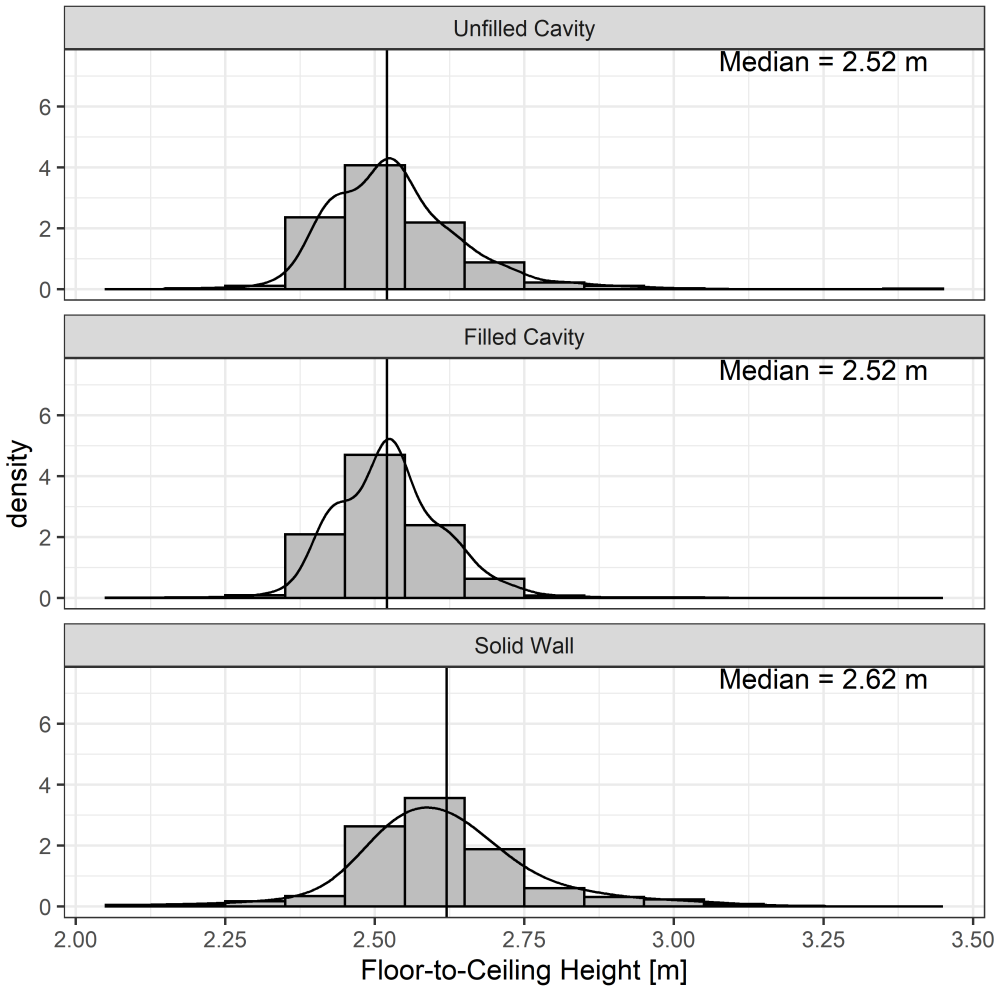


Figure H.9: Histogram of the Floor-to-Ceiling Height measurements for semi-detached dwellings in the EHS, separated by wall construction. This is the average of the main bedroom and living room measurement plus 0.125 m. The solid vertical line indicates the median.

provides a good description for all but one data point. For the solid wall, a few more points deviate away within the Q-Q plot and P-P plot, yet the distribution provides a good description of most data points.

Table H.3: Distributions fitted to floor-to-ceiling height data categorised per wall type. They are ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j).

Wall	Distributions	AIC	Δ_j	w_j	P1	P2
	lnorm	-2777	0	0.96	meanlog = 0.93	sdlog = 0.034
	gamma	-2771	7	0.04	shape = 840	rate = 330
FCW	norm	-2755	22	0.00	mean = 2.5	sd = 0.087
	invweibull	-2525	252	0.00	shape = 27	scale = 2.5
	weibull	-2251	526	0.00	shape = 24	scale = 2.6
	invweibull	-1108	0	1.00	shape = 27	scale = 2.5
	lnorm	-1059	48	0.00	meanlog = 0.93	sdlog = 0.043
UCW	gamma	-1047	60	0.00	shape = 520	rate = 210
	norm	-1021	87	0.00	mean = 2.5	sd = 0.11
	weibull	-599	509	0.00	shape = 15	scale = 2.6
	lnorm	-363	0	0.62	meanlog = 0.96	sdlog = 0.054
	gamma	-361	1	0.34	shape = 340	rate = 130
SW	norm	-357	5	0.04	mean = 2.6	sd = 0.14
	invweibull	-261	102	0.00	shape = 16	scale = 2.5
	weibull	-256	107	0.00	shape = 16	scale = 2.7

Table H.4: Distributions fitted to the floor area factor dataset grouped by wall type. These are ranked in decreasing order of goodness of fit based on the Akaike Information Criterion (AIC), difference in AIC (Δ_j) and Akaike weights (w_j). The corrected AIC was used for both groups. P1 and P2 represent the parameters of the fitted distribution, stated to two significant figures.

Wall	Dist.	AIC	Δ_j	w_j	P1	P2
	invweibull	-17.00	0.00	0.94	shape = 5.5	scale = 0.74
	lnorm	-11.00	6.00	0.05	meanlog = -0.2	sdlog = 0.23
FCW	gamma	-8.00	9.00	0.01	shape = 17	rate = 20
	norm	0.00	17.00	0.00	mean = 0.84	sd = 0.23
	weibull	3.00	20.00	0.00	shape = 3.5	scale = 0.93
	invweibull	-13.00	0.00	0.89	shape = 5	scale = 0.76
	lnorm	-9.00	4.00	0.10	meanlog = -0.17	sdlog = 0.24
SW	gamma	-4.00	9.00	0.01	shape = 16	rate = 18
	norm	8.00	21.00	0.00	mean = 0.88	sd = 0.25
	weibull	12.00	25.00	0.00	shape = 3.2	scale = 0.97

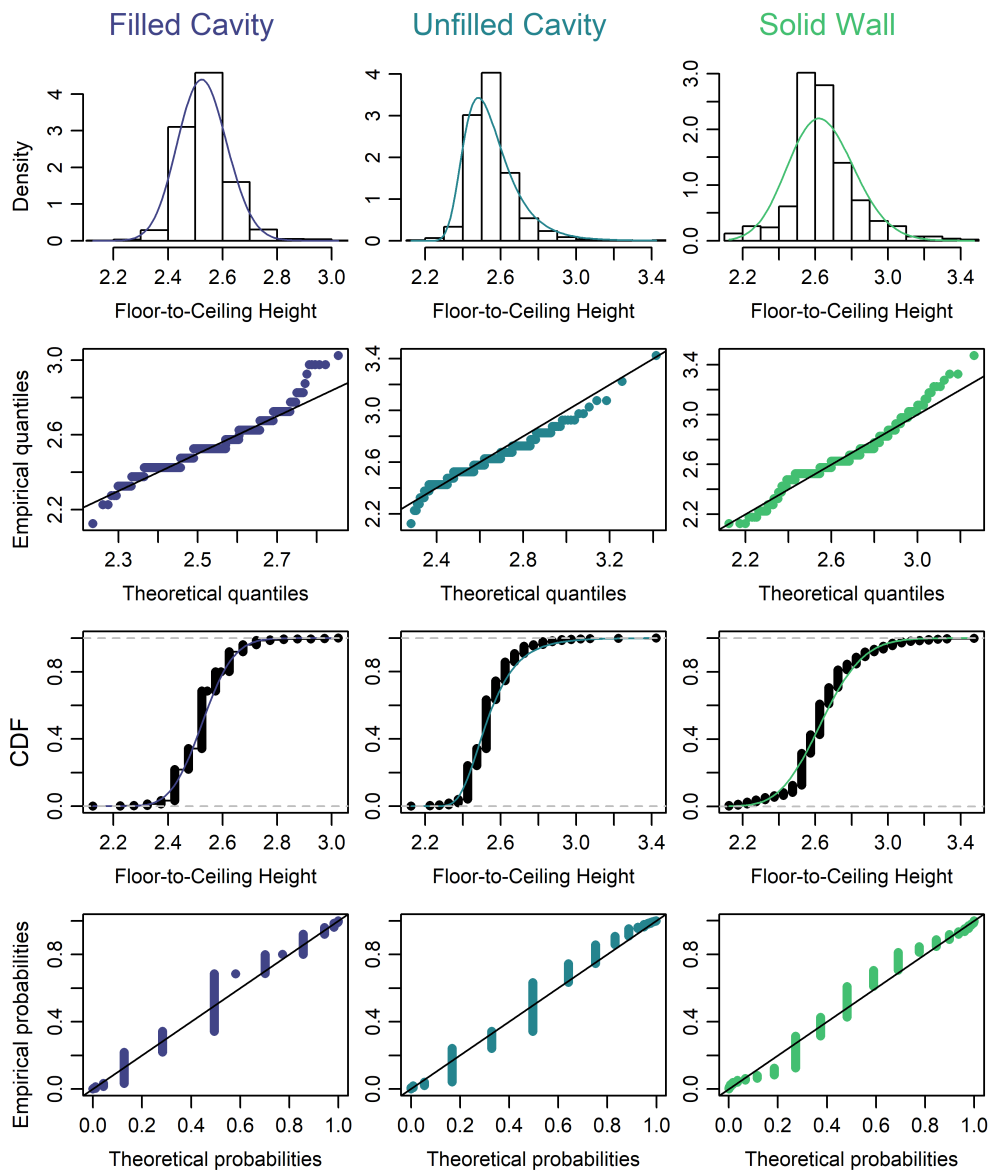


Figure H.10: Goodness of fit plots the floor-to-ceiling height of semi-detached homes within the 2011 English Housing Survey. A lognormal was fitted to the filled cavity and solid wall subgroup, while an inverse Weibull was fitted to the unfilled cavity wall subgroup.

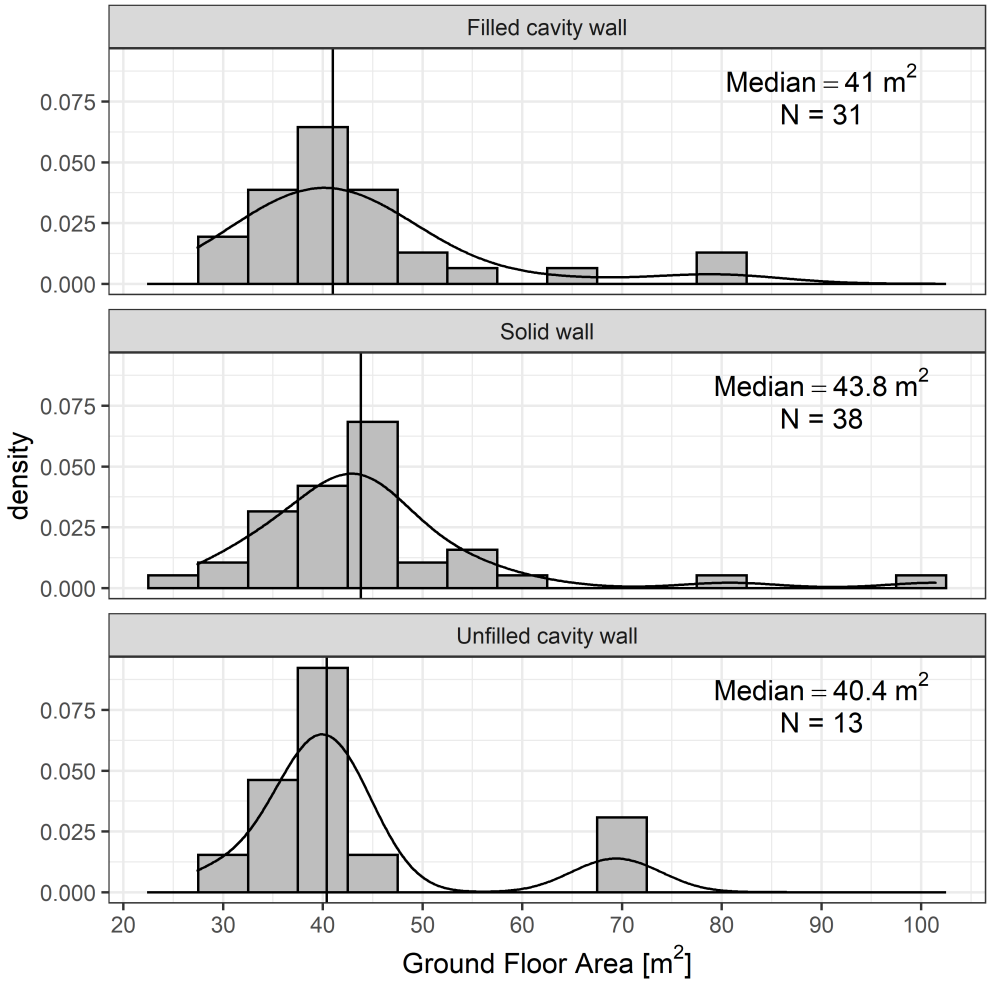


Figure H.11: Histogram of the floor area measurements for semi-detached dwellings in the 4M dataset, separated by wall construction. The solid vertical line indicates the median.

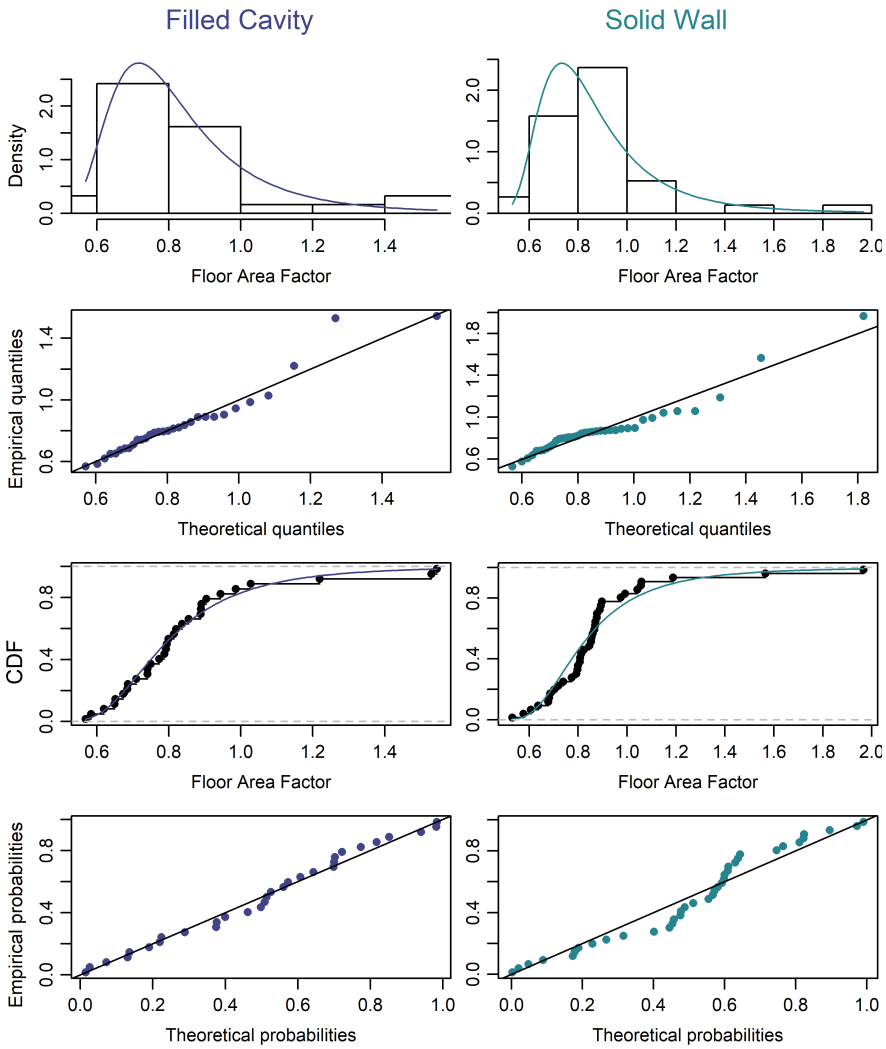


Figure H.12: Goodness of fit plots floor area factor of semi-detached homes in 4M. The filled cavity and solid wall groups were both fitted with an inverse Weibull distribution.