Bayesian Calibration Of Residential Building Clusters Using A Single Geometric Building Representation

Martin Heine Kristensen[1*], Ruchi Choudhary[2], Rasmus Høst Pedersen[3], Steffen Petersen[1]
[1]Department of Engineering, Aarhus University, 8000 Aarhus C, DK
[2]Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK
[3]AffaldVarme Aarhus, Bautavej 1, 8210 Aarhus V, DK
[*]Corresponding author (mhk@eng.au.dk)

Abstract

For a homogeneous cluster of single-family dwellings, an archetype model incorporating simple scalable geometry and an hourly dynamic building energy model was set-up to represent its energy performance. Using metered annual energy use for a random sample of 450 buildings in the cluster, the archetype model was calibrated in a Bayesian regression framework using the floor area as common scale for regression of the physics-based input to the hourly dynamic energy model. In this process, posterior estimates of seven selected building parameters shared by buildings within the cluster were inferred. The calibrated archetype model was used to make predictions of annual building energy use with a normalized mean bias error (NMBE) of 2.3 % and a coefficient of variation of the root mean squared error (CVRMSE) of 26.5 %.

Introduction

Building energy modeling (BEM) often relies on physics-based principles for representing the thermodynamic mechanisms of buildings (see Kavgic et al. (2010) for a review on physics-based approaches). To have better estimates of the inputs to such building models, it is often advantageous to calibrate them using appropriate field data. In the case where a vast amount of buildings is to be modeled for analysis, e.g. the housing stock of a larger city, it might be practical to model a group of similar buildings as a homogeneous cluster, also known as archetype modeling. However, as emphasized by Reinhart and Davila (2016), the inevitable presence of heterogeneity among buildings within an archetype can make it difficult to identify a deterministic best fit of physical model parameters in the calibration process. Using instead probabilistic calibration, one may incorporate and propagate parameter uncertainty more appropriately including both aleatory uncertainty (i.e. heterogeneity across a building cluster) and epistemic uncertainty (i.e. uncertainty about the true value of parameters).

Booth et al. (2012) used a monthly average quasi-steady-state energy model in combination with the Bayesian calibration framework by Kennedy and O'Hagan (2001) to set up a probabilistic model and calibrate uncertain model parameters for a group of 35 identical flats using energy performance certificates data and metered energy use. In Booth et al. (2013), this framework was expanded to the scale of housing stock models using synthetic/regressed calibration data from a macro scale and GIS-recorded building geometry. In a recent study by Sokol et al. (2017), a residential building stock was subdivided into archetypes and calibrated using the Bayesian framework and individual EnergyPlus BEMs to set up an aggregated urban building energy model (UBEM). They used monthly and annually metered calibration data, in combination with simple building-specific property information from tax assessments and GIS-based estimates of the geometry, to achieve a a mean error of 44 %-47 % and a CVRMSE of 58 %-66 % for a validation sample of 2263 buildings.

In all of these studies, information about the geometry of the individual buildings of the cluster was available (or inferred); however, in many cases, this information may be sparse or difficult to collect. In this study, we therefore investigate whether a single geometric and scalable building representation – an archetype model – modeled using an hourly dynamic BEM can adequately represent the variations within a cluster of similar residential buildings. Using the Kennedy and O'Hagan (2001) calibration methodology that relies on Bayesian propagation of uncertainties, we train the model using metered district heating consumption data, and information about the construction year and building footprint area from 450 buildings belonging to the archetype, and validate it against an unseen sample of 150 buildings to test the reliability in the inferred results. As a result, both the aleatory and epistemic uncertainties in model parameters are quantified for the cluster to perform highly accurate archetype model predictions.

Method

Characteristics of the building cluster

A building cluster consisting of detached single-family dwellings from the Danish building stock was iden-

tified based on the typological characterizations defined by the European research project TABULA (Wittchen and Kragh, 2012; Loga et al., 2016). The cluster represents buildings constructed between 1979-1998 in Denmark, a period with rather uniform building traditions and non-varying building codes. The buildings within the cluster were therefore assumed to exhibit similar building-physical properties. In fact, the value and state of unknown technical properties, whether it be the U-values or the air tightness of the envelope, etc., were assumed independent and identically distributed samples drawn from the same population. Pictures of typical buildings represented by the cluster are shown in Figure 1.

Figure 1: Examples of typical buildings of the cluster.

A total sample of $N = 2000$ existing buildings situated in the district heating grid of Aarhus, Denmark, were sampled from the cluster. Information about the construction year, building footprint area, number of floors and envelope materials for each building i within the cluster sample was collected from the Danish Building and Dwelling Register (BDR), which is publicly available (Table 1). The aggregated annual district heating energy use, E^{DH}, for the last three years (2013-2015) was collected for all N buildings from the local district heating supplier (AffaldVarme Aarhus). As indicated in (1), E^{DH} consists of energy use for hydronic space heating, E^{SH}, and energy use for on-site domestic hot water preparation, E^{DHW}. The data does not allow any distinction between how much energy is spent on space heating and DHW, respectively.

$$E_i^{DH} = E_i^{SH} + E_i^{DHW}. \tag{1}$$

The expected mean annual energy use per building, $\overline{E^{DH}}$,was computed as the direct average of the metered consumption over the last three years (Table 1) in accordance with (2). It was assumed that this three-year average represented the true mean.

$$\overline{E_i^{DH}} = \frac{1}{3}\sum_{y=1}^{3} E_{i,y}^{DH}. \tag{2}$$

Geometric building representation

The actual geometry of the individual buildings in the sample was unknown. The individual building geometry was instead estimated using a rectangular box (Figure 2) with dimensions based on simple rules applied to known information about the floor area, A_{floor}, and number of floors, n_{floors}, (Table 1), and our a-priori beliefs about the length-to-width-ratio, ρ_{LW}, and floor height, h_{floors}, (Table 2). In the case of two or three-storey buildings with unequal floor areas, A_{floor} was averaged over the floors to fit the geometric model. The four facades were assumed to face the four cardinal directions (North, East, South, West) while the unknown window area in each facade was allowed to vary for each building (Table 2).

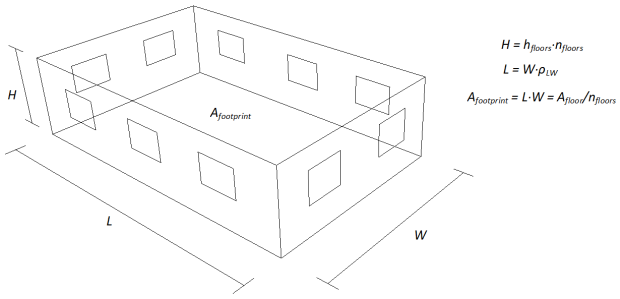

Figure 2: Geometric model of archetype building.

Building energy model for space heating and domestic hot water

The energy use for space heating, E_{SH}, was modeled using the simple hourly dynamic model of ISO 13790:2008 (International Organization for Standardization, 2008), treating the building as a single thermal zone. The Danish building code energy requirements in force at the time of construction (BR77, BR82 and BR85), historical surveys of the Danish building stock (Bøhm et al., 2009; Bergsøe, 2015), and our educated guesses were used to estimate the majority of the technical and occupation-related inputs required to run the model (Table 2). Ventilation was assumed a mix of infiltration and opening of windows. Based on studies by Rijal et al. (2007), the airflow through windows was modeled hourly as a percentage of maximum design airflow using a logistic transformation of a linear regression on the outdoor temperature (3) where T_{out} is the outdoor temperature, and a and b are empirical regression coefficients (Table 2).

$$p = \frac{\exp\left(aT_{out} + b\right)}{1 + \exp\left(aT_{out} + b\right)}. \tag{3}$$

Internal heat loads from occupants were modeled as a scalable day profile for the activity level constructed with a variable period without any presence called Away time (Figure 3).

Energy use for DHW, E_{DHW}, was modeled using a simple linear model proportional to the consumed amount of hot water under the assumption that the rate of consumption was approximately constant

Table 1: Data quantiles of known cluster characteristics ($N = 2000$).

Name	Unit	Quantiles				
	-	**2.5%**	**25%**	**Median**	**75%**	**97.5%**
Construction year	-	1979	1982	1987	1995	1998
Number of floors*, n_{floors}	-	1	1	1	1	1
Floor area, A_{floor}	m^2	98	130	149	170	238
Energy use, $\overline{E^{DH}}$	MWh/year	7.20	11.85	14.65	17.75	28.60
Energy use intensity**, EUI	kWh/m^2/year	54.7	83.4	97.9	112.8	155.3

* 97.8 % of the buildings had one floor, 2.15 % had two floors, and 0.05 % had three floors.
** Calculated quantity, $EUI = \overline{E^{DH}}/A_{floor}$.

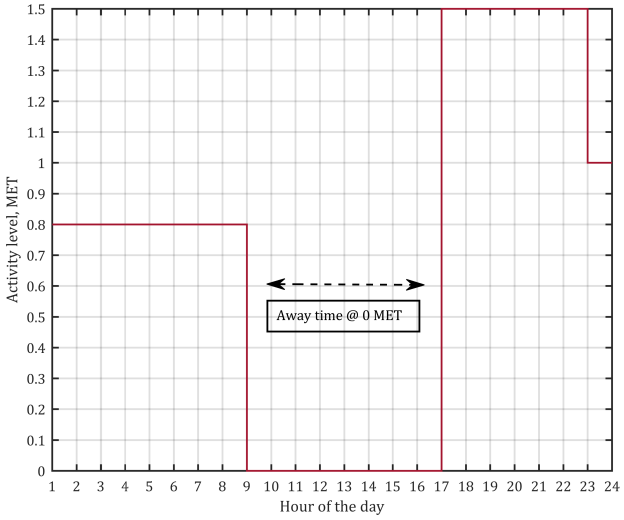

Figure 3: Day profile for occupant heat loads with variable period without presence.

throughout the year:

$$E_{DHW} = \rho c V_{occ.} n_{occ.} \left(T_{flow} - T_{mains}\right) + \sum_{t=1}^{L} H_{circ.}^{(t)} \left(T_{flow} - T_{air}^{(t)}\right). \quad (4)$$

The room air temperature, T_{air}, was modeled dynamically in the ISO 13790 model, and the volumetric heat capacity of the water, ρc, was given as 4140 kJ/(m^3K). The remaining model parameters were unknown, i.e. $V_{occ.}$ (hot water consumption per occupant per year), $n_{occ.}$ (number of occupants), T_{flow} (flow temperature), T_{mains} (mains supply temperature), and $H_{circ.}$ (specific heat loss in the circulation system). An additional time varying term was added to account for residual hot water circulation loss to the building, assuming that the temperature difference between hot water and room air was proportional. This loss was also added as an internal heat gain in the dynamic model for calculation of energy for space heating.

The simulated district heating energy use should represent the true mean energy use, hence average weather conditions of the outdoor temperature and global solar irradiation were needed as boundary conditions. For this end, the Danish design reference year (DRY) was applied; a set of hourly weather conditions composed from historical measurements in Denmark used for designing and testing buildings (Jensen and Lund, 1995).

Sensitivity analysis for selecting calibration parameters

Prior to calibration of the archetype model, 32 input parameters were unknown. For each parameter, probability density functions (PDFs) were assigned to reflect our beliefs about their values; distribution types *Uniform(min, max)*, *Beta(shape1, shape2)*, *Gamma(shape, scale)*, and *Normal(mean, variance)* were applied (Table 2). As these distributions rely on a-priori information, they are referred to as *a-priori distributions* or simply *priors*.

A probabilistic sensitivity analysis using the method of Sobol' (Sobol', 1993) was conducted to rank the 32 unknown model input parameters, given their prior distributions, in descending order of importance (Table 2). Based on recommendations by Kristensen and Petersen (2016), who analyzed the performance of three different sensitivity analysis methods on the ISO 13790 energy calculation models, the Sobol' total effects index, ST_i, was applied as a measure of explaining individual parameters' combined effect on model output variability. A total of 15 000 Monte Carlo iterations were performed to obtain convergence of the Sobol' algorithm, $\sum(ST_i) \approx const.$ No correlations between model parameters were taken into account.

Ideally, one ought to calibrate all uncertain model parameters; however, as this is computationally infeasible, a limited number of parameters must be chosen. We selected the seven highest ranked parameters from the sensitivity analysis, which in total accounted for approx. 75 % of the model output variability (Figure 4). In descending order of importance, the selected parameters were: the heating set point, U-value (windows), appliances heat load, infiltration rate @50 Pa, window-wall ratio (North), window-wall ratio (South), and U-value (walls). The remaining 25 parameters were left uncalibrated with their prior uncertainty specification being propagated through the

model.

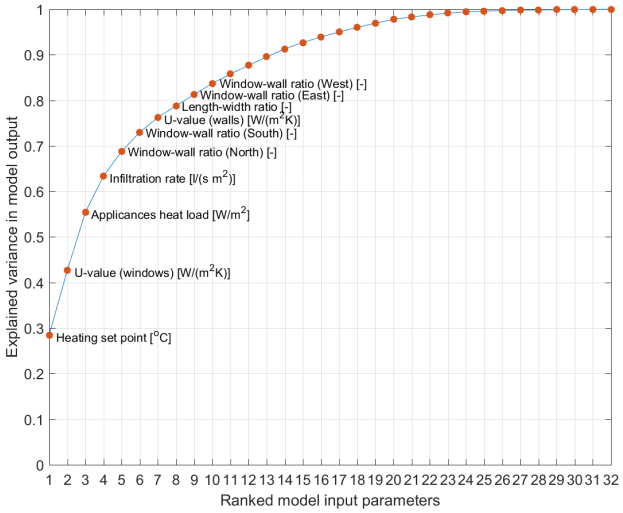

Figure 4: Cumulative sum of explained variance in model output (ST_i) by ranked input parameters.

Calibration framework

The Bayesian approach used in this study followed the approach proposed by Kennedy and O'Hagan (2001). It explicitly incorporates uncertainty in model inputs, uncertainty due to limited numbers of simulation runs, and discrepancy between the model outputs and the actual energy-consuming mechanism of the buildings. We represented the measured energy use data (Table 1) statistically as

$$y_i^{meas} = \zeta\left(\boldsymbol{x}_i^{meas}\right) + e_i^{meas} \quad i = 1, \ldots, n, \qquad (5)$$

where y_i^{meas} is the field data for the i-th building, i.e. measured annual energy use in kWh/year, each observed by a p-dimensional vector of known explanatory design points, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, for which $x \in \mathbb{R}^p$. In this study, we applied the known heated floor area as the only explanatory variable; hence $p = 1$. The field data, y_i^{meas}, was assumed to be realizations from the true energy-consuming process, $\zeta\left(\boldsymbol{x}_i\right)$, at the observed design points, $\boldsymbol{x}_i^{meas}$, with an unknown observation error, e_i^{meas}. We modeled the true process as:

$$y_i^{meas} = \eta\left(\boldsymbol{x}_i^{meas}, \boldsymbol{\theta}\right) + \delta\left(\boldsymbol{x}_i^{meas}\right) + e_i^{meas} \quad i = 1, \ldots, n, \qquad (6)$$

using the building energy model as an emulator, $\eta(\boldsymbol{x}_i, \boldsymbol{\theta})$, evaluated at the $(p+q)$-dimensional input vector $(\boldsymbol{x}_i^{meas}, \boldsymbol{\theta})$ comprising the vector of known design points, $\boldsymbol{x}^{meas}$, and additional calibrated parameters, $\boldsymbol{\theta} \in \mathbb{R}^q$, represented by the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)$. For this study, we took the top-seven ranked model parameters (Figure 4) as the calibration parameters; hence $q = 7$. As no model is perfect, a stochastic model bias was introduced through $\delta(\cdot)$ independently of $\eta(\cdot, \cdot)$ to account for discrepancy between the model and the true physical process.
The inclusion of a noise-term, e_i^{meas}, allowed for different observations of y at identical settings of $\boldsymbol{x}$. In addition to measurement error, the true archetype energy-consuming process may be inherently unpredictable and stochastic due to e.g. occupant behavior. Thus, strictly speaking, e_i^{meas} includes any residual variability that cannot be decomposed in addition to measurement error. This also includes violations of the assumption of cluster homogeneity. We denote these different sources of uncertainty (archetype heterogeneity, measurement error, and chance variability) *aleatory uncertainty.*
The *true* or *best achievable* distributions of the calibration parameters, $\boldsymbol{\theta}$, are to be inferred in the calibration process, and are thus unknown. They are referred to as *a-posteriori* distributions or simply *posteriors.* The posterior distributions represent the *epistemic parameter* uncertainty, that is uncertainty due to modeling assumptions of the BEM, our estimates of the prior distributions, and the amount of calibration data available for a precise identification of the true parameter values. We represented the calibration parameters in the simulations in terms of their priors, $\boldsymbol{t} = (t_1, t_2, \ldots, t_q)$, given in Table 2. Evaluating the BEM at randomly selected design points, $\boldsymbol{x}_j^{sim}$, and at random samples of the prior distributions, $\boldsymbol{t}_j$, yields simulated annual energy use data denoted y_j^{sim}.

$$y_j^{sim} = \eta\left(\boldsymbol{x}_j^{sim}, \boldsymbol{t}_j\right) + e_j^{sim} \quad j = 1, \ldots, m. \qquad (7)$$

As proposed by Higdon et al. (2004), we added a small numerical white noise error-term, e_j^{sim}, in the statistical representation (7) to secure the positive definiteness of the later covariance matrix (16) necessary for the Gaussian process (GP) regression. We modeled both e_i^{meas} and e_j^{sim} as i.i.d. Gaussian distributed noise:

$$e_i^{meas} \sim N\left(0, \sigma_{meas}^2\right), \qquad (8)$$

$$e_j^{sim} \sim N\left(0, \sigma_{sim}^2\right). \qquad (9)$$

For the evaluation of the model, $\eta(\cdot, \cdot)$, the remaining 25 non-calibrated input parameters were sampled at random for each simulation run j. As the value of the functions $\eta(\cdot, \cdot)$ and $\delta(\cdot)$ are known only at the applied design points, they were modeled as Gaussian processes (GP) represented in terms of multivariate Gaussian distributions (10)-(11), allowing an infinite number of different regressions to be fit. Other formulations are available though, e.g. multiple linear regression as proposed by Li et al. (2016). A key property of the applied GP regression model is that it obtains a perfect fit to the training data. However, due to the inclusion of the observation error, e_i^{meas},

Table 2: Prior distributions and ranking of 32 unknown model parameters using the Sobol' sensitivity analysis method.

Parameter	Unit	PDF	Quantiles					Sensitivity Analysis	
			2.5%	25%	Median	75%	97.5%	STi*	Rank
Geometry									
Length-width ratio	-	Unif(0.3;3.0)	0.4	1.0	1.7	2.3	2.9	2.6%	8
Room height	m	Unif(2.1;3.1)	2.12	2.35	2.60	2.85	3.08	1.2%	-
Window-wall ratio (North)	-	Beta(3;3)	0.10	0.30	0.40	0.50	0.75	5.4%	5
Window-wall ratio (East)	-	Beta(3;3)	0.10	0.30	0.40	0.50	0.75	2.5%	9
Window-wall ratio (South)	-	Beta(3;3)	0.10	0.30	0.40	0.50	0.75	4.2%	6
Window-wall ratio (West)	-	Beta(3;3)	0.10	0.30	0.40	0.50	0.75	2.4%	10
Window frame fraction	-	Beta(8;24)	0.12	0.20	0.24	0.30	0.41	0.5%	-
Transmission									
Temp. adjustment factor (ground)	-	Beta(10;1.1)	0.78	0.87	0.91	0.94	0.98	0.1%	-
U-value (floors)	$W/(m^2\,K)$	Gam(7;0.029)	0.08	0.15	0.19	0.24	0.37	1.7%	-
U-value (walls)	$W/(m^2\,K)$	Gam(7;0.043)	0.12	0.22	0.29	0.37	0.56	3.3%	7
U-value (basement)	$W/(m^2\,K)$	Gam(4;0.088)	0.10	0.22	0.32	0.45	0.77	1.4%	-
U-value (roofs)	$W/(m^2\,K)$	Gam(7;0.029)	0.08	0.15	0.19	0.24	0.37	2.1%	-
U-value (windows)	$W/(m^2\,K)$	Gam(50;0.05)	1.86	2.25	2.48	2.73	3.24	14.2%	2
SHGC (windows)	-	Beta(45;32.6)	0.49	0.56	0.60	0.64	0.71	0.4%	-
Internal shading coef.	-	Unif(0.2;0.9)	0.22	0.38	0.55	0.73	0.88	0.1%	-
Internal heat capacity	$kJ/(m^2\,K)$	Gam(10,17)	59	138	225	312	391	1.0%	-
Effective mass area	m^2/m^2	Gam(70;0.04)	2.2	2.6	2.8	3.0	3.5	<0.1%	-
Heat conduction (mass)	$W/(m^2\,K)$	Gam(350;0.026)	7.1	8.3	9.1	9.8	11.3	<0.1%	-
Heat transfer coef. (surf.-air)	$W/(m^2\,K)$	Gam(350;0.01)	3.1	3.3	3.4	3.6	3.8	<0.1%	-
Ventilation									
Infiltration airflow @ 50 Pa	$l/(s\,m^2)$	Gam(5;0.50)	0.97	2.02	2.80	3.76	6.15	7.9%	4
Design airflow (windows)	$l/(s\,m^2)$	Gam(10;0.04)	0.19	0.31	0.39	0.48	0.68	0.9%	-
Regression coef. a	-	Unif(0.1;0.4)	0.11	0.17	0.25	0.33	0.39	0.3%	-
Regression coef. b	-	Unif(-4;-1)	-3.92	-3.25	-2.50	-1.75	-1.07	1.9%	-
Occupation									
Occupant density	m^2/pers	Gam(10;5.18)	25	40	50	62	88	1.9%	8
Occupant heat load	W/(MET pers)	Gam(50;1.60)	60	70	80	90	100	<0.1%	-
Appliances heat load	W/m^2	Gam(40;0.75)	21	27	30	33	40	12.8%	3
Away time @ 0 MET	h/day	Unif(0;12)	0.3	3.0	6.0	9.0	11.7	0.5%	-
Room heating set point	°C	Norm(20.5;22)	16.6	19.2	20.5	21.8	24.4	28.5%	1
Domestic hot water									
DHW flow temperature	°C	Gam(200;0.263)	42.7	48.9	52.3	56.0	63.3	0.9%	-
Mains temperature	°C	Gam(100;0.125)	10.2	11.6	12.5	13.3	15.1	<0.1%	-
Specific circulation loss	W/K	Gam(25;0.32)	5	7	8	9	11	<0.1%	-
Hot water consumption	$m^3/(pers\,year)$	Gam(25;0.75)	9.7	12.9	14.8	16.9	21.4	1.1%	-

* Sobol' total effects index, $0 \leq ST_i \leq 1$

the training data was treated as noisy measurements letting instead the GP model identify and fit the *true* mean energy use at each value of $\boldsymbol{x}$ by attributing fluctuations from the archetype mean energy use to the noise-term.

$$\eta(\cdot,\cdot) \sim N\left(0, \Sigma_\eta\right). \tag{10}$$

$$\delta(\cdot) \sim N\left(0, \Sigma_\delta\right). \tag{11}$$

By standardizing the data $(y, \boldsymbol{x}, \boldsymbol{t}, \boldsymbol{\theta})$ to a range $[0;1]$, the two GPs were specified with constant mean functions of zero and squared exponential covariance functions (12)-(13), allowing a smooth and stationary representation of the process less prone to numerical problems:

$$\Sigma_\eta = \sigma_\eta^2 \exp\left(-\sum_{d=1}^{p} \beta_{\eta,d} \left|x_{i,d} - x'_{i,d}\right|^2\right) \cdot \exp\left(-\sum_{d'=1}^{q} \beta_{\eta,p+d'} \left|t_{i,d} - t'_{i,d}\right|^2\right), \tag{12}$$

$$\Sigma_\delta = \sigma_\delta^2 \exp\left(-\sum_{d=1}^{p} \beta_{\delta,d} \left|x_{i,d} - x'_{i,d}\right|^2\right). \tag{13}$$

In (12)-(13), σ_η^2 and σ_δ^2 control the marginal function variance of the energy-consuming process, i.e. variation that is explained by the GP regression function, while β_η and β_δ are weighting parameters for the d-th dimension of the input space controlling the strength of each model parameter as a predictor. The larger β is, the more dimension d is weighted in the summation of the Euclidean distances. We refer to these variance and weighting parameters as hyperparameters of the GPs. They were unknown and hence to be inferred in the calibration process. The total number of

unknown model parameters were thus comprised by calibration parameters $\boldsymbol{\theta}$, variance parameters σ^2_{meas}, σ^2_{sim}, σ^2_η and σ^2_δ, and weighting parameters β_η and β_δ.

Defining a $(n+m)$ observation vector, $\boldsymbol{z} = (\boldsymbol{y}^{meas,T}, \boldsymbol{y}^{sim,T})^T$, containing n field observations $\boldsymbol{y}^{meas} = (y_1^{meas}, \ldots, y_n^{meas})^T$, and m simulation runs $\boldsymbol{y}^{sim} = (y_1^{sim}, \ldots, y_m^{sim})^T$, the regression model becomes the following by combining (5) and (6):

$$\begin{aligned}\boldsymbol{z} &= \begin{bmatrix} y_1^{meas} \\ \vdots \\ y_n^{meas} \\ y_1^{sim} \\ \vdots \\ y_m^{sim} \end{bmatrix} \\ &= \begin{bmatrix} \eta\left(\boldsymbol{x}_1^{meas}, \boldsymbol{\theta}\right) + \delta\left(\boldsymbol{x}_1^{meas}\right) + e_1^{meas} \\ \vdots \\ \eta\left(\boldsymbol{x}_n^{meas}, \boldsymbol{\theta}\right) + \delta\left(\boldsymbol{x}_n^{meas}\right) + e_n^{meas} \\ \eta\left(\boldsymbol{x}_1^{sim}, \boldsymbol{t}_1\right) + e_1^{sim} \\ \vdots \\ \eta\left(\boldsymbol{x}_m^{sim}, \boldsymbol{t}_m\right) + e_m^{sim} \end{bmatrix}. \end{aligned} \tag{14}$$

Applying Bayes theorem, the joint posterior distribution of the parameters, conditional on the augmented observation vector $\boldsymbol{z}$, is obtained:

$$\begin{aligned} &p(\boldsymbol{\theta}, \sigma^2_{meas}, \sigma^2_{sim}, \sigma^2_\eta, \sigma^2_\delta, \beta_\eta, \beta_\delta | \boldsymbol{z}) \\ &\quad \propto L(\boldsymbol{z}|\boldsymbol{\theta}, \sigma^2_{meas}, \sigma^2_{sim}, \sigma^2_\eta, \sigma^2_\delta, \beta_\eta, \beta_\delta) \\ &\quad p(\boldsymbol{\theta}) p(\sigma^2_{meas}) p(\sigma^2_{sim}) p(\sigma^2_\eta) p(\sigma^2_\delta) p(\beta_\eta) p(\beta_\delta). \end{aligned} \tag{15}$$

The likelihood of $\boldsymbol{z}$ conditional on the parameters, $L(\cdot|\cdot)$, is computed from a multivariate Gaussian distribution with zero mean function and a combined $(n+m)\cdot(n+m)$ covariance matrix Σ_z:

$$\Sigma_z = \Sigma_\eta + \begin{bmatrix} \Sigma_\delta & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \mathbf{I}_n \sigma^2_{meas} & 0 \\ 0 & \mathbf{I}_m \sigma^2_{sim} \end{bmatrix}, \tag{16}$$

where $\mathbf{I}_n$ and $\mathbf{I}_m$ are the $n \cdot n$ and $m \cdot m$ identity matrices, respectively, to put the Gaussian noise-terms on the diagonal.

For this study, we performed 1000 simulations, $\boldsymbol{y}^{sim}$, using draws from the prior distributions ($m = 1000$), and applied a sub-sample of 450 randomly selected field data points, $\boldsymbol{y}^{meas}$, for calibration ($n = 450$) from the $N = 2000$ cluster sample.

Prior distributions

Prior information is a very important aspect of the Bayesian framework as it influences the calibration output through a weighted fitting of the prior information and data at hand. For the calibration parameters, $\boldsymbol{\theta}$, we applied the distributions defined in Table 2 to reflect our a-priori beliefs about their values. The standardized variance hyperparameters, σ^2_{meas}, σ^2_{sim}, σ^2_η and σ^2_δ, were parametrized as precision parameters, $\lambda = 1/\sigma^2$, for which Gamma distributions were applied constraining the values to $\lambda \in \mathbb{R}|0 < \lambda < \infty$. The weighting parameters, β_η and β_δ, were parametrized as $\rho = \exp(-\beta/4)$, constraining the values to $\rho \in \mathbb{R}|0 < \rho < 1$, making it convenient to apply Beta distributions (Table 3).

For the parametrised weighting parameters, ρ, most of the prior mass is placed on values of ρ near 1 indicating vague parameter strength.

MCMC algorithm and convergence for posterior inference

The multi-dimensional joint posterior distribution cannot be obtained analytically. Therefore, we employed a random walk Markov Chain Monte Carlo (MCMC) algorithm – the Metropolis-Hastings algorithm (Gelman et al., 2014) – whose equilibrium distribution is indeed an approximation of the joint posterior distribution. Four chains were run in parallel with randomly dispersed starting points in the parameter space to draw samples from the posterior distribution. For each chain, 14 000 MCMC samples were drawn with the first 7000 samples of the chains being considered cool, meaning that information about the starting point might still prevail. Samples from this cold period were thus discarded leaving only the warm part of the chains for analysis.

Convergence in the warm chains was monitored in terms of the potential scale reduction factor, $\hat{R}$, for which $\hat{R} \in \mathbb{R}|1 < \hat{R} < \infty$. It is an estimate of the scale with which the variations in the inferred parameter distributions might be reduced if the simulations were continued in the limit $n \to \infty$ ($\lim_{n\to\infty} \hat{R} \to 1$) (Gelman et al., 2014). $\hat{R}$ accounts for the within-chain and between-chain variance in the warm chains, simultaneously evaluating both the mixing and the stationarity of it. For $\hat{R} < 1.1$, a stable and converged estimation was considered for each parameter, respectively.

Results

The MCMC algorithm is considered converged for both the calibration parameters and the hyperparameters ($\hat{R} < 1.1$ for all parameters). The calibration parameters are shown in Figure 5 in terms of their prior and posterior (calibrated) distributions.

Using the posterior parameter distributions, 1000 draws from the GP regression model are shown in Figure 6 with posterior uncertainty given as 95% uncertainty bands. In Figure 6A, predictions from the combined regression model, $y = \eta\left(\boldsymbol{x}, \boldsymbol{\theta}\right) + \delta\left(\boldsymbol{x}\right) + \epsilon$, are shown; in Figure 6B, isolated predictions of the physics-based ISO 13790 emulator, $\eta\left(\boldsymbol{x}, \boldsymbol{\theta}\right)$, are shown; in Figure 6C, isolated predictions of the model bias term, $\delta\left(\boldsymbol{x}\right)$, explaining the inadequacy of the physical model are shown; and in Figure 6D, the pos-

Table 3: Standardized prior distributions for hyperparameters of covariance functions.

Hyperparameter	PDF	Constraints		Quantiles				
		Lower	Upper	2.5%	25%	Median	75%	97.5%
λ_η	Gam (10,0.10)	0.1	∞	0.5	0.8	1.0	1.2	1.7
λ_δ	Gam (10,3.33)	0	∞	16.0	25.7	32.2	39.7	56.9
λ_{meas}	Gam (10,20.0)	0	∞	95.9	154.6	193.4	238.3	342.0
λ_{sim}	Gam (10,1000)	100	2e5	4799	7721	9662	11907	17054
ρ_η	Beta(1.0;0.5)	0	1	0.05	0.44	0.75	0.94	1.0
ρ_δ	Beta(1.0;0.4)	0	1	0.06	0.51	0.82	0.97	1.0

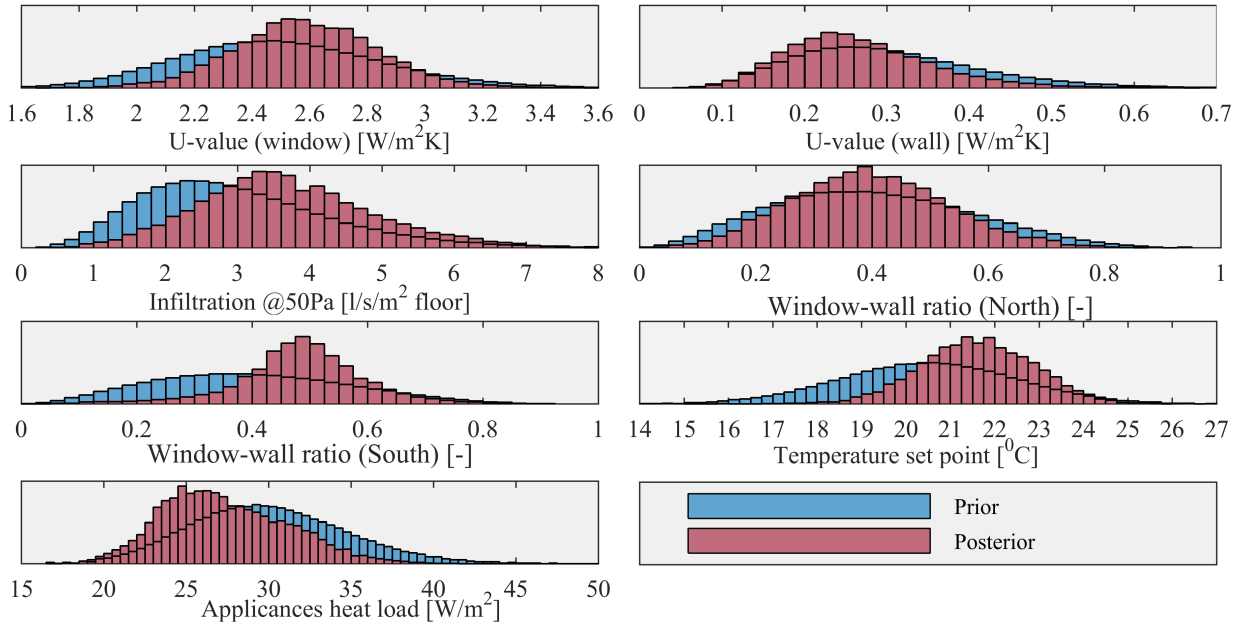

Figure 5: Prior PDFs (Table 2) and posterior PDFs (1000 draws) for calibration parameters.

terior distribution of the standard deviation of the observation error, σ_{meas}, is shown. It should be noted that the bias term only makes up a negligible part of the overall model predictions (approx. zero across the entire range of x), leaving predictions of the true energy-consuming process of the archetype building, $\zeta(\boldsymbol{x}) = \eta(\boldsymbol{x}, \boldsymbol{\theta}) + \delta(\boldsymbol{x})$ (noise-free predictions), to be based mainly on the physics-based term of the model. Following ASHRAE Guideline 14 (ASHRAE, 2014), the normalized mean bias error (NMBE), the coefficient of variation of the root mean squared error (CVRMSE), and the coefficient of determination (R^2) are used to assess model fit. These are supplemented with the mean absolute percentage error (MAPE) and shown in Table 4.

Discussion

The application of a common scalable geometric building model to fit all buildings of the cluster proves to be appropriate for the given field data. The regression model ascribes no significant value to the model bias term, $\delta(x)$, indicating that the field data conforms to the physics-based model (geometrical model in combination with hourly dynamic ISO 13790 model and DHW model). Given the constraints of the archetype model, the seven calibrated parameters are thus regarded true for the cluster.

As opposed to previous studies applying monthly quasi-steady-state housing stock modeling (Booth et al., 2012, 2013), the application of hourly dynamic modeling allows a more meaningful interpretation of the calibration parameters by permitting a better inclusion of time-varying parameters. However, additional work is required to improve the representation of individual occupant schedules, user pattern, and other highly stochastic and time-varying mechanisms. As long as the model does not fully and accurately capture the energy-consuming phenomena, the calibrated parameters can only be considered true when applied in connection with the applied model.

Goodness of fit

Regressing average annual building energy use as a function of heated floor area leaves approx. 60 % unexplained variance (R^2 = 42%). This residual variability the model has successfully identified and captured as i.i.d. Gaussian distributed observation error, e^{meas}, leaving only the true mean building energy use at a given floor area, $\zeta(x)$, to be fitted. Hence, in average, the regression model predicts the energy use of

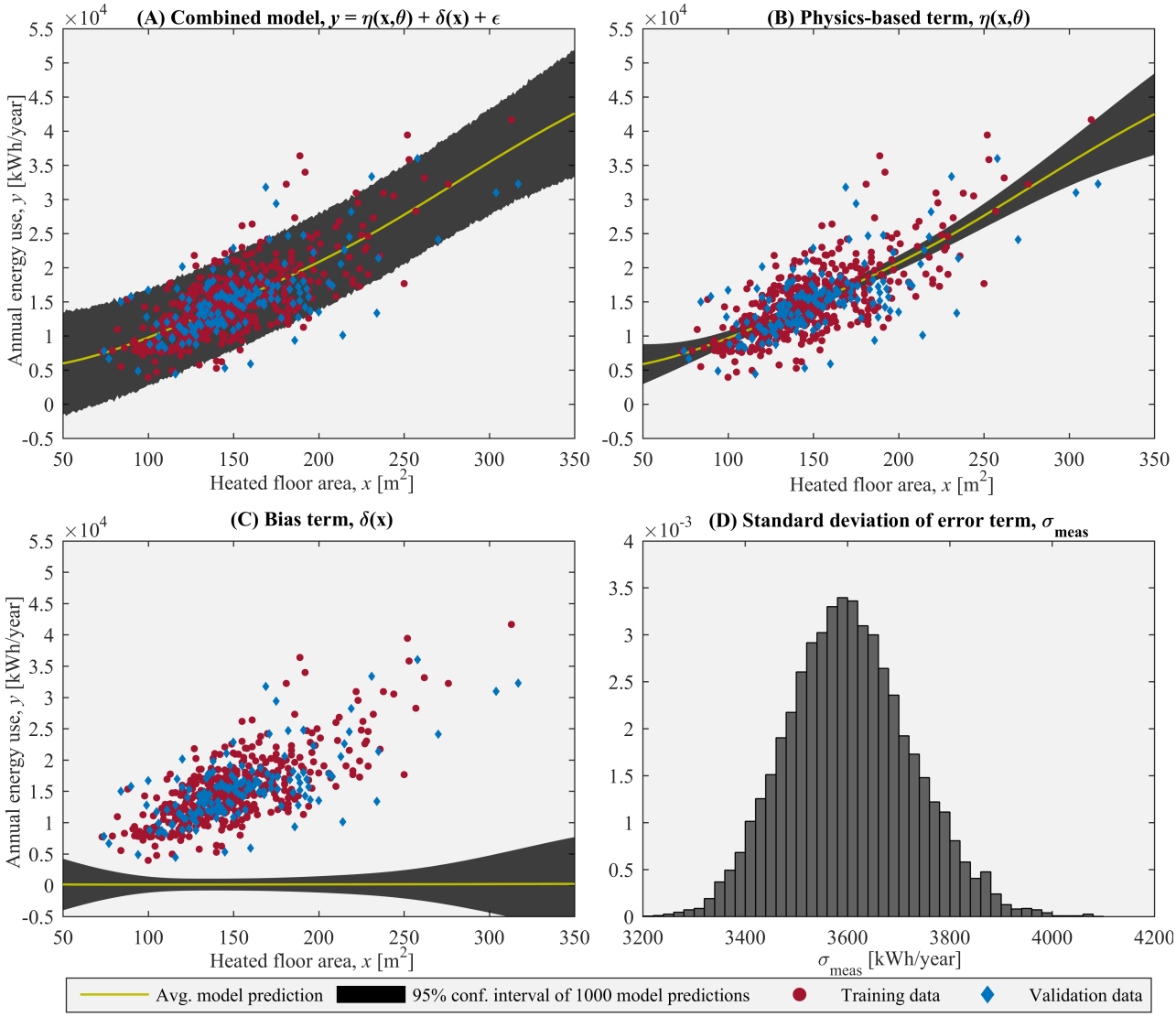

Figure 6: Archetype model predictions using 1000 draws from posterior distributions. (A) Combined regression model. (B) Physics-based term (ISO 13790 emulator) . (C) Bias term. (D) Histogram of posterior distribution of error term standard deviation.

Table 4: Measures of model predictive accuracy; NMBE (normalized mean bias error), MAPE (mean absolute percentage error), CVRMSE (coefficient of variation of the RMSE), and R^2 *(coefficient of determination). All measures were computed using posterior mean predictions of the true energy-consuming process,* $\zeta(\boldsymbol{x}) = \eta(\boldsymbol{x}, \boldsymbol{\theta}) + \delta(\boldsymbol{x})$ *(noise-free predictions).*

Dataset	**n**	**NMBE ± 95% conf. interval**	**MAPE ± 95% conf. interval**	**CVRMSE**	R^2
Training set	450	-0.3% ± 2.2%	20.0% ± 1.9%	24.1%	61.6%
Validation set	150	+2.3% ± 4.3%	21.9% ± 4.5%	26.5%	41.9%

the archetype within the ranges of the validation data (50-350 m^2) to an NMBE of approx. 2.3 % for the validation set. The NMBE measure holds for the average of large-sample predictions, e.g. aggregated predictions on the urban scale, but as the NMBE measure is subject to cancellation errors, the model predictive accuracy of single-building predictions may be more intuitively interpreted in terms of the MAPE, a measure of absolute error, resulting in approx. 22 %. This means that while the aggregated prediction error of a large sample of the archetype buildings is approx. 2.3 % in average, single-building predictions will be approx. 22 % uncertain in average. The model precision, i.e. the consistency of the model predictive accuracy, is assessed in terms of the CVRMSE, a normalized measure of variability between field data and model predictions across the ranges of the validation data (50-350 m^2). A CVRMSE score of 26.5 % seems

satisfactory given the fact that measurements originate from different buildings, and hence should not be confused with the ASHRAE Guideline 14 (ASHRAE, 2014) requirements for monthly calibrations (15 %) and hourly calibrations (30 %) as such requirements only cover single-building calibrations, where data is not subjected to the same level of heterogeneity and hence ought to be calibrated more easily.

Sources of error

As elaborated upon in the section *Calibration framework*, the observation error, e^{meas}, includes a mixture of uncertainties: geometrical heterogeneity between buildings of the archetype (even though the floor area is similar, the surface-to-volume ratio might not be), heterogeneity of technical properties between buildings (U-values, air tightness, etc.), variations due to occupant-related building operation (heating set point temperature, schedules, venting, domestic hot water, etc.), measurement error, and any residual variability that cannot be decomposed and ascribed to the parameters of the Gaussian process models. We assume the majority of this error to be driven by chance variability among buildings, predominantly occupant-related variation. However, also the assumption of archetype homogeneity is suspected to contribute significantly.

In order to fully understand the effect of these individual sources of uncertainty and ultimately secure a better and more reliable parameter calibration, additional work has to be performed to address the sources of error individually. In this context, the homogeneity assumption could be investigated in a comparison of single-building calibrations against the entire archetype calibration.

To further reduce model uncertainty, one could include additional explanatory input dimensions in $\boldsymbol{x}$, i.e. other design points known for each observation besides the heated floor area, e.g. number of occupants registered on the address, number of heating and/or cooling degree days in the year, level of refurbishment, etc. The application of Gaussian processes to model such diverse data spaces enables a very flexible framework that will encompass both linear and non-linear phenomena without relying on a specific model structure and, at the same time, allow the modeler to embed his prior knowledge if he wants to. This feature in particular is highly beneficial for sparsely distributed input spaces. However, we have to acknowledge that modeling the two-dimensional data presented here might have been done using a more parsimonious model, e.g. a linear model, without any noticeable difference in model predictive accuracy.

Scaling archetype predictions to the urban level

A natural next step would be to set up similar archetype models for the remaining building stock and use these for composing urban-scale building energy models to assess the aggregated energy use and subsequently the retrofit potential of such case studies. However, as advocated for by Reinhart and Davila (2016), the approach of scaling archetype models to the urban level by multiplying each individual archetype by the number of buildings in the urban area ignores the urban context, i.e. the presence of shading, local wind patterns, etc. Hence, one has to find a way to reliably compose aggregated models and, at the same time, account for the presence of such context-specific uncertainties, e.g. by inclusion of GIS-data.

Assessing retrofit potential

The probabilistic nature of the archetype model makes it suitable for assessing the averaged effect of building retrofits under uncertainty. This could be done by setting priors on selected physical parameters to reflect the anticipated, but uncertain, value of the parameters in question after implementation of energy conserving measures (ECM). Using this framework, the uncertainty of the ECM would then be propagated through the model to reflect its effect on the annual energy use of an average building represented by the archetype. The details of such an analysis, and how to actually conduct it, we leave for future work; however, inspiration for a similar study can be found in Booth and Choudhary (2013) who performed a retrofit analysis for the UK housing stock under uncertainty.

Conclusion

Using a Bayesian calibration technique it was demonstrated how a homogeneous cluster of single-family dwellings could be fitted to a single archetype geometry and hourly dynamic building energy model. The probabilistic treatment of the model allowed a quantification of the epistemic uncertainties embedded in the calibration parameters and the aleatory uncertainty in the cluster homogeneity assumption. The application of a single archetype model to fit the cluster showed to be adequate as all weight was attributed to the physics-based model, leaving only insignificant influence to the statistical model bias term. The calibrated archetype model can be used to make predictions of annual building energy use with a NMBE of 2.3 % and a CVRMSE of 26.5 %.

Acknowledgements

The research was conducted as part of the 'Resource Efficient Cities Implementing Advanced Smart City Solutions' (READY) project, work package 3, financed by the 7th EU Framework Programme (FP7-Energy project reference: 609127). Furthermore, the authors would like to thank the district heating company in Aarhus, AffaldVarme Aarhus, for supplying the building energy data that forms the basis of the study. Especially business development manager

Adam Brun deserves our gratitude, without whose help and cooperation this study would not have been possible.

References

ASHRAE (2014). *ASHRAE Guideline 14-2014 – Measurement of Energy, Demand, and Water Savings*. ASHRAE.

Bergsøe, N. C. (2015). *SBi 2015:25 Tæthed i eksisterende bygninger – Analyse af målte værdier (in Danish)*. Danish Building Research Institute, Aalborg University.

Bøhm, B., F. Schrøder, and N. C. Bergsøe (2009). *SBi 2009:10 Varmt Brugsvand – Måling af forbrug og varmetab fra cirkulationsledninger (in Danish)*. Danish Building Research Institute, Aalborg University.

Booth, A. T. and R. Choudhary (2013). Decision making under uncertainty in the retrofit analysis of the uk housing stock: Implications for the green deal. *Energy and Buildings 64*, 292–308.

Booth, A. T., R. Choudhary, and D. J. Spiegelhalter (2012). Handling uncertainty in housing stock models. *Building and Environment 48*, 35–47.

Booth, A. T., R. Choudhary, and D. J. Spiegelhalter (2013). A hierarchical bayesian framework for calibrating micro-level models with macro-level data. *Journal of Building Performance Simulation 6*(4), 293–318.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis* (3 ed.). CRC Press.

Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing 26*(2), 448–466.

International Organization for Standardization (2008). *ISO 13790:2008 - Energy performance of buildings - Calculation of energy use for space heating and cooling* (2 ed.).

Jensen, J. M. and H. Lund (1995). *Design Reference Year, DRY - Et nyt dansk referenceår (in Danish)*. Technical University of Denmark. Announcement number 281.

Kavgic, M., A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, and M. Djurovic-Petrovic (2010). A review of bottom-up building stock models for energy consumption in the residential sector. *Building and Environment 45*, 1683–1697.

Kennedy, M. C. and A. O'Hagan (2001). Bayesian calibration of computer models. *Journal of Royal Statistical Society. Series B (Statistical Methodology) 63*(3), 425–464.

Kristensen, M. H. and S. Petersen (2016). Choosing the appropriate sensitivity analysis method for building energy model-based investigations. *Energy and Buildings 130*, 166–176.

Li, Q., G. Augenbroe, and J. Brown (2016). Assessment of linear emulators in lightweight bayesian calibration of dynamic building energy models for parameter estimation and performance prediction. *Energy and Buildings 124*, 194–202.

Loga, T., B. Stein, and N. Diefenbach (2016). TABULA building typologies in 20 european countries – making energy-related features of residential building stocks comparable. *Energy and Buildings 132*, 4–12.

Reinhart, C. F. and C. C. Davila (2016). Urban building energy modeling – a review of a nascent field. *Building and Environment 97*, 196–202.

Rijal, H. B., P. Tuohy, M. A. Humphreys, J. F. Nicol, A. Samuel, and J. Clarke (2007). Using results from field surveys to predict the effect of open windows on thermal comfort and energy use in buildings. *Energy and Buildings 39*(7), 823–836.

Sobol', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment 1*(4), 407–414.

Sokol, J., C. C. Davila, and C. F. Reinhart (2017). Validation of a bayesian-based method for defining residential archetypes in urban building energy models. *Energy and Buildings 134*, 11–24.

Wittchen, K. B. and J. Kragh (2012). *SBi 2012:01 Danish building typologies – Participation in the TABULA project*. Danish Building Research Institute, Aalborg University.