

# A data-driven framework for characterising building archetypes: A mixed effects modelling approach

**Palmer Real, Jaume; Møller, Jan Kloppenborg; Li, Rongling; Madsen, Henrik**

*Published in:*
Energy

*Link to article, DOI:*
10.1016/j.energy.2022.124278

*Publication date:*
2022

*Document Version*
Publisher's PDF, also known as Version of record

Energy 254 (2022) 124278

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

# A data-driven framework for characterising building archetypes: A mixed effects modelling approach

Jaume Palmer Real [a, *], Jan Kloppenborg Møller [a], Rongling Li [b], Henrik Madsen [a, c]

[a] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[b] Department of Civil Engineering, Technical University of Denmark, Denmark
[c] FME-ZEN, Norwegian University of Science and Technology, Norway

ARTICLE INFO

*Article history:*
Received 10 February 2022
Received in revised form
19 April 2022
Accepted 12 May 2022
Available online 20 May 2022

*Keywords:*
Building archetype
Thermal characterisation
Mixed-effects modelling
Data-driven modelling

ABSTRACT

Building archetypes are a common solution to study the energy demand of cities and districts. These are generally based on building information such as construction year and function. However, there can be large differences in the energy demand of buildings of the same archetype due to factors such as the preferences of occupants, quality of the building construction, and unrecorded renovations. This work uses a non-linear mixed effects model to capture these random differences. The model uses weather measurements to generate the daily heating load of buildings for the whole year. The model is generated and tested using data from 56 Norwegian apartments. Results show that 91% of measurements from an out-of-sample test set fall inside the 95% prediction interval. Additionally, the model allows us to compute a proxy of the heat loss coefficient, which characterises the heating performance of the population of apartments. Finally, two sub-categories of apartments are identified by clustering the model estimates for the studied population. The model is general, computationally light and uses existing data that are commonly collected in many buildings. The suggested method offers a more robust and reliable method to segment building archetypes using only weather data and energy demand.



## 1. Introduction

Around 55% of the world's population live in cities, and this number is expected to increase to 70% by 2050 [1]. This implies higher energy demand in concentrated areas, thus meticulous planning is necessary to guarantee sustainable growth. Moreover, a city is a complex system, and to increase its sustainability, it is necessary to understand its components and how they interact. Buildings are a key factor to manage when planning the energy use in cities; they represent 40% of the total energy demand in urban areas. Therefore, increasing buildings' efficiency will reduce total energy use considerably. In addition, buildings can be used to balance the energy supply grid by using strategies such as demand response facilitated by smart infrastructures [2]. When modelling districts and cities, the inclusion of individual buildings is challenging, as this increases the complexity of the model. A potential solution to this is using urban building energy models (UBEM), which aim to divide the building stock into categories (*segmentation*) and capture their attributes to simulate a typical consumption of each category of building (*characterisation*) [3].

In general, the categorisation is based on qualitative attributes of the buildings, such as the year of construction, location and the functionality. In each category, a building model is calibrated to simulate the energy consumption of the specific category [4]. Often, the building models used are based on previously identified archetypes, which are deterministic models that disregard the variability of the heat consumption inside each category. However, buildings that are grouped based on qualitative or quantitative characteristics, such as their usage or year of construction, present significantly different heat responses [5,6]. The causes behind these differences in energy use inside a building can be difficult to identify. They may include occupant behaviour, different geometry, or renovations that have not been declared [7,8]. The effects of occupants have been evaluated through measurements of CO2 concentration [9], tracking appliance usage [10] or survey activity diaries [11]. Assessing the effects of geometry and renovation requires data about the building construction that may not be up to date or even accessible. Thus, monitoring the possible sources of

* Corresponding author. Anker Engelunds vej 1, Building 101A, 2800, Kongens Lyngby, Denmark.
*E-mail address:* jpre@dtu.dk (J. Palmer Real).

https://doi.org/10.1016/j.energy.2022.124278

**Nomenclature**

*Acronyms*

| | |
|---|---|
| ES | Energy signature |
| EUI | Energy use index, kWh/m$^2$ |
| GMM | Gaussian mixture model |
| HLC | Heat loss coefficient, kW/$^o$C |
| ME | Mixed effects |
| UBEM | Urban building energy model |

*Indices*

| | |
|---|---|
| $i$ | Observation/measure |
| $k$ | Individual building |

*Variables*

| | |
|---|---|
| $\Phi$ | Heat load, kW/m$^2$ |
| $\Phi^{sol}$ | Solar irradiation, W/m$^2$ |
| $T^{out}$ | Outdoor temperature, $^o$C |
| $W^s$ | Wind speed, m/s |

uncertainty is not always an option.

This study proposes using random effects to model the energy consumption of a population of similar buildings. The focus is set to characterise the heating load and generate stochastic simulations; thus, we pursue a reduced-order model that can be easily interpreted. Instead of seeking a model structure that relies on extensive measured variables to incorporate the causes behind differences in building consumption, this study accepts those differences and aims to quantify their impact using limited data that is easy to access.

### 1.1. Review of hierarchical methods in building modelling

Random effects are used to model and quantify random differences between individuals in a population. Often, they are used to model the inner sample variations in clinical studies [12]. In buildings science, a similar example is found in the work by Rupp et al. [13], where they used random effects to model impact of the occupants on the heating of an office in a sub-tropical city in Brazil. In their work, they identified parameters that determine the level of comfort of the office workers, such as the habit of drinking hot beverages or wearing warm clothing. Since these attributes depend on individual preferences, random effects were added to the model to account for them.

Random effects were also explored by Capozzoli et al. [14] when they used a linear mixed effects model (LMEM) to simulate the annual energy consumption of healthcare buildings in northern Italy. The mixed effects structure allowed them to characterise a big ensemble of buildings with a single model and capture their common attributes, despite the buildings presenting qualitative differences. However, their model is only able to simulate annual values, and it is designed for coarse energy benchmarking. The model proposed in the present work uses daily values and takes into account the weather's influence, thereby offering a richer simulation of the heating demands of building categories.

Palmer et al. [15] presented a linear model that used random effects to simulate stochastic hourly profiles of building categories. The model was an extension of the work done by Lindberg, Bakker, and Sartori [16] that developed a linear fixed effect model to simulate the above-mentioned energy profiles. The model by Palmer et al. [15] was linear and the uncertainty was purely Gaussian, which limited the overall performance of the simulation tool.

Mixed effects models are often formulated as hierarchical models, where a random variable that describes a subset of a population is nested inside a broader and more general model. Cerezo et al. [17] compared methods to characterise building archetypes to simulate yearly consumption. Starting with a physics-based archetype model, they assigned a probability distribution to its most uncertain parameters; later, they calibrated these parameters using a Bayesian approach. Their work showed that a model based on this stochastic calibration provided more reliable simulation results than a purely deterministic method. The work by Cerezo et al. [17] was continued by Sokol, Davila, and Reinhart [18], using a similar approach to a different group of buildings in a different climate. In this way, they validated the concept of using a hierarchical structure to define building archetypes. Additionally, they repeated the experiment using yearly and monthly values of energy consumption and found that simulating monthly values and later aggregating provides more accurate estimation of the distribution of heating loads.

Kristensen, Hedegaard, and Petersen [19] suggested a hierarchical approach to model a population of Danish houses. In their work, they used a complex building model that returns the hourly energy use. Given the model complexity, it was only feasible to calibrate a subgroup of the model parameters for a segmented building population. De Jaeger, Lago, and Saelens [20] proposed a stochastic characterisation of the thermal performance of buildings by estimating the probability distribution of the U-values for different Flemish building categories. To perform such a study, they needed to have access to a region-wide energy certificate database.

Gholami et al. [21] used a Bayesian calibration method to tune 11 different building archetypes from a neighbourhood in the Italian city of Bologna. Their results showed robust long-term prediction with an improvement in the computational requirements. Still, their study focused on the annual energy use index (EUI) and relied on a model structure provided by the building modelling software Energy+.

A trend can be observed in the previous studies reviewed here: they focus on assigning a density distribution to a subset of parameters using established archetype building models, and these distributions are then calibrated, so the building models account for random differences within building categories. In the present study, the differences between buildings are captured by the random effects, as part of a reduced-order model. The model is based on a non-linear sigmoid-based energy signature that, as stated by Nageler et al. [22], gives reliable simulation results when heating measurements are available. The model uses daily energy consumption and weather data as input and can simulate the daily heating load continuously for the whole year. Since the uncertainty caused by buildings' differences is captured, it is separated from other sources of noise, which renders the model fit more reliable.

The proposed model highlights the potential of mixed effects (ME) models to study building populations. Instead of limiting the use of hierarchical models for calibration of known building models, ME models can be developed in more general model structures to satisfy specific goals; in this case, the simulation of daily heating load and the characterisation of the weather response of a category of buildings. The flexibility of this framework is a valuable asset to develop reliable and representative models for large districts and cities.

### 1.2. Paper outline

This paper is organised as follows. Section 2 introduces the modelling method where, first, the non-linear model is presented

for one single building; then, the model is extended with the addition of random effects to model a building population. To introduce the reader to the concept of random effects, Section 2.2.1 presents an example studying the annual energy demand of a population of schools. Section 3 presents the main findings of this work and is divided in two sub-sections: first, a fixed effects (FE) non-linear model is presented to study the heating load pattern of an individual building and a detailed description of the model is given to assess the quality of the fit. The FE model is extended with the addition of random effects, which converts it into a mixed effects (ME) model. Fig. 1 schematically presents the workflow of this section, where one can see that the outcome of the ME model is two-fold. On the one hand, the ME model allows a richer study of the known set of buildings used to train the model (*profiling*); on the other hand, the ME model is used to generate new observations of buildings (*sampling*), i.e., simulating unknown buildings. Section 4 uses the outcome of the profiling to refine the segmentation of the initial population of buildings. Finally, Section 5 discusses the major outcomes of this work.

## 2. Methods

This section introduces the model structure to model one single building. Later, the model is extended to account for the whole category.

### 2.1. One building: a continuous energy signature

The heating needs of buildings change throughout the year as they are affected by weather conditions, which as a first approximation can be represented by the outdoor temperature; thus, the outdoor temperature is considered the main driver of the heating demand. In absence of cooling, the heating load of a building presents the following trend: when the outdoor temperature is high enough and there are no other heating requirements, the heating load of a building is zero; as temperature decreases, the load starts to increase; eventually, in the coldest period of the year, the heating load curve will flatten out because the heating system reaches its capacity. Qualitatively, this heating behaviour matches a monotonic increase between two plateaus: one at zero heating demand and one at the maximum heating capacity. This S-shape behaviour can be modelled by a sigmoid curve, a model found in numerous and diverse systems. Such curves are continuous functions that are characterized using a small number of parameters.

A specific type of sigmoid is the Gompertz curve. This model has typically been used to model population growth in biological systems [23], although it has also been used to model wind power curves [24]. This model has the following structure

$$y = Ae^{-e^{-C(x-Q)}}, \tag{1}$$

which describes a continuous curve that starts at $y = 0$ (at $x = \infty$) and eventually reaches a horizontal asymptote. This curve is characterised by three parameters: $A > 0$ represents the upper asymptote (at $x = -\infty$), $C$ gives the growth rate from 0 to $A$, lastly, $Q$ acts as a horizontal offset of the curve. The function from Equation (1) can be seen in Fig. 2, where different combinations of its parameters have been plotted to visualise their effects on the shape of the curve.

In this study, Equation (1) is used to model the heating load curve of a building, where $y$ represents the heating load of a building, $\Phi$, and $x$ is the outdoor temperature, $T^{\text{out}}$. Then, Equation (1) becomes

$$\Phi_i = A\exp[-\exp(-C(T_i^{\text{out}} - Q))] + \epsilon_i, \tag{2}$$

where the notation is changed to improve the readability. In Equation (2), the sub-index $i$ represents the $i$th observation of the heating load and the outdoor temperature; in addition, a noise term is added, which is represented by the random variable $\epsilon_i \sim N(0, \sigma^2)$.

The parameters $\{A, C, Q\}$ describe how the daily heating demand changes with the outdoor air temperature; still, this temperature is coupled with other climatic variables that affect the heating load of a building. In this work, we modelled these effects by defining $\{A, C, Q\}$ as a function of the weather conditions. As introduced in Fig. 2, each of these three parameters defines a distinct attribute of the heating load curve, so each parameter has been handled separately as follows:

- The parameter $A$ represents the heating capacity, thus it is defined as constant regardless of the weather conditions.
- The parameter $C$ characterises the slope of the heating load curve. This dependence echoes the heat loss coefficient, which is often influenced by the wind conditions [25]. For this reason, the

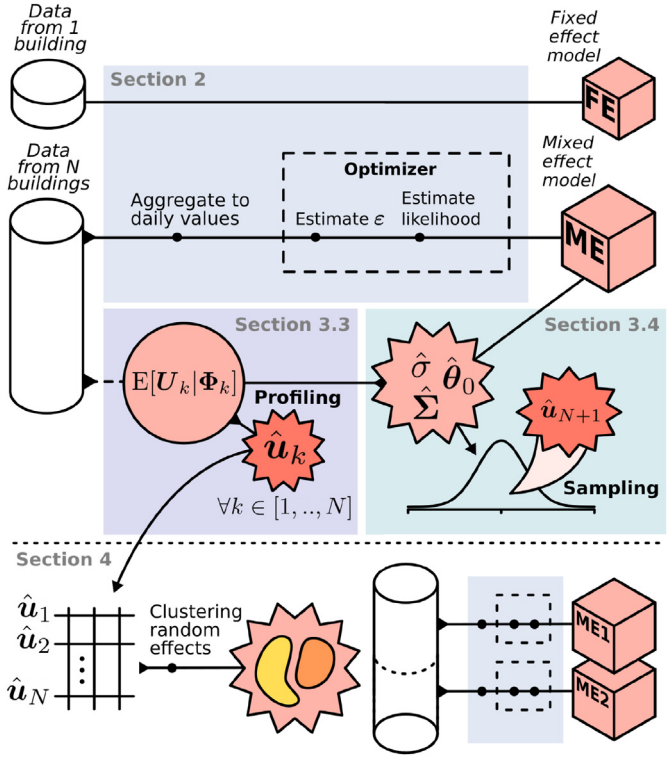


**Fig. 1.** Flowchart showing the main blocks of this work.

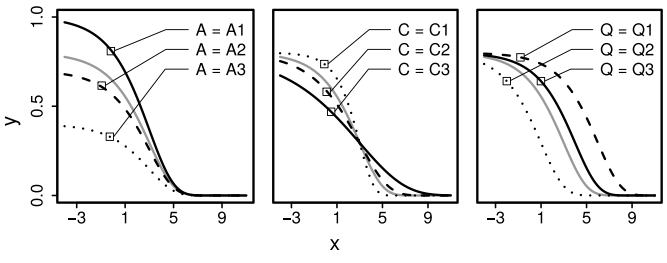


**Fig. 2.** Comparison of different Gompertz curves for arbitrary values of its defining parameters.

parameter $C$ is modelled as a linear function of the wind speed, $W^s$, such that

$$C \rightarrow C(W^s) = \alpha + \beta W^s. \tag{3}$$

- The parameter $Q$ horizontally shifts the heating load curve, which represents the passive heat gains and losses in the building. In particular, the effects of solar irradiation are a major contributor of the heat gains in highly insulated buildings [26]. For this reason, the parameter $Q$ is defined as a function of the solar irradiation, $Q(\Phi^{sol})$. This relationship might not have an explicit expression, since the effects of the solar irradiation depend on variables such as the incidence angle or the shading of nearby objects. Here, we use B-spline curves to capture the non-linear effects of the solar irradiation on the heating of a building. Thus, the parameter $Q(\Phi^{sol})$ is defined as

$$Q \rightarrow Q(\Phi^{sol}) = \sum_{i}^{n} b_i B(\Phi^{sol}), \tag{4}$$

where $n$ is the chosen number of spline curves, $B(\cdot)$ the basis spline function and $b_i \; \forall i \in [1, .., n]$ are scalar parameters representing the weight of each spline curve. The use of spline curves offers flexibility to model complex relationships between variables; more information on their application is found in the work of Rasmussen et al. [27].

The newly defined $\{A, C(W^s), Q(\Phi^{sol})\}$ introduced a lower level of parameters, namely $\boldsymbol{\theta} = \{A, \alpha, \beta, b_1, \ldots, b_n\}$ which represents the fixed effects parameter vector of the proposed model:

$$\Phi_i = A \exp\left[ -\exp\left[ (\alpha + \beta W_i^s)\left( T_i^{out} - \sum_{i}^{n} b_i B(\Phi_i^{sol}) \right) \right] \right] + \epsilon_i. \tag{5}$$

Notice that Equation (5) has the same non-linear structure as Equation (1). Thus, for the sake of clarity, the final model can be re-written as

$$\Phi_i = A \exp\Big[ -\exp\Big( -C(W_i^s)\Big(T_i^{out} - Q(\Phi_i^{sol})\Big)\Big)\Big] + \epsilon_i, \tag{6}$$

where the parameters $\{C(W^s), Q(\Phi^{sol})\}$ are given by Equations (3) and (4).

#### 2.1.1. Interpretability

The model introduced in Equation (6) captures the dependence of the heating load on the outdoor temperature, wind speed and solar irradiation. Often, the dependence between the heating load of a building and the weather conditions is modelled using a piece-wise differentiable model known as the *energy signature* (ES), which is a well known method to assess the thermal performance of buildings [28]. For buildings without cooling, such as the ones studied in this work, this model has the following expression

$$\Phi_i = \begin{cases} HLC \cdot (T_b - T_i^{out}) + \varepsilon_i & \text{if heating period} \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where $\{T_b, HLC\}$ are model parameters and $\varepsilon_i$ represents independent and identically normally distributed residuals. In Equation (7), the parameter $HLC$ stands for the *heat loss coefficient*, a performance indicator to evaluate the thermal insulation of the envelope of the building. Thus, the energy signature assumes a linear relationship between the heating load and the outdoor temperature in the heating regime. The change of regime in Equation (7) depends on the relationship between the outdoor air temperature and the parameter $T_b$: if $T^{out} < T_b$ the building requires heat to maintain comfortable indoor conditions; typically in Norwegian buildings $T_b \approx 17^oC$ [29]. Hence, the classic ES requires prior knowledge from the modeller to be able to separate the different heating regimes and adjust the fitting of the curve [30].

The model introduced in this work is continuous and presents a smooth transition between heating regimes. As shown in Equation (6), the model has the following structure

$$\Phi_i = f(T_i^{out}, \Phi_i^{sol}, W_i^s) + \epsilon_i, \tag{8}$$

where $f(\cdot)$ captures the weather dependence of the heating load. This makes it possible to define

$$g(T_i^{out}, \Phi_i^{sol}, W_i^s) = \frac{\partial f}{\partial T^{out}}, \tag{9}$$

a closed form continuous function that is completely described by the parameter $\{A, C(W^s), Q(\Phi^{sol})\}$.

The function in Equation (9) describes the change rate between the heating load and the outdoor temperature, and it can be used to compute a proxy of the classic $HLC$. Yet, $HLC$ is a constant parameter, whereas $g(\cdot)$ is a continuous function that is defined for the whole range of weather variables. As seen in Fig. 2, the Gompertz curve presents three different regions: two plateaus at the ends, and a slope in the middle. Near the inflection point, the middle region of the Gompertz function is well approximated by a linear model. Then, we define $HLC^* = g(T^{out^*})$, where $T^{out^*}$ is the outdoor temperature at the inflection point of $f(\cdot)$. Hence, physical information about the performance of the envelope of the building can be computed directly from Equation (6).

### 2.2. One category: randomness at the building level

This section introduces random effects to the model of Equation (6). To illustrate this concept, Section 2.2.1 presents an example where random effects are used to evaluate the differences in annual consumption of a population of schools. In the example, a simple model is developed to introduce the ME framework. The choice of schools is arbitrary, thus, this model could be applied to other building categories. If the reader is familiar with this type of modelling, they can skip this section and jump to Section 2.2.2, where the model from Equation (6) is extended using random effects.

#### 2.2.1. A mixed-effects example

The energy usage index (EUI) is a metric that summarises the annual energy consumption of a building per unit area. Buildings that have similar characteristics will have a similar EUI. If we are interested in estimating the mean value EUI of a population of similar schools, a model (M0) would be

$$EUI_i = \mu + \upsilon_i, \tag{10}$$

where $i$ denotes the $i$th observation of EUI and $\upsilon_i \sim N(0, \sigma_0^2)$ represents residual noise. Equation (10) contains only one fixed parameter, $\mu$, also known as a *fixed effect*. Notice that, since the model structure is chosen for its simplicity, it is assumed that the EUI does not depend on any other variable and it is distributed around the mean value, $\mu$.

However, in reality, the yearly heating consumption of different schools is different from the mean value. Then, for individual buildings, an alternative model (M1) would be

$$EUI_{i,k} = \mu + U_k + \varepsilon_{i,k}, \tag{11}$$

where, $\varepsilon_{i,k} \sim N(0, \sigma_1^2)$ and $\mu$ still represents the mean value EUI for the given population. The added term, $U_k$, captures the deviations of the individual $k$th building around the mean value. Notice that, since $U_k$ is added to Equation (11), the residual term $\varepsilon_{i,k}$ accounts only for the deviations of measurements of $EUI_{i,k}$ coming from the same individual $k$th building.

In order to characterise a population of buildings, we study how $U_k$ varies, rather than its individual values. Thus, it is modelled as a random variable $U_k \sim N(0, \sigma_u^2)$, and it is called the *random effect* of Equation (11). Notice that now, M1 is described by three parameters $\{\mu, \sigma_1^2, \sigma_u^2\}$. Since M1 contains both fixed and random effects, it is called a *mixed effects* model; for more details about this family of models please refer to Madsen and Thyregod [31].

Fig. 3 shows the results after fitting M0 and M1 using data from different Norwegian schools ($N = 21$). Sub-figure A) compares the estimated models over the distribution of the used data; sub-figure B) offers a visual comparison of the variances of the three random variables $v_i$, $\varepsilon_{i,k}$ and $U_k$. In order to ensure the data were normally distributed, the data were transformed using the Box-Cox transformation [32].

$$h(y_i) = \frac{y_i^{-0.5} - 1}{-0.5}. \tag{12}$$

Notice that the mean value of both M0 and M1 are the same, and the main difference is found in the variances $\sigma_0^2$ and $\sigma_1^2$. Since M1 includes the random effects, $U_k$, to account for the differences in individual $EUI_{i,k}$ the residuals $\varepsilon_{i,k}$ are smaller. This can be clearly observed in sub-figure B) where it is shown that in model M1, most of the noise is caused by $U_k$.

Fitting model M1 with data from 21 schools allows us to use Equation (11) to estimate the value of the individual values of $u_k$ $\forall\, k \in [1, \ldots, 21]$. As explained in Chapter 5 of Madsen and Thyregod [31], the random effects are estimated by

$$\hat{u}_k = E[U_k | EUI_i = \boldsymbol{y}_k] = \omega\hat{\mu} + (1-\omega)\bar{y}_k \tag{13}$$

where $\omega = 1/(1 + n\gamma)$, with $n$ being the number of observations of the $k$th building, and $\gamma = \sigma_u^2/\sigma_1^2$. It is important to highlight that $\hat{u}_k$ is an estimated value, whereas $U_k$ is a random variable. In addition, as introduced in Fig. 1, given the estimated parameters, it is possible to sample values from $N(0, \hat{\sigma}_u^2)$. Introducing these sampled values to Equation (11) makes it possible to simulate the energy usage index for similar schools that have no available energy data.

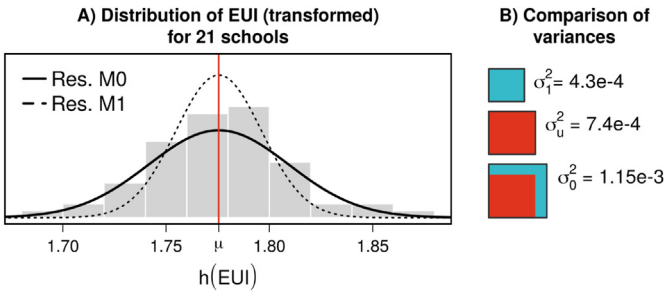


**Fig. 3.** Comparison of uncertainty for models M0 and M1. Figure A) shows the distribution of the residuals. Figure B) compares the different estimated variances, where the size of the rectangles is proportional to the variance.

### 2.2.2. Including the inner-category randomness in the heating load curve

The heating load curve of a single building can be modelled with Equation (6). This model is extended to represent an entire category of buildings by including random effects, defined in this work by the random variable $\boldsymbol{U} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. This constitutes a non-linear mixed effects model where on the one hand, the mean heating load curve of a category of buildings is characterised by $\boldsymbol{\theta}$ and on the other hand, the deviations from the mean curve of the $k$th building are described by $\boldsymbol{u}_k = \{u_{A,k}, u_{\alpha,k}, u_{\beta,k}, u_{Q,k}\}$, which is a sample from $\boldsymbol{U}$.

The final mixed effects model is

$$\Phi_{i,k} = (A_0 + u_{A,k})\exp\Big[-\exp\Big[(\alpha + u_{\alpha,k} + (\beta + u_{\beta,k})W_i^s) \times \Big(T_{i,k}^{\text{out}} - (1 - u_{Q,k})\sum_i^n b_i B(\Phi_{i,k}^{\text{sol}})\Big)\Big]\Big] + \epsilon_{i,k}, \tag{14}$$

where, $\boldsymbol{u}_k$ is included in Equation (6) under the following assumptions:

- The variable $u_{A,k}$ accounts for the differences in the upper limit of the heating load between individual buildings.
- The variables $u_{\alpha,k}$ and $u_{\beta,k}$ capture the small differences in the effects of the wind on an individual building.
- The influence of the solar irradiation is modelled as a sum of B-spline curves to adapt to possible non-linear effects. The small differences around the mean value of $Q(\Phi^{\text{sol}})$ are captured by adding $u_{Q,k}$.

Recalling the non-linear structure of the Gompertz curve, Equation (14) can be re-parameterised, defining the three high-level parameters as:

$$\begin{cases} A \rightarrow A_k = A_0 + u_{A,k} \\ C \rightarrow C_{i,k} = \alpha_0 + u_{\alpha,k} + (\beta_0 + u_{\beta,k})W_i^s \\ Q \rightarrow Q_{i,k} = (1 - u_{Q,k}) \cdot \sum_j^n b_j B(\Phi_i^{\text{sol}}) \end{cases} \tag{15}$$

Notice that, for the sake of clarity, a sub-index is added to the fixed effects so that $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}_0$. Adding the newly defined parameters in Equation (14) yields the final model,

$$\Phi_{i,k} = A_k \exp\Big[-\exp[C_{i,k}(T_i^{\text{out}} - Q_{i,k})]\Big] + \epsilon_{i,k}, \tag{16}$$
$$\underbrace{\qquad\qquad}_{f_{\text{ME}}(T_{i,k}^{\text{out}}, \Phi_i^{\text{sol}}, W_i^s)}$$

which characterises the daily heating load of a building category using only the continuous function $f_{\text{ME}}(\cdot)$.

### 2.3. Data description

All data used in this work comes from the TREASURE database supplied by SINTEF in the framework of the Flexbuild project. This database contains hourly measures of the heating load, outdoor temperature, wind speed and solar irradiation for nearly 300 Norwegian buildings. The consumption data was collected by a company that provides energy management services (EMS), and the measurements range from 2009 to 2018; the meteorological data was extracted from the Norwegian Meteorological Institute (MET). The data was cleaned so each measured building has continuous measurements that span from 1 year up to 3 years. The data quality is good, only the wind speed and solar irradiation presented minor

gaps in their measurements; the missing measurements were addressed using linear interpolation. Since this work only focuses on static characteristics, the data has been aggregated to daily values.

The data set also includes building information, for example, the geographic regions, building floor area, and functions. Additionally, the buildings are labelled according to their energy efficiency in one of the categories: E, T, R. Category E refers to buildings with efficiency near the *Passivhaus* standards; category T refers to buildings that have been recently renovated and are compliant with the Norwegian standards TEK10 [33]; finally, the buildings labelled R do not comply with any of the above two standards. Some of the buildings use district heating and others use electric heaters. Regardless of the heating sources, this study focus on the space heating load normalised using the area of each building with the unit of $kW/m^2$.

The proposed model is developed using a subset of the main database containing measurements from 56 apartment buildings. In order to validate the results, the data were split randomly into a training set, containing 41 buildings, and a testing set containing the remaining 15 buildings. The training set data are from 2013, 2017 or 2018, depending on the availability of the data for individual buildings; to ensure that the training set is balanced, only one year of data is chosen for each building. All measurements from the testing set are from 2017.

### 2.4. Parameter estimation

The parameters of the model in Equation (16) are estimated by maximizing the likelihood function,

$$L(\boldsymbol{y};\boldsymbol{\Theta}) = f_Y(y_1, y_2, \ldots, y_n;\boldsymbol{\Theta}), \tag{17}$$

where $n$ is the number of observations and $f_Y(\cdot)$ is the density function of the model. Thus, the likelihood function quantifies the probability of observing $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]$ given the parameters $\boldsymbol{\Theta}$.

In this work, $\boldsymbol{y}$ contains observations of the daily averaged heating load for 41 different buildings, so Equation (17) becomes

$$L(\boldsymbol{y};\boldsymbol{\Theta}) = \prod_{i=1}^{41} \cdot f_Y(\boldsymbol{y}_i;\boldsymbol{\Theta}), \tag{18}$$

where $\boldsymbol{y}_i$ contains the observations of the $i$th building. For models that contain only fixed effects, $\boldsymbol{\Theta}$ includes the fixed effects and the parameters that characterise the chosen model distribution; typically, when the model is assumed to be Gaussian, such distribution is completely characterised by the variance of the residuals. The inclusion of random effects adds a new level of uncertainty which requires re-writing the density function such that

$$f_Y(\boldsymbol{y}_i, \boldsymbol{u}_k;\boldsymbol{\Theta}) = f_{Y|\boldsymbol{u}_k}(\boldsymbol{y}_i|\boldsymbol{u}_k;\boldsymbol{\Theta}_1) f_U(\boldsymbol{u}_k;\boldsymbol{\Theta}_2), \tag{19}$$

where $\boldsymbol{\Theta} \equiv \boldsymbol{\Theta}_1 \cup \boldsymbol{\Theta}_2$ and $\boldsymbol{u}_k$ is a vector of length $q$ containing the random effects of the $k$th building.

The formulation presented in Equation (19) is called *hierarchical likelihood*, which splits the likelihood function in two terms, $f_U(\cdot)$ and $f_{Y|\boldsymbol{u}_k}(\cdot)$, highlighting the hierarchical structure of a mixed effects model. As depicted in Fig. 4, there is an underlying random variable, $\boldsymbol{U}$, that follows the distribution $f_U(\cdot)$; an observation of such random variable, $\boldsymbol{u}_k$, conditions the upper layer of the model that follows the distribution $f_{Y|\boldsymbol{u}_k}(\cdot)$. More details about the hierarchical function can be found in Chapter 5 of Madsen and Thyregod [31].

As presented in Section 2.2.2, in this work both the individual noise and the population noise are normally distributed, with

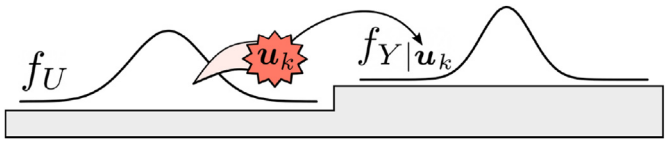


**Fig. 4.** Schematic representation of the hierarchical likelihood structure from Equation (19).

$$f_{Y|\boldsymbol{u}_k}(\boldsymbol{y}_k;\boldsymbol{u}_k,\boldsymbol{\theta}_0,\sigma_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}^n} \exp\left[ -\frac{1}{2} \frac{\sum_{i=1}^{n} \left( y_{i,k} - f_{\mathrm{ME}}(\boldsymbol{u}_k, T_{i,k}^{\mathrm{out}}, \Phi_{i,k}^{\mathrm{sol}}, W_{i,k}^{s}) \right)^2}{\sigma_1^2} \right] \tag{20}$$

and

$$f_U(\boldsymbol{u}_k;\boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi}^q \det \boldsymbol{\Sigma}} \exp\left[ -\frac{1}{2} \boldsymbol{u}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{u}_k \right]. \tag{21}$$

where $\boldsymbol{\Theta_1} \equiv \{\boldsymbol{\theta}_0, \sigma_1, \boldsymbol{u}_k\}$ since Equation (20) contains the function $f_{\mathrm{ME}}(\cdot)$, and $\boldsymbol{\Theta_2} \equiv \{\boldsymbol{\Sigma}\}$.

The final expression of the likelihood function is obtained by combining Equations (18) and (20)-(21), which leaves a complicated non-linear product. To reduce the computational requirements, the negative logarithm of the likelihood function is computed, so the product from Equation (18) becomes a summation, and the maximisation becomes a minimisation. The final objective function is

$$l(\boldsymbol{y}, \boldsymbol{u}_k; \boldsymbol{\theta}_0, \sigma_1, \boldsymbol{\Sigma}) = -\log\left[ \prod_{i=1}^{41} f_{Y|u_k}(\boldsymbol{y};\boldsymbol{\theta}_0,\sigma_1) f_U(\boldsymbol{u}_k;\boldsymbol{\Sigma}) \right], \tag{22}$$

which is still a complex non-linear function to minimise. To ease this optimisation, in this work the TMB package is used [34]. This package runs in R and uses the Laplace approximation to calculate the function $l(\cdot)$ and estimates its parameters.

### 2.5. Simulation and validation framework

The model proposed in this work is intended for simulation purposes. Recalling Fig. 1, the simulation is done by sampling realisations from the random effects distribution. Then, using outdoor air temperature, wind speed and solar irradiation measurements for a given period, it is possible to simulate the daily heating load of a building.

To study the uncertainty introduced by the differences among individual buildings, the region that includes the 95% of sampled buildings is computed. This uncertainty region is computed using a Monte Carlo approach, where we simulate numerous realisations and select the region containing 95% of the sampled space. The performance of the proposed model is evaluated using the *reliability* metric, which is computed by measuring the % of test measurements that fall inside the 95% uncertainty region.

## 3. Results

This section introduces the results of fitting the model in Equation (6) using data from one single building. Then, the results of modelling a whole category of buildings with Equation (16) are presented. The results at category level are divided into two subsections: first, the model fit is analysed studying the model

results compared to the training set; later the model is validated comparing the results with a set of measurements from out-of-sample buildings. Finally, the proposed model is used to study the thermal performance of a category of buildings.

### 3.1. One apartment

Fig. 5 shows the fit of the training set for one apartment building. It can be seen that the model accurately captures the fluctuations of the daily heating load and presents a small residual noise.

Evaluating the residuals of such a model confirms the quality of the fit. Fig. 6 shows the residuals are centred around zero and present no significant trends when compared with the three weather variables. However, it is noticeable that the variance decreases slightly for the outdoor temperature at the warmer side of the plot. This shrinking of the variance is caused by the change of heating behaviour during summer days: when the outdoor temperature is very high the heating load is very close to zero. The residuals term from Equation (6) has a constant variance that is unable to adapt to such behavioural changes. Moreover, it can be noticed that the residuals are slightly narrower for high values of the solar irradiation; this is caused by the coupling between bright days and high outdoor air temperature. In addition, given that the data comes from Norway, it is seen that most of the daily data points concentrate in the lower end of the solar irradiation, so it is more likely to find larger residuals in that region. Similarly, focusing on the wind speed, extreme values present small residuals, but, given the low number of observations in this range, this is not considered as a potential bias; as seen in Fig. 7, changes in the wind speed will have a limited impact over the heating load.

Additionally, Fig. 6 shows an exponential decay of the auto-correlation function (ACF) and a significant partial auto-correlation (PACF) in lag one. This indicates that the model omits some structure (presumably an AR(1)); yet, given the lag-one auto-correlation is only around 0.5, this is ignored. It is important to recall that the proposed model is static and does not take into account the effects of heating inertia, which might be what causes this lag-one auto-correlation. Still, the results of the model are satisfactory for the purpose of this study.

The relation between the estimated heating load and the outdoor temperature is shown in Fig. 7. There, three different curves are plotted for different weather conditions to evaluate how the model adapts to changes in the weather variables. The figure includes two sub-plots that show the dependence of parameters $\{C(W^s, Q(\Phi^{sol}))\}$ on the weather variables.

The $C$ parameter grows slightly with the wind speed, confirming that the heat loss increases as wind speed increases; however, the small slope suggests that changes in the wind speed will not cause significant changes in the heating load. The parameter $Q$ follows a sharply decreasing trend that flattens, then slightly increases during days with high solar irradiation. The initial decrease suggests that, as the solar irradiation increases (see trend from point I to point II), the heating load curve is shifted to the left (see solid and dashed curve in Fig. 7 left), since the building demands less heating due to the solar gain. The increase of $Q$ (see trend from point II to point III), suggests that, for very bright days, the internal gains decrease slightly. This is however a small and possibly not significant effect. This could be caused by some behavioural changes of the occupants during very bright days, such as an increase in the heat loss caused by window opening.

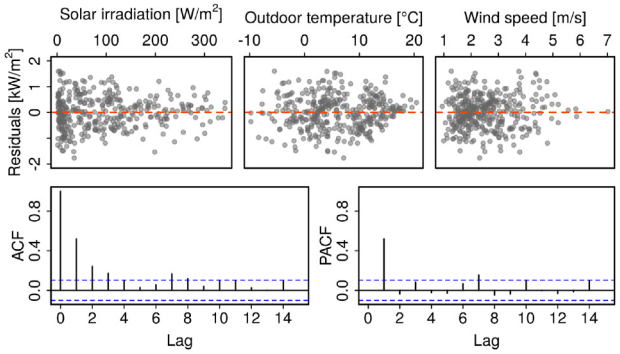


**Fig. 6.** Residuals analysis of an arbitrary apartment building. The top row of plots show the dependence of the residuals with two main inputs. The bottom row shows the ACF and PACF.

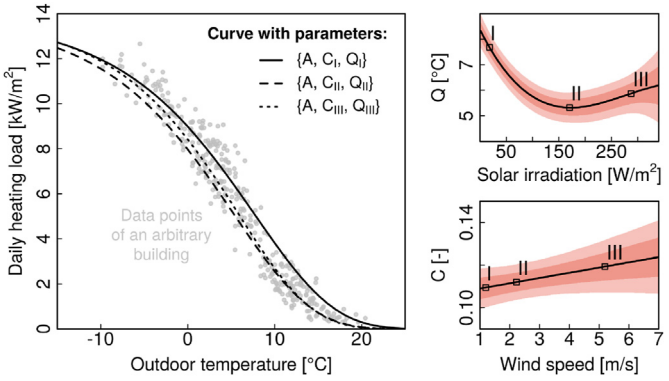


**Fig. 7.** Main figure shows the dependence of $f_{ME}(\cdot)$ on the three weather variables. Two sub-figures show the dependence of parameters $C$ and $Q$ on wind speed and solar irradiation, respectively.

### 3.2. Modelling results of all apartments

To model the apartment category, random effects are introduced as described by Equation (16). After computing the Akaike information criterion (AIC) for models with different numbers of splines, the final choice was to use four spline curves. Using four spline curves yields the final model with seven fixed effects, $\hat{\theta}_0$, which can be seen in Table 1. The p-values confirm that all parameters are significant, proving the dependence of the model on the three

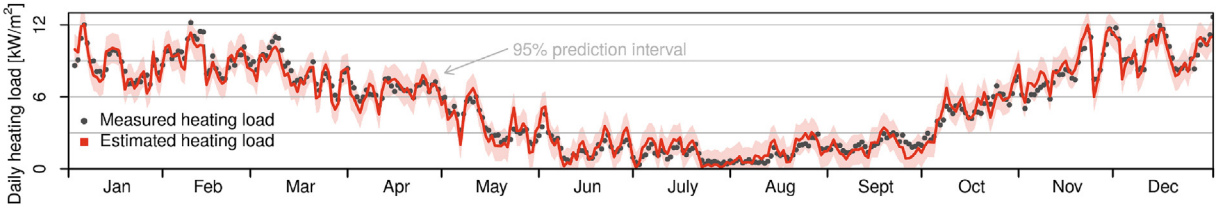


**Fig. 5.** Yearly evolution of the simulated heating load for one building compared to measurements of the heating load during that period.

**Table 1**
Fixed effects estimates (left). Diagonal of the covariance matrix of random effects, $\boldsymbol{\Sigma}$ (right).

| **Fixed effects** | | | | **Random effects** | |
|---|---|---|---|---|---|
| | Mean | Std. dev. | p-value | | Std. dev. |
| $\hat{A}_0$ | 23.60 | 1.08 | < 1e-16 | $\hat{u}_A$ | 1.64 |
| $\hat{\alpha}_0$ | 8.44e-2 | 2.8e-3 | < 1e-16 | $\hat{u}_\alpha$ | 1.59e-2 |
| $\hat{\beta}_0$ | 8.82e-4 | 5.2e-4 | 8.75e-2 | $\hat{u}_\beta$ | 2.10e-3 |
| – | – | – | – | $\hat{u}_Q$ | 2.89e-1 |
| $\hat{b}_1$ | 5.91 | 0.30 | < 1e-16 | – | – |
| $\hat{b}_2$ | 0.29 | 0.16 | 7.0e-2 | – | – |
| $\hat{b}_3$ | 1.52 | 0.23 | 8.53e-11 | – | – |
| $\hat{b}_4$ | 3.10 | 0.27 | < 1e-16 | – | – |

weather variables. Yet, $\beta_0$ presents a small value with the largest uncertainty, suggesting the effects of the wind speed are weak.

The random effects of the model are contained in a four-dimensional random vector, $\boldsymbol{u}_k$, that is distributed following a multivariate normal $\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. The right side of Table 1 includes the standard deviation of each component of $\boldsymbol{u}_k$. It can be observed that $\hat{u}_\alpha$ $\hat{u}_\beta$ present a large standard deviation, highlighting that the wind dependence varies significantly among the training buildings.

As explained in Section 2.2.1, using the measurements in the training set, it is possible to estimate the individual values of $\boldsymbol{u}_k$ for all the observed buildings. Studying the quantiles of $\boldsymbol{u}_k$ $\forall k \in [1, \ldots, 41]$ confirms the estimated random effects follow a Gaussian distribution, as can be seen in Fig. 8.

The fixed effects shown in Table 1 give the mean value of the heating load curve which can be seen in Fig. 9. As the parameters describing the curve depend on the solar irradiation and wind speed, to improve visualisation, three snapshots are plotted. Each sub-plot contains the observations for a range of solar irradiation, and the mean curve uses the median solar irradiation and the median wind speed of that range. There is still a large uncertainty around the mean, which is caused by the differences among the buildings. This is confirmed in Fig. 9 by including the individual building curves from the buildings in the training set.

Thus, the random differences from building to building add a new layer of uncertainty that is captured by the random effects, hence, it is separated from residual noise. To visualise this result, two models were compared: the proposed model with fixed effects and random effects, denoted by $f_{\mathrm{ME}}(\cdot)$; and another model with only fixed effects, that is, setting $\boldsymbol{U} = \boldsymbol{0}$, denoted by $f_0(\cdot)$. The comparison of $f_{\mathrm{ME}}(\cdot)$ and $f_0(\cdot)$ can be seen in Fig. 10, where the regions defined by the 95% prediction interval are highlighted for both models.

In the model without random effects, the uncertainty is constant around the mean, since the only source of uncertainty comes from the normally distributed residuals. When using the random effects, the uncertainty changes with the outdoor temperature due to the non-linear structure of the model. In the cold end of the temperature spectrum, the uncertainty is significantly wider when compared to the model without random effects, that is, during days with low outdoor temperatures the heating load differs more from one building to the other. When the outdoor temperature is high, the uncertainty is lower since all buildings have lower heating loads. The uncertainty region of $f_{\mathrm{ME}}$ includes more data points, which renders that model as a better representative of the studied population.

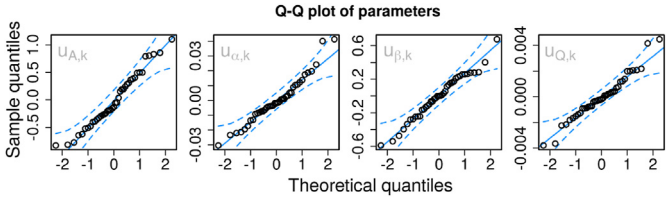


**Fig. 8.** Quantile plots of the estimated random effects. All of the quantile points fall inside the 95% prediction interval confirming that the random effects follow a normal distribution.

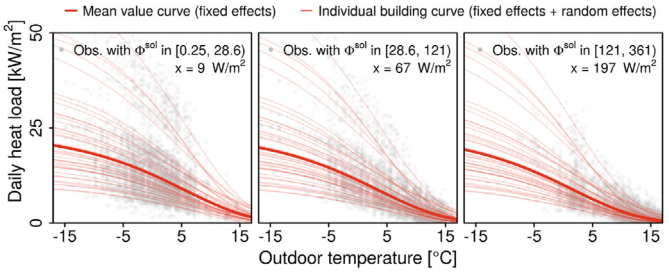


**Fig. 9.** Three snapshots comparing the mean curve and the individual building curves for different values of the solar irradiation.

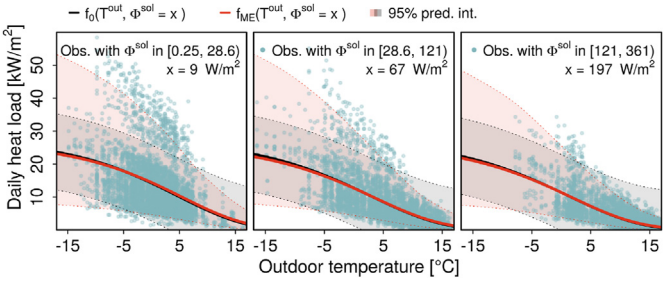


**Fig. 10.** Comparison of the uncertainty region of the ME model and a model without random effects.

### 3.3. Predicting observed buildings

The estimated fixed effects, $\hat{\boldsymbol{\theta}}_0$, and random effects, $\hat{\boldsymbol{u}}_k$, completely characterise the buildings in the training set. Then, their only source of uncertainty is the residual noise, $\epsilon_{i,k}$. Fig. 11 shows the estimated individual curves of 6 buildings; to improve readability, these curves are computed with a fixed solar irradiation. Notice that the 95% prediction interval is much narrower, when compared to Fig. 10, due to the absence of the building uncertainty. Still, most of the measured data points are included in the uncertainty region, highlighting the good fit of the model at the building level.

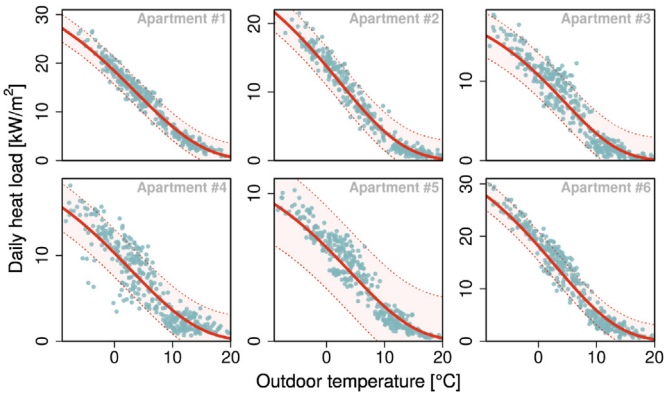


**Fig. 11.** Heating load curves of six different buildings. The solar irradiation and the wind speed are set as constants to ease readability.

### 3.4. Simulating unobserved buildings

If the 41 buildings of the training set are a representative sample of the apartment population in Norway, then the uncertainty region from Fig. 10 marks the region where 95% of observations from any unobserved Norwegian apartment buildings will fall. Although it is not possible for the authors to examine the representativeness of these buildings, we can compare the uncertainty region to measurements from 15 apartment buildings that were not part of the training set. The daily heating load measurements of those test buildings compared to the uncertainty area of $f_{ME}$ depicted in Fig. 10. Given this region, it is possible to compute the *reliability* measure, defined as the percentage of data points that fall inside the 95% prediction interval. Fig. 12 shows the distribution of the individual scores, where one will note that most of the test buildings show a reliability over 90%.

### 3.5. Assessing the thermal performance of the studied category

The estimated parameters of Equation (16) were used to compute the proxy heat loss coefficient, *HLC** described in Section 2.1.1. Given the presence of random effects, *HLC** is not described by a single parameter but a distribution that characterises the studied category of buildings. The density distribution of *HLC** was estimated using a Monte Carlo approach and is shown in Fig. 13. It presents a wide bell curve with a long tail in the higher side of the *HLC* range.

For each building, in both the training and testing set, the classic heat loss coefficient are computed following Equation (7). The distribution of these individual values is included in Fig. 13 as a dot plot. Despite the low number of points, the computed values resemble the distribution, indicating that *HLC** is a valid alternative for simulating the thermal performance of the envelope of an unobserved apartment building.

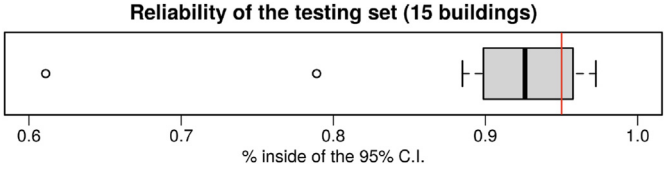


**Fig. 12.** Boxplot that shows the distribution of reliability measure for the 15 buildings in the testing set.

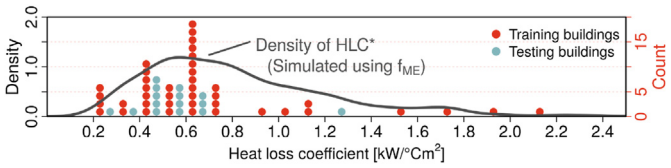


**Fig. 13.** Distribution of the *HLC** computed using model $f_{ME}$. The distribution is compared to a dot plot of the classic *HLC* computed for the buildings in testing and training sets.

## 4. Model application: refining segmentation

As described in Section 3, with the proposed model it is possible to estimate the random effects, $\hat{\boldsymbol{u}}_k$, for each building in the training set. This vector, along with $\boldsymbol{\theta}_0$, completely characterises a building's response to the weather conditions, providing a deeper understanding of the set of buildings used to train the model. In this work, we propose using $\hat{\boldsymbol{u}}_k$ to group buildings based on the likeness of their heating curve. This approach offers a data-driven alternative to the conventional segmentation procedure that groups buildings using qualitative data that might be outdated or missing.

The buildings are grouped using a fuzzy analysis clustering (FANNY method) [35] and two sub-categories of buildings are found, named C1 and C2. Table 2 presents the available details of the buildings contained in each cluster. Notice that, using only qualitative details, no clear line can be drawn to separate C1 and C2, since both include buildings with similar attributes. Additionally, both clusters have buildings with missing details.

As outlined in Fig. 1, these two clusters are studied separately by splitting the original training data set and repeating the fitting process described in Section 3. This results in having two different models: ME1 and ME2. Both models follow the structure presented in Equation (16) and are governed by $f_{\text{ME1}}$ and $f_{\text{ME2}}$ respectively. Hence, ME1 and ME2 are described by a different set of parameters $\boldsymbol{\theta}_0^{(l)}$, $\Sigma^{(l)}$ and $\sigma^{(l)}$ for $l \in [1,2]$. Table 3 compares the fixed effects after fitting both models. It is seen that ME2 presents a higher heating capacity and higher wind dependence than ME1, suggesting that C1 contains buildings that are better insulated. This was already hinted at in Table 2, where we noted that C1 contained buildings compliant with high efficiency standards. Furthermore, notice that the solar gains, reflected by the sum of splines, are higher for buildings in C1, which indicates again that the impact of solar irradiation is more significant in low-energy buildings.

To assess how this finer segmentation represents the population of apartments, the models ME1 and ME2 are compared to the test set presented in Section 3.4. Since it is unknown which model represents each test building, the reliability of both models is computed for each of the 15 buildings. When the measurements of

**Table 2**
Summary of characteristics found in the different buildings contained in C1 and C2.

| Function | C1 (24 buildings) | C2 (17 buildings) |
|---|---|---|
| | Apartment block | |
| Construction years | 2010, 2012–2016 | Unknown, 1998, 2005–2006, 2011, 2013, 2015–2016 |
| # of units | 4–154 | 9–62 |
| Location | Unknown, Harstad, Jakobsli, Ranheim, Trondheim | Harstad, Heimdal, Moss, Ranheim, Trondheim |
| Area [m²] | 352–17457 | 640–5775 |
| Eff. label | R, T, E | R |

**Table 3**
Fixed effects for the two clusters. Notice that the term $\sum_i^{n=4} b_i B(\Phi^{sol})$ is defined for a range of solar irradiation; to improve interpretability, only the values for $x = 50W/m^2$ are given.

| | Train. | $A^{(l)}$ | $\alpha^{(l)}$ | $\beta^{(l)}$ | $\sum_i^{n=4} b_i^{(l)} B(x)$ |
|---|---|---|---|---|---|
| ME1 | C1 | 2.80 | 9e-2 | 1.2e-4 | 4.03 |
| ME2 | C2 | 3.66 | 7.72e-2 | 2.02e-3 | 3.72 |

**Table 4**
Comparison of NMBE and CVRMSE metrics of the whole data set, C1 and C2.

| | | NMBE [%] | CVRMSE [%] |
|---|---|---|---|
| ASHRAE 14-2014 | | $\leq \pm 5$ | $\leq 15$ |
| **Model ↓** | **Test set ↓** | | |
| ME1 | 13 buildings (C1) | 1.12 | 6.29 |
| ME2 | 2 buildings (C2) | −26.30 | 50.75 |

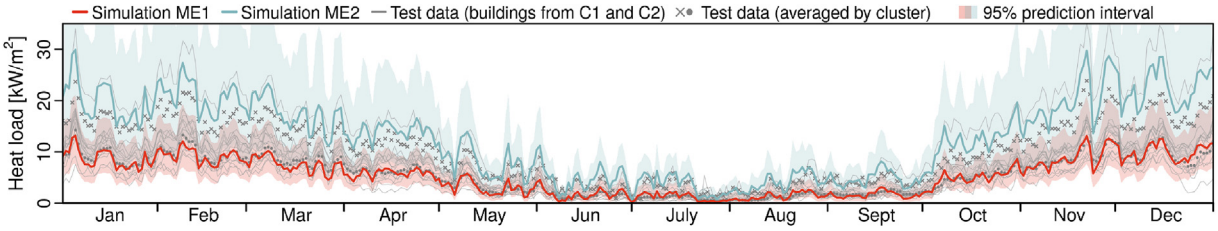


**Fig. 14.** Simulated yearly evolution of the heating load for the two clusters compared to the daily mean of the 15 test buildings.

a test building show high reliability with model ME1, it is assumed that the building belongs to C1 and vice-versa. The results show that 13 of the 15 test buildings are better described by ME1. Hence, the test set is split in two: one subset contains 13 buildings to test ME1 and the other subset contains only two buildings to test ME2.

Fig. 14 shows one year of simulated heating load using ME1 and ME2. It is seen that the heating load simulated using ME2 is always higher that the trend simulated using ME1 and presents more pronounced fluctuations; similarly, the prediction interval of ME1 is wider. The figure includes measured data from the individual test buildings, as well as their daily averages by cluster. The averaged test data of buildings from C1 (marked with •) falls very close to the simulated mean with ME1, suggesting the model is a good representative of the sub-category. On the other hand, the average of C2 test buildings (marked with × ) falls lower than the ME2 simulation; however, it is important to recall that this average is computed using only two test buildings, which does not allow us to have an accurate heating trend for testing purposes.

Finally, the accuracy of each fitted model is quantified following ASHRAE guidelines 14−2014 [36]. These guidelines propose the NMBE and CVRMSE metrics to evaluate the model performance, and give boundaries for each metric to guarantee a good fit. For each sub-category, the monthly average of its test buildings is computed and compared to the simulated monthly heating load. The results can be seen in Table 4, which confirms model ME1 as a good representative of C1. The results for ME2 are significantly worse due to the low number of test buildings.

## 5. Conclusion

This work presents a methodology to model the heating load of archetypes of buildings using existing weather and energy meter data. The results from this model can be used to refine segmentation of a population of buildings. First of all, a non-linear model was introduced which captures the weather dependency of buildings. Using this model, reliable results were presented simulating the daily heating load of a single building during the period of one year. The resulting simulation is continuous and adapts to the typical heating regime change from heating season to non-heating season.

To model an entire category of buildings, the model was extended with the addition of random effects; in this work, a population of apartments was modelled. Results showed that the simulated heating load accurately follows the measured trend of the buildings in the training set. The non-linear model structure was able to adapt to the regime changes of the heating load during the year, which cause high variance during colder periods compared to warmer ones. Thus, the model shows a high uncertainty region during winter months which narrows as the heating load approaches zero during the summer. This uncertainty is caused by the random differences between apartments, and quantifying it allowed us to compute the region where 95% of measurements of the heating load will fall. This region was validated using measurements from 15 out-of-sample buildings, capturing 91% of these test data. Thus, given the weather conditions of an arbitrary period, the simulation using ME provides a reliable estimation of the range where the heating load of any apartment might fall during such a period.

The model is based on known physical phenomena and is easy to interpret. The estimated parameters give direct insights into the effects of outdoor temperature, wind speed and solar dependence. In addition, a proxy of the heat loss coefficient can be computed through these parameters. The stochastic nature of the proposed model allowed us to estimate how this thermal performance is distributed for the studied category.

One of the major challenges of working with models that aim to represent urban areas is finding a way to accurately segment the building stock. Working with a large enough data set, the proposed model allows identifying sub-categories based on the estimated random effects, which offers a richer description of the building landscape. This method is purely data driven and does not require having access to qualitative data of the building (such as the geometry or year of construction) to segment a population of buildings. The model completely characterises the thermal response for a climatic year, and the buildings are directly grouped based on estimated random effects. In addition, if partial qualitative information about the studied set of buildings is available, this method can be used to fill the gaps. In the presented work, it was possible to classify buildings lacking key information, such as their construction year or clear efficiency labelling. The results indicate that this method can also be used to complement and validate segmentation that uses the classic archetype approach.

When working with mixed effects (ME) models, specially with non-linear ME, it is recommended to use a low number of random effects to ensure the computational feasibility. In this work, it was possible to add a random effect for each major parameter since the proposed model has a low order structure. In case of using a more complex model, it will be necessary to evaluate which parameter would be more affected by random effects. Issues can also arise

when dealing with long high-frequency data sets; for instance, in the proposed model, the trials to address the significant auto-correlation in the residuals were unsuccessful due to computational limitations.

Segmenting a population of buildings based on data-driven methods requires to be able to interpret the cause behind the newly found categories. In the studied case, the model was interpretable and the differences between sub-categories were easy to identify, which allowed to recognise a sub-category containing mostly low-energy buildings. However, the other sub-category could not be properly validated since its test set was too small and proved not to be representative. This arises the question of how to categorise the buildings that are not part of the training set.

Nevertheless, the results of this work are satisfactory and suggest that mixed effects are an effective modelling framework to develop urban models and adapt to the modelling needs. In this case, pursuing generality, the model is relatively simple as it uses daily values. Still, the results indicate that a mixed effects approach can be applied to more complex applications.

## Author statement

**Jaume Palmer Real:** Conceptualization, Methodology, Software, Visualization, Writing **Jan Kloppenborg Møller:** Methodology, Validation, Reviewing and Supervision **Rongling Li:** Writing, Reviewing and Supervision **Henrik Madsen:** Supervision, Reviewing and Validation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was done as a part of the Flexbuild project, which provided a clean and complete data set to work with. Furthermore, we wish to acknowledge the FME-ZEN project (Research Council of Norway - Project No. 257660) and SEM4Cities (Innovation Fund Denmark - Project No. 0143-0004) for inspiring and supporting this work.

## References

[1] Urban Development. Last Updated, https://www.worldbank.org/en/topic/urbandevelopment/overview. [Accessed 20 April 2020].
[2] Chen Yongbao, Xu Peng, Gu Jiefan, Schmidt Ferdinand, Li Weilin. Measures to improve energy demand flexibility in buildings for demand response (DR): a review. In: Energy and buildings, vol. 177; 2018. p. 125–39.
[3] Ali Usman, Haris Shamsi Mohammad, Hoare Cathal, Mangina Eleni, O'Donnell James. Review of urban building energy modeling (UBEM) approaches, methods and tools using qualitative and quantitative analysis. In: Energy and buildings, vol. 246; 2021. p. 111073.
[4] Ali Usman, Haris Shamsi Mohammad, Hoare Cathal, Mangina Eleni, O'Donnell James. A data-driven approach for multi-scale building archetypes development. In: Energy and buildings, vol. 202; 2019. p. 109364.
[5] Palmer Jaume, Rasmussen Christoffer, Li Rongling, Leerbeck Kenneth, Michael Jensen Ole, Wittchen Kim B, Madsen Henrik. Characterisation of thermal energy dynamics of residential buildings with scarce data". English. In: Energy and buildings, vol. 230; 2021. https://doi.org/10.1016/j.enbuild.2020.110530. issn: 0378-7788.
[6] Aksoezen Mehmet, Daniel Magdalena, Hassler Uta, Kohler Niklaus. Building age as an indicator for energy consumption. In: Energy and buildings, vol. 87; 2015. p. 74–86.
[7] Gabriel Happle, Fonseca Jimeno A, Schlueter Arno. A review on occupant behavior in urban building energy models. In: Energy and buildings, vol. 174; 2018. p. 276–92.
[8] Jong-Hwan Ko, Kong Dong-Seok, Huh Jung-Ho. Baseline building energy modeling of cluster inverse model by using daily energy consumption in office buildings. In: Energy and buildings, vol. 140; 2017. p. 317–23.
[9] Wolf Sebastian, Møller Jan Kloppenborg, Alexander Bitsch Magnus, Krogstie John, Madsen Henrik. A Markov-switching model for building occupant activity estimation. In: Energy and buildings, vol. 183; 2019. p. 672–83.
[10] Hamed Nabizadeh Rafsanjani, Ahn Changbum R, Chen Jiayu. Linking building energy consumption with occupants' energy- consuming behaviors in commercial buildings: non-intrusive occupant load monitoring (NIOLM). In: Energy and buildings, vol. 172; 2018. p. 317–27.
[11] Flett Graeme, Kelly Nick. A disaggregated, probabilistic, high resolution method for assessment of domestic occupancy and electrical demand. In: Energy and buildings, vol. 140; 2017. p. 171–87.
[12] DerSimonian Rebecca, Kacker Raghu. Random-effects model for meta-analysis of clinical trials: an update. In: Contemporary clinical trials, 28.2; 2007. p. 105–14. https://doi.org/10.1016/j.cct.2006.04.004. issn: 1551-7144, https://www.sciencedirect.com/science/article/pii/S1551714406000486.
[13] Rupp Ricardo Forgiarini, Andersen Rune Korsholm, Toftum Jørn, Ghisi Enedir. Occupant behaviour in mixed-mode office buildings in a subtropical climate: beyond typical models of adaptive actions. In: Building and environment, vol. 190; 2021. p. 107541.
[14] Capozzoli Alfonso, Piscitelli Marco Savino, Neri Francesco, Grassi Daniele, Serale Gianluca. A novel methodology for energy performance benchmarking of buildings by means of Linear Mixed Effect Model: the case of space and DHW heating of out-patient Healthcare Centres. In: Applied energy, vol. 171; 2016. p. 592–607.
[15] Palmer Jaume, Møller Jan Kloppenborg, Rasmussen Christoffer, , Karen Byskov Lindberg, Sartori Igor, Madsen Henrik. Simulating heat load profiles in buildings using mixed effects models. In: Journal of physics: conference series. (Lyngby, Denmark). Web of Science; 2021.
[16] Karen Byskov Lindberg, Bakker Steffen J, Sartori Igor. Modelling electric and heat load profiles of non-residential buildings for use in long-term aggregate load forecasts. In: Utilities policy, vol. 58; 2019. p. 63–88.
[17] Cerezo Carlos, Sokol Julia, AlKhaled Saud, Reinhart Christoph, Al-Mumin Adil, Ali Hajiah. Comparison of four building archetype characterization methods in urban building energy modeling (UBEM): a residential case study in Kuwait City. In: Energy and buildings, vol. 154; 2017. p. 321–34.
[18] Sokol Julia, , Carlos Cerezo Davila, Reinhart Christoph F. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. In: Energy and buildings, vol. 134; 2017. p. 11–24.
[19] Martin Heine Kristensen, Rasmus Elbæk Hedegaard, Petersen Steffen. Hierarchical calibration of archetypes for urban building energy modeling. In: Energy and buildings, vol. 175; 2018. p. 219–34.
[20] De Jaeger Ina, Lago Jesus, Saelens Dirk. A probabilistic building characterization method for district energy simulations. In: Energy and buildings, vol. 230; 2021. p. 110566. https://doi.org/10.1016/j.enbuild.2020.110566. issn: 0378-7788, https://www.sciencedirect.com/science/article/pii/S0378778820333521.
[21] Gholami M, Torreggiani D, Tassinari P, Barbaresi A. Narrowing uncertainties in forecasting urban building energy demand through an optimal archetyping method. In: Renewable and sustainable energy reviews, vol. 148; 2021. p. 111312.
[22] Nageler P, Koch Andreas, Mauthner Franz, Leusbrock Ingo. Thomas Mach, Christoph Hochenauer, and Richard Heimrath. "Comparison of dynamic urban building energy models (UBEM): sigmoid energy signature and physical modelling approach. In: Energy and buildings, vol. 179; 2018. p. 333–43.
[23] Kathleen M C Tjørve, Even Tjørve. In: The use of Gompertz models in growth analyses, and new Gompertz-model approach: an addition to the Unified-Richards family. Plos one, vol. 12; 2017. p. 1–17. https://doi.org/10.1371/journal.pone.0178691.
[24] Madsen Henrik. Models and methods for predicting wind power. English; 1996.
[25] O'Grady Malgorzata, Lechowska Agnieszka A, Harte Annette M. Quantification of heat losses through building envelope thermal bridges influenced by wind velocity using the outdoor infrared thermography technique. In: Applied energy, vol. 208; 2017. p. 1038–52. https://doi.org/10.1016/j.apenergy.2017.09.047. issn: 0306-2619, https://www.sciencedirect.com/science/article/pii/S0306261917313284.
[26] Foteinaki Kyriaki, Li Rongling, Heller Alfred, Christensen Morten Herget, Rode Carsten. Dynamic thermal response of low-energy residential buildings based on in-wall measurements. In: E3S web of conferences, vol. 111. EDP Sciences; 2019. p. 4002.
[27] Rasmussen Christoffer, Frolke Linde, Bacher Peder, Madsen Henrik, Rode Carsten. English. In: Solar energy, vol. 195; 2020. p. 249–58. https://doi.org/10.1016/j.solener.2019.11.023. issn: 0038-092X.
[28] Hammarsten Stig. A critical appraisal of energy-signature models. In: Applied energy, 26.2; 1987. p. 97–110.
[29] Rønneseth Øystein, Sartori Igor. Possibilities for supplying Norwegian apartment blocks with 4th generation district heating (ZEN Report No. 8). Tech Rep SINTEF; 2018. p. 9–13.

[30] Rasmussen Christoffer, Bacher Peder, Cali Davide, , Henrik Aalborg Nielsen, Madsen Henrik. Method for scalable and automatised thermal building performance documentation and screening. In: Energies, 13.15; 2020. p. 3866.
[31] Madsen Henrik, Paul Thyregod. Introduction to general and generalized linear models. Chapman Hall; 2011.
[32] Madsen Henrik. Time series analysis. Chapman Hall; 2008. 978-1420059670 0.
[33] Byggtekniskforskrift (TEK 10). Standard. Direktoratet for byggkvalitet. 2011. Feb.
[34] Kristensen Kasper, Bell Brad, Hans Skaug, Magnusson Arni, Berg Casper, Nielsen Anders, MAechler Martin, Michelot Theo, Brooks Mollie, Forrence Alex, , Christoffer Moesgaard Albertsen, Cole Monnahan. Template model builder: a general random effect tool inspired by. ADMB; 2021.
[35] Kaufman Leonard, Peter J Rousseeuw. Finding groups in data: an introduction to cluster analysis, vol. 344. John Wiley & Sons; 2009.
[36] Measurement of Energy. Demand, and water savings. Standard. ASHRAE; 2014. Dec.