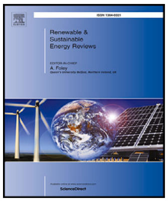
Renewable and Sustainable Energy Reviews 158 (2022) 111976

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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



On the accuracy of Urban Building Energy Modelling

A. Oraiopoulos [*], B. Howard

School of Architecture, Building and Civil Engineering, Loughborough University, LE11 3TU, Leicestershire, United Kingdom

ARTICLE INFO

Keywords:
Urban Building Energy Modelling
Systematic analysis
Singular taxonomy
Calibration
Validation
Accuracy
Error

ABSTRACT

The growing demand for energy in urban areas has led to the development of a variety of methodologies for modelling energy in buildings at large scale. However, their accuracy has yet to be thoroughly reviewed. This paper presents a systematic analysis of urban building energy models, that have been validated against measured data, using a singular taxonomy based on key attributes that could influence a model's accuracy: application, scale, input data, computational method, calibration and validation methods. The analysis showed that the accuracy of urban building energy models is multi-dimensional, considered at a variety of temporal resolutions, spatial resolutions and measures of error, with the results demonstrating that there is no single key attribute that governs it. At the aggregate spatial and annual temporal resolutions, the accuracy, often reported in a single percent error value, can be as low as 1%, while for individual buildings at the annual resolution, the tails of the distribution of errors can reach 1000%. Models using non-calibrated physics-based computational methods were more likely to report overly large errors, while those employing Bayesian calibration consistently reported lower errors at the hourly temporal resolution, demonstrating the positive impact of calibration and in particular the Bayesian approach, on the models' accuracy. Overall, the review has highlighted that more transparent and consistent reporting of accuracy is necessary and further research is essential for improving the evaluation of accuracy in modelling methodologies, if modern challenges are to be met through emerging applications such as energy systems integration and climate resilience.

1. Introduction

The energy crisis in the 1970's had a profound effect on the need to implement energy conservation strategies on a large scale, to ensure security of supply for nations across the world. Policy makers realised the importance of energy studies and as a consequence the quantity of research on energy demand on a large scale increased rapidly, with universities, government agencies and companies developing models for the evaluation of energy policy [1]. The early Urban Energy Models (UEMs), predicting energy demand on a large scale, were strictly analytical, basing forecasts on the statistical relationships between demographic data, often produced by nation-wide surveys, and energy consumption records that state-owned utilities would hold [2]. The output of these models was used mainly for the development and evaluation of national policies on the energy sector [1] and were also applied to estimate the impact of conservation measures on energy demand reduction [3,4], since it was recognised that buildings offered a large potential for demand reduction. However, the very low spatial and temporal resolution of the input data, would not allow buildings to be modelled in great detail.

This drawback was addressed partly by developing models predicting the energy demand of individual buildings, shifting the research interest to the so called Building Energy Models (BEMs). Initially these were simple command-line interfaces that would calculate the dynamic exchanges of energy within a building and between a building and its external environment [5]. These were then improved to include plant and mass flow modelling and over the 1990s they were advanced, and focus was given to their validation [6]. By the early 2000's, BEMs were developed to sophisticated software programs that perform multiple calculations in a fraction of time, to produce energy demand forecasts with great accuracy [7]. However, these models did not have the capability of producing outputs at large scale, due to the difficulty of modelling the complexity of buildings and the interactions between them and their surrounding environment, at scale. Therefore, research on the Urban Building Energy Modelling (UBEM) field expanded, and continues to until today, in an attempt to address this limitation. New tools as well as combinations of existing software, have been developed and tested to accommodate a plethora of research questions. Over the years these have been captured in an increasing amount of review articles making it necessary to justify the need for this review, set the review's scope and identify its relevance.

Previous reviews have covered the field of Urban Building Energy Modelling from different perspectives in order to explore their unique

[*] Corresponding author.
E-mail address: a.oraiopoulos@lboro.ac.uk (A. Oraiopoulos).

https://doi.org/10.1016/j.rser.2021.111976
Received 5 March 2021; Received in revised form 15 November 2021; Accepted 1 December 2021
Available online 20 January 2022

Abbreviations

ANN	Artificial Neural Network
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BEM	Building Energy Model
CVRMSE	Coefficient of Variation of Root Mean Square Error
EPC	Energy Performance Certificate
HVAC	Heating Ventilation Air Conditioning
LiDAR	Light Detection and Ranging
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MBE	Mean Bias Error
NMBE	Normalised Mean Bias Error
RMSE	Root Mean Square Error
SVM	Support Vector Machine
UBEM	Urban Building Energy Model/Modelling
UEM	Urban Energy Model

objectives. In earlier reviews, the lack of available data, specifically the low temporal resolution of publicly accessible data sets, was a common theme that was highlighted as a hindrance to the field [8,9]. In more recent years, reviews documented that technological advances in computer software, sensor equipment and cloud computing have allowed researchers to take energy systems as well as the urban environment into consideration [10–13]. Further, the availability of higher temporal and spatial resolution data has enabled the use of statistical methods, and in particular machine learning techniques, for the modelling of energy demand on a large scale [14–17]. However, previous reviews have not provided a detailed focus on the accuracy of the methodologies in modelling building energy demand on a large scale.

The challenges of the field were listed by Keirstead et al. [18], including the uncertainty imposed by input data, mainly due to measurement errors, but did not reflect what would be the impact on the accuracy of models. Hong et al. [19] did not include the accuracy of UBEM as one of the challenges in their 10-question based analysis of the field. Johari et al. [20] acknowledged that UBEM studies often lack validation but did not expand on this any further. Li et al. [21] dedicated a section of their work on the calibration and validation of models, yet with little focus on the accuracy of large scale studies, mostly reporting errors from single building simulations. Using qualitative terms, Abbasabadi and Ashayeri [22] explored the accuracy of data driven models, but rather in comparative manner, between different data driven models. Chalal et al. [23] analysed the prediction accuracy of certain methodological approaches in both building and urban spatial resolutions, but only by using the qualitative terms “fair” and “high”, without defining these quantitatively. In a more contextual form, Sousa et al. [24] provided various models with scores, taking accuracy into account, though without the evaluation of a practical application (e.g. assessing predictions against measured data), solely by considering if the models had been tested and validated or not. In the only review that presented the accuracy of UBEM studies in quantitative terms, Reinhart and Cerezo [25] reported the errors of models at the spatial resolution at which they were validated, either at single building or at the aggregate, with errors ranging from 1%–99%, acknowledging the important role of the models’ application (e.g. peak load analysis) when considering the accuracy of UBEM. More recently Qian Ang et al. [26], explored the field through use cases and focused on producing guides for developing UBEMs for specific applications. The authors reported the accuracy of models, yet further exploration into the large inaccuracies of the presented studies was not provided, leaving a gap in the knowledge of the field.

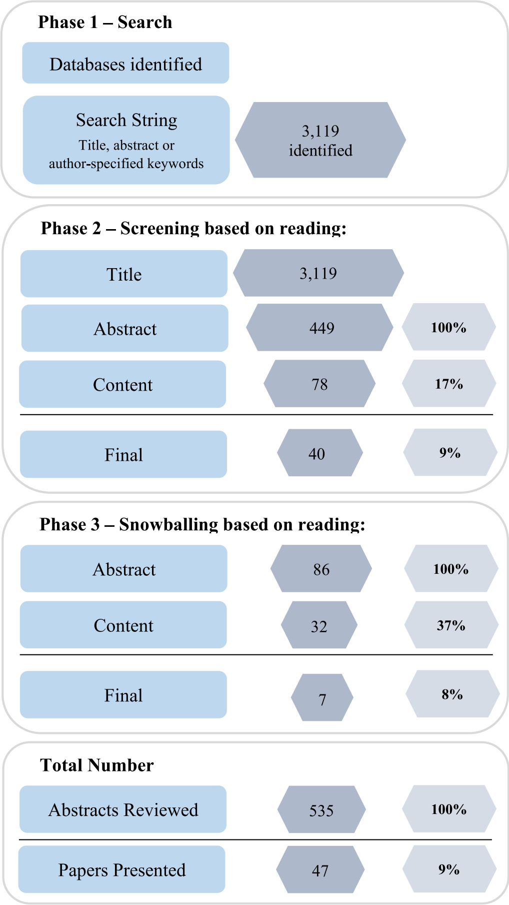


Fig. 1. Schematic of evidence selection phases illustrating the number of documents at each phase and percentages of remaining studies, counting from the first abstract reading in Phase 2.

The current work contributes to the field by addressing this gap and providing a systematic in-depth analysis of studies that have quantified the error, against measured data, of their urban building energy models. This is needed as UBEM is being called upon to answer the urgent calls for how to develop net-zero greenhouse gas emission cities and communities. As the first step towards net-zero emissions is reducing energy demand, UBEM is being called not to only provide estimates of relative changes in demand but estimates of absolute changes i.e. not just 40% reduction but a reduction of 300 GWh. Further with the introduction of distributed generation and demand side response, UBEM is being asked to provide estimates of energy demand, and the effects of any intervention, at higher spatial and temporal resolutions. Therefore, it is imperative that the UBEM community, understand the accuracy UBEMs have been able to achieve in practice, in order to direct how modelling efforts should be improved to meet these new challenges.

2. Methodology

2.1. Literature identification

The review methodology for the identification of the appropriate literature was completed in three phases (see Fig. 1).

The main databases for the literature searches were Scopus, Google Scholar and Science Direct. These are some of the biggest and most widely used databases throughout the field of engineering in academia. Scopus is the largest abstract and citation database of peer-reviewed literature, with nearly 70 million records and over 34,000 peer-reviewed

Table 1
Qualitative decision criteria for the inclusion or exclusion of studies.

Decision	Criteria	Phase 1	Phase 2	Phase 3
Inclusion	Peer reviewed	x		
	Written in English language	x		
	Accessible online or in British Library	x		
	Title relevant to UBEM field		x	
	Modelling is a prime objective		x	
	Prediction/Forecasting of building energy demand		x	x
	Data for more than one building		x	x
	Analysis includes validation against measured data		x	x
	Earlier (or later) study with same model but different data scale		x	x
Exclusion	Any studies published after January 2021	x	x	x
	Not a forecasting, rather a mapping study		x	
	Does not include energy demand as input/output		x	
	Conference paper unless novel, significant findings		x	x
	Earlier (or later) study with same model, different data but same scale		x	x
	Earlier (or later) study with same data in different Journal		x	x

journals, updated daily [27]. Google Scholar is the largest freely accessible web search engine that indexes the full text or metadata of scholarly literature [28] with an estimated 389 million records. Science Direct is a database hosting over 12 million pieces of content from 3500 academic journals of which over 1.2 million are open access [29]. The searches were based on the title, abstract or author-specified keywords containing all or some of the following search words: urban, city, large, scale, building, housing, energy, electricity, heating, cooling, modelling, model, bottom-up, top-down. Furthermore, Loughborough University's Library resources, with access to over 200 databases (such as ArXiv, SprinkerLink and Web of Science), were utilised in complementary manual searches throughout the completion of the presented work.

In the first phase 3119 studies were initially identified using the search words and string : "TITLE-ABS-KEY (urban OR large AND scale OR city OR city AND scale) AND TITLE-ABS-KEY (building OR housing) AND TITLE-ABS-KEY (energy OR electricity OR heating OR cooling) AND TITLE-ABS-KEY (modelling OR model) OR TITLE-ABS-KEY (bottom AND up OR top AND down)" in the online databases and sources of information. The selection of these keywords for this initial search was based on previous literature reviews in the field. It has to be noted however, that additional manual searches were performed after having assessed phases one and two, leading to further additions to the list of reviewed studies, as more knowledge was gained while conducting this research. In order for these 3119 studies to be assessed, they had to be written in English, peer reviewed and available either online or in the British Library. Furthermore, any studies published after January 2021 were not included, as outlined in Table 1.

In the second phase these were screened in three separate stages. Firstly, based on the relevance of the title to the field of Urban Building Energy Modelling 2670 studies were omitted, leaving 449 abstracts for review. In the second screening stage, the 449 abstracts were read by the author and 371 were omitted based on several criteria, leaving 78 for full review. In the third and final screening stage of this second phase, having read the 78 selected papers, it was decided that only 40 would be appropriate for the scope of this paper.

In the third phase of the review, from the 78 documents that were fully reviewed, 86 references were selected for further reading and out of these 7 were included in this paper, bringing the total number of the papers presented in this work to 47. It has to be noted that four of these studies included the development and analysis of two different urban building energy models. Therefore, although the number of studies presented as the core of this work is 47, the number of models analysed is 51.

Several criteria were applied during the three phases for the inclusion or the exclusion of studies. First and foremost the studies had to present data from more than one buildings in their analysis which had to include validation against measured data. The prime objective of the study had to be modelling and the prediction of building energy demand. In a similar manner, papers presenting mapping studies without having the building energy demand as output were excluded. To ensure the quality of the included studies, these had to be published in peer reviewed journals unless a significant finding was presented in other formats. Also studies replicating their methodology by applying the same modelling approach to different data of a similar scale were excluded as well as those replicating research in different journals. Table 1 outlines all the selection criteria in all three phases.

Finally it has to be noted, that the final number of studies that fall within the scope of this study (47) was less than 10% of the total number of abstracts (449 + 86 = 535) that were considered as relevant for review based on the criteria presented in Table 1.

2.2. Systematic analysis

The studies that were found to be in line with the scope of this paper were reviewed in detail and their content was systematically analysed, using a singular taxonomy based on the following key attributes that can affect the accuracy of the developed UBEM: application, scale, input data, computational method, calibration method and validation method. The following paragraphs will describe each of these elements as they form the main categorisation of the reviewed papers.

2.2.1. Application

The identification of UBEM applications was based on inferences made from key words in the main body of text of the reviewed papers, often linked to hypothetical arguments with regards to the intended uses of the models. Whilst these models would most likely be used as inputs to further analysis, these intentions were not alluded to in the works themselves. The applications were therefore classified in four main categories: (a) energy efficiency retrofit analysis; (b) energy demand quantification; (c) energy systems integration; (d) climate resilience. Energy efficiency retrofit analysis describes works that aimed to use the model to evaluate different energy efficiency measures throughout the building stock, e.g. changing all windows from single glazing to triple glazing. Energy demand quantification describe works whose aim was simply the quantification of energy demand. Energy systems integration describe works whose intent is to use the models to develop designs or plans for future energy systems such as distributed generation, district heating, or estimating the effects of demand response on local electrical infrastructure. Lastly, climate resilience describes works whose aim was to determine the ability of the local building stock to cope and adapt to changing climate with respect to both energy consumption and internal temperatures. The application a UBEM is used for should provide an expectation for its accuracy, since, for example, the requirements for an annual planning assessment will differ from the design of a local energy system. Therefore, the application was included as a category of this review to see if the intended use of the developed UBEMs effected the achieved accuracy.

2.2.2. Scale

In UBEM, it is often unclear what constitutes a block of buildings, a neighbourhood, a district, a community or a city. To assist the analysis, this work has divided the studies based on a general framework that comprises of three different scales; the micro, the meso and the macro [30]. The micro scale is assigned to studies that have a dataset containing more than one building up to hundreds (<1000 buildings), the meso scale to those comprising of thousands (1000–999,999 buildings) and the macro scale is assigned to those that include millions of buildings (≥1,000,000). The number of buildings considered in the analysis determines the intrinsic variability in the energy consumption data. Therefore, this phenomena must be considered when exploring the accuracy of UBEM as it could indicate the predictability of the underlying energy demand.

2.2.3. Input data

The input data requirements include diverse sets of information, depending on the type of model developed. Here the input data have been divided to those related to: geometry, fabric, systems, controls, occupancy, energy and temperature. The geometry refers to the representation of the buildings, the fabric to their thermal properties, the systems to the heating ventilation and air conditioning (HVAC) parameters, the controls to the operational variables of the systems, the occupancy to the presence of occupants and their behavioural actions, the energy to the energy consumption input data and the temperature to the internal temperature time series data (the external weather data were not analysed across the studies and therefore did not form part of the analysis). Moreover, the relevant information of input data has been categorised as measured, estimated or assumed, depending on the method with which the data were collected and also to unknown and not applicable, for absent or not required data respectively. “Measured” was assigned to data that have been collected using empirical measurements (i.e. lidar, energy use monitoring, utility measured readings), “estimated” was assigned to data that have been inferred using national registers, surveys, census data, online mapping tools (i.e. openstreetmap, google maps), and “assumed” was assigned to information that has been clearly stated forming part of the assumptions in the modelling. Information that was regarded as essential for the computational method but was completely absent from the text in the reviewed studies was classified as “unknown”. Finally data that was regarded as not required for the computational method to operate was assigned as “not applicable”.

2.2.4. Computational method

The computational method characterises the mathematical structure used for defining the relationships between various attributes. This work considers three categories of approaches: the statistical methods that make use of statistical relationships between input and output variables, physics-based methods that are based on energy balance and heat transfer equations, and hybrid methods that use a combination of statistical and physics-based approaches. Taking into consideration the computational method enables a comparison of the accuracy achieved in practice for each type of model.

2.2.5. Calibration method

With computational method and input data defined, the final steps of UBEM are the processes of calibration and validation. These approaches can vary significantly depending on the chosen computational method i.e. physics-based versus statistical methods, but is an essential aspect of delivering accurate models and is therefore considered.

The calibration method is a process whereby, various inputs to a model are fine-tuned so that the predicted values of the output match closely those obtained by experimentation (i.e. measurements) [31]. The type of computational method often determines the approach followed in the calibration process, with three main techniques being identified, the Bayesian calibration where certain parameters are given a probability distribution instead of fixed value, the bespoke calibration where certain parameters are altered until the output of the model matches a set of measured data, and the statistical training where part of the dataset is used to calculate the parameters of the model’s variables.

2.2.6. Validation method

The validation method refers to the sole process of comparing the results of a model to measured data. The objective is to examine the performance of the model given a certain dataset. There are three main validation methods, the apparent, the internal and the external, based on whether the measured dataset used for comparing against the output of the model, has been part of the calibration process [32,33]. In apparent validation, the data used for the model’s validation have also been used for its calibration, in internal validation the dataset is divided to data used for validation and data used for calibration and in external validation the dataset used for validation is entirely different to that used for calibration.

The remainder of the review, presents a discussion on the characteristics of the validated studies, the accuracy they were able to achieve and the implications of these studies for the UBEM field.

3. Results and discussion

This work, examined studies that present UBEM validated against measured data. In the literature search, from the 535 abstracts that were reviewed only 47 studies were found to include validated models, representing a mere 9% of the research in the field. Table 2 summarises the findings of the systematic analysis for all the 51 models described in the 47 studies presented in this paper, by classifying each of the models with respect to its application, scale, input data, computational method, calibration and validation methods as well as providing the reported accuracy of the models. It serves as the main reference point throughout this paper, allowing immediate comparison between all the analysed models.

This section reports on the distribution of the attributes in UBEM studies (application, scale, input data, computational methods, calibration and validation), by exploring these with respect to the achieved accuracy of the models. First though, it is important to present the key elements in reporting accuracy as found in the reviewed studies: the spatial resolution, the temporal resolution and the measures of errors.

3.1. Reporting accuracy: Spatial resolution, temporal resolution and measures of error

The systematic analysis of this work revealed that the accuracy of the output in UBEM studies is multidimensional, reported at varied temporal resolutions and spatial resolutions, using a number of different measures of error.

3.1.1. Spatial resolution

Spatial resolution refers to whether the error of the output has been calculated for a cluster of buildings (aggregate) or on a per building basis. Overall, the accuracy of UBEM can be high (1%) when evaluating results on an aggregate spatial resolution (i.e. for many buildings together)[55]. However, this decreases as the models are tested on finer spatial resolution (i.e. single buildings). As the spatial resolution changes from many buildings to individual ones, the error in the results has been reported to increase from less than 10% [37,41,53] up to tenfold [60] or more [47]. In some cases the reported error can be up to 1000% [61]. Additionally, the variation in the results of UBEM, when evaluating the accuracy in individual buildings, can have a large spread, with reporting error ranges of 2.5%–262% [67]. Most importantly, when reporting the error on a per building case, the result is often not presented as a distribution across the sample, but rather as a mean figure or at best as a range, masking the details of the output and therefore increasing the uncertainty and failing to achieve full transparency in the results.

Table 2

UBEM review results (ERA: Energy Efficiency Retrofit Analysis, EDQ: Energy Demand Quantification, ESI: Energy Systems Integration, RES: Climate Resilience; M: Measured, E: Estimated, A: Assumed, U: Unknown, NA: Not Applicable; P: Physics-based, S: Statistical, H: Hybrid; ST: Statistical Training, Be: Bespoke, BA: Bayesian, NS: Not Specified; Ap: Apparent, Ex: External, In: Internal, a: annual, m:monthly, d:daily, h:hourly).

Reference			Input data										Accuracy in aggregate				Accuracy in single building			
	Application	Scale	Geometry	Fabric	Controls	Systems	Occupancy	Energy	Temperature	Computation	Calibration	Validation	(%)	R^2	CVRMSE	NMBE	(%)	R^2	CVRMSE	NMBE
Reiter 1980 [4]	ERA	Meso	E	E	NA	NA	NA	M	U	P	ST	In	d: 6.26							
Reiter 1980 [4]	EDQ	Meso	NA	NA	NA	NA	NA	M	NA	S	ST	In	d: 5.94							
Shimoda et al. 2007 [34]	ERA	Meso	E	A	A	E	E	NA	U	P	NS	Ex	a: 1							
Strzalka et al. 2011 [35]	ERA	Micro	M	A	A	U	U	NA	NA	P	NS	Ex	a: >30				a: 3–26			
Aranda et al. 2012 [36]	EDQ	Micro	M	E	U	E	E	M	U	S	ST	Ap		a: 0.6			a: 27–40			
Lorimer 2012 [37]	EDQ	Meso	NA	NA	NA	NA	M	M	NA	S	Be	Ex	a: −4 to +5	a: 0.4						
Ren et al. 2012 [38]	ERA	Meso	E	E	A	A	E	NA	NA	P	NS	Ex	a: −9	a: 0.5			a: 11, d: 10, h: <300			
Howard et al. 2012 [39]	ERA	Micro	E	NA	NA	NA	NA	M	NA	S	ST	In	a: ±20							
Booth et al. 2012 [40]	EDQ	Macro	E	E	NA	NA	NA	M	NA	P	BA	Ex	d: 0.005							
Filogamo et al. 2014 [41]	ERA	Macro	E	E	U	E	U	NA	NA	P	NS	Ex	a: −7.8							
Mastrucci et al. 2014 [42]	ERA	Meso	E	E	NA	NA	E	M	NA	S	ST	In	a: 2–5	a: 0.7–0.8			a: ±20			
Fonseca et al. 2015 [43]	ERA	Meso	E	E	U	E	E	M	NA	H	NS	Ex	a: 1–19				a: 4–66			
Quan et al. 2015 [44]	EDQ	Meso	E	U	U	U	E	NA	NA	P	NS	Ex					a: −800 to +800		a: 69–85	
Nouvel et al. 2015 [45]	ERA	Meso	E	U	U	U	U	M	NA	P	NS	Ap	*MAPE: 49							
Nouvel et al. 2015 [45]	ERA	Meso	E	U	U	U	NA	M	NA	S	NS	Ap	*MAPE: 26							
Cerezo et al. 2016 [46]	ESI	Meso	E	E	U	E	E	NA	NA	P	Be	Ex	a: 5–94				a: 5–20			
Osterbring et al. 2016 [47]	ERA	Micro	E	E	A	E	U	NA	NA	P	NS	Ex	a: 3				a: up to 100			
Ma and Cheng, 2016 [48]	EDQ	Meso	E	E	U	U	E	M	NA	S	ST	In	*MSE: 0.75–0.99							
Nageler et al. 2017 [49]	ESI	Mi-Me	E	E	A	E	A	NA	NA	P	NS	Ex					a: −64 to +75			
Buffat et al. 2017 [50]	ERA	Mi-Me	E	E	U	E	E	NA	NA	P	NS	Ex						a: 0.1–0.8		
Sokol et al. 2017 [51]	ERA	Meso	E	E	E	U	E	M	NA	P	BA	Ex					a: 47, m: 44		a: 66, m: 58	
Nouvel et al. 2017 [52]	EDQ	Meso	E	E	A	U	E	NA	NA	P	NS	Ex	a: 10–30							
Olivo et al. 2017 [53]	EDQ	Macro	E	E	U	A	U	NA	NA	P	NS	Ex	a: 5				a: 75	a: 0.73		
Kontokosta and Tull, 2017 [54]	EDQ	Me-Ma	E	E	NA	NA	NA	M	NA	S	ST	In	*LAR: 0.3–0.8				*MAE: 1.1–1.5			
Alhamwi et al. 2018 [55]	ESI	Micro	E	NA	NA	NA	E	M	NA	S	NS	Ex	a: 1, h: 30	d: 0.83–0.94						
Kristensen et al, 2018b [56]	EDQ	Meso	E	E	NA	E	NA	M	NA	S	ST	Ex		a: 0.35	a: 32	a: 0.3				
Nageler, et al. 2018 [57]	EDQ	Micro	E	E	U	E	U	M	NA	P	Be	Ex		a: 0.92	a: 21.4			a: 0.68–0.92	a: 24.9–40.2	
Nageler, et al. 2018 [57]	ESI	Micro	NA	E	NA	NA	NA	M	NA	S	Be	Ex		a: 0.97	a: 12.5			a: 0.87–0.96	a: 17.6–25.4	
Nagpal and Reinhart 2018 [58]	ERA	Micro	E	A	A	A	A	M	NA	P	NS	Ex					a: <200	a: 0.96		
Nagpal and Reinhart 2018 [58]	ERA	Micro	E	A	A	A	A	M	NA	H	NS	Ex					a: <1000	a: 0.85		
Zhang et al. 2018 [59]	EDQ	Micro	A	U	A	A	A	M	NA	P	NS	Ex	h: <30							
Kristensen et al. 2018a [60]	EDQ	Micro	E	E	A	NA	NA	M	U	P	BA	Ex			h: 7.8 ±2.9	h: 2.9 ±6.2			h: 28.7–120.1	h: −38.7–112.9
Wang et al. 2018 [61]	EDQ	Micro	E	E	E	E	E	NA	NA	P	NS	Ex	a: 1.05				a: <1000			
Panao and Brito, 2018 [62]	ESI	Meso	E	E	U	U	A	M	NA	P	Be	Ex	a: −26							
Nutkiewicz et al, 2018 [63]	ERA	Micro	E	E	E	E	E	M	NA	H	ST	Ex			m:11.4 d:14.4 h:25.6				m:27.9 d:31.3 h:46.0	
Koschwitz et al. 2018 [64]	ESI	Micro	NA	NA	NA	NA	A	M	NA	S	ST	In	*MSE: 52.5							
Moghadam et al. 2018 [65]	ERA	Micro	E	E	U	E	E	M	NA	S	ST	In		d: 0.8						
Katal et al. 2019 [66]	RES	Meso	E	A	E	U	A	NA	NA	P	NS	Ex	a: 47				a: up to > 100			
Krayem et al. 2019 [67]	RES	Meso	E	E	U	A	E	M	NA	P	Be	Ex					m: 2.5–262			
Kim et al. 2019 [68]	ERA	Meso	E	E	A	E	A	NA	NA	H	NS	Ex							m: 3.3–51.2	m: −14.8 to 3.7
Xu et al. 2019 [69]	EDQ	Micro	E	E	NA	NA	NA	M	NA	S	ST	In		m: 0.86						
Yi and Peng, 2019 [70]	RES	Micro	E	E	U	A	A	M	NA	S	Be	In		m: 0.97						
Gassar et al. 2019 [71]	EDQ	Macro	E	NA	NA	NA	NA	M	NA	S	ST	In		a: 0.91–0.99						
Hedegaard et al, 2019 [72]	ESI	Micro	E	A	A	E	E	M	NA	P	BA	Ex			h: 5.6	h: −1.39			h: up to 53.7	h: −8.47 to 15.3
Krati et al. 2020 [73]	ERA	Macro	E	E	U	U	E	M	NA	P	Be	Ex	a: 2							
Streltsov et al. 2020 [74]	EDQ	Meso	E	NA	NA	NA	NA	E	NA	S	ST	Ex		a:0.48–0.97				a:0.28–0.36		
Jahani et al. 2020 [75]	EDQ	Meso	E	E	E	E	E	M	NA	H	NS	In	a:9, m:1–10		m: 7.8	m:4.5				
Tardioli et al. 2020 [76]	EDQ	Meso	E	E	E	E	E	M	NA	P	BA	In	a: 2–8.2							
Roth et al. 2020 [77]	EDQ	Macro	E	E	NA	NA	NA	M	NA	H	ST	In	*MAPE h:5–10							
Yang et al. 2020 [78]	ERA	Meso	E	E	E	E	A	M	NA	P	NS	Ex			h:31					
Fernandez et al. 2020 [79]	EDQ	Meso	E	E	A	E	A	M	NA	P	Be	Ex	a: 1–10							

3.1.2. Temporal resolution

Temporal resolution refers to the discrete resolution of the measured data used for validation with respect to time. If a model is capable of producing results in hourly (or even sub-hourly) resolution, but the available measured data are in annual form, the evaluation of the accuracy of the model will be limited to the annual resolution. This has been a challenge since early studies [4], restricting the performance evaluation of the developed models. As data became more accessible and models were evaluated in more than one temporal scales, results revealed the difficulty in maintaining high accuracy as temporal resolutions change. The most common temporal resolution used to report the accuracy of UBEMs is the annual, with 35 models reporting error in this temporal resolution, 7 reporting in monthly, 6 reporting in daily, and 6 reporting in hourly. Whilst the majority of papers reported errors at a single temporal resolution, a few reported at multiple, enabling a description of the accuracy as the temporal resolution increased. Studies indicated that when switching from annual to hourly temporal resolution, errors can increase from fivefold [62] to more than 20 times [38,55].

3.1.3. Measures of error

The accuracy of UBEM is defined as the error in the output of the model when compared to measured data and is reported using statistical measures. These measures can range from a simple percentage difference between measured and simulated (modelled) data, to more complex mathematical formulas designed to indicate specific faults in the model. This work has explored all the measures of error reported in the presented studies including: coefficient of determination (R^2), root mean square error (RMSE), coefficient of variation of RMSE (CVRMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), mean biased error (MBE), normalised MBE (NMBE). In most of the earlier studies (until 2016) the chosen approach was to report a single value for the difference between the measured and the simulated annual energy demand, as a percentage [4,34,41]. This evaluated the accuracy of estimates of energy demands of hundreds of buildings by a single measure. Another widely used method was to report the percentage of the results that lays within a ±20% margin of error from the measured values [39,47,54]. This approach still considered annual measures of demand but reported some indication of the spread of the distribution of the results. However, as the interest in UBEM increased, researchers started applying multiple measures of error, to get a better understanding of the performance of models, amongst which were the coefficient of variation (CV), the root mean square error (RMSE) and the coefficient of determination (R^2) [37].

The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Guideline 14-2002 for measurement of energy and demand saving recommendations [80] was first applied in UBEM by Quan et al. [44], introducing the normalised mean bias error (NMBE) and the coefficient of variation of the root mean squared error (CVRMSE) to the field of UBEM. The use of these has been criticised by Hedegaard et al. [72], reporting that these measures alone are not sufficient in evaluating the performance of models. In more recent years, researchers have explored applying an increased number of measures of error (five or more), to provide a better and more transparent evaluation of their developed models. Among the measures that have been used are the mean absolute percentage error (MAPE) and the dimensional mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) [69–71], which as non-normalised measures cannot be used for comparison across studies.

Despite the increased use of multiple measures of error, the vast majority of those used, do not capture important features of models' strengths or drawbacks, emphasising mainly in the mean behaviour of the output of a UBEM. This way they fail considerably in highlighting any discrepancies in terms of hourly shifting or in terms of peaks or troughs (e.g. hourly or daily maximum energy demand), which are key for applications looking at the security of supply but also for the accurate evaluation of demand side response implementation programmes. Nonetheless, the increasing use of multiple measures of error within a study in order to examine the accuracy of a model, transforms the performance metric of UBEM from a single statistic to a suite of multiple measures of error, increasing the depth of research in UBEM accuracy and partly adjusting this scientific field to modern challenges.

3.2. Impact of key attributes on UBEM accuracy

The systematic analysis of the literature was performed using a singular taxonomy which identified the key attributes in UBEM. These were evaluated both in terms of their characteristics and their relation to the accuracy reported in the studies, as described in the following paragraphs.

3.2.1. Application

The applications of UBEM are an essential attribute that should form a central part in the development of the models. However, setting a specific application as the explicit use of a UBEM could suggest an expectation for the anticipated accuracy of the model. In most of the reviewed studies, a specific application could not be clearly inferred from the text but rather hypothetically an intended use for the model was sometimes stated. Overall, energy demand quantification (EDQ in Table 2) was the intended application in 21 of the models presented here, while 18 models looked at energy efficiency retrofit analysis (ERA). Energy systems integration (ESI) was only assessed in 7 models and climate resilience (RES) was the application of least concern amongst the presented models.

When looking at the aggregate and annual spatial and temporal resolutions respectively, it can be observed that there are little differences between applications in terms of accuracy. Models applied for energy demand quantification can present errors of 1%–10% [37,76,79] but even up to 30% [52], ranges that appear also for energy efficiency retrofit analysis, with lower annual aggregate values around 1%–2% [34,73] but as high as 20% [43] or even 30% [35]. The same can be found in models intended for energy systems integration, with errors as low as 1% [55] but as high as 24% [62] and similarly for the case of climate resilience (47% [66]).

Examining each application, for energy demand quantification, Nageler et al. [57], for their model, using a physics-based computational approach and intended to explore physical densification and demand forecast, reported values of CVRMSE between 24.9–40.2% for individual buildings. Kristensen et al. [60] developed a UBEM that managed to explain approximately about 50% of the variability in the predicted annual heating energy use of randomly selected buildings, reporting CVRMSE values of up to 120% for single buildings. For energy efficiency retrofit analysis, Nutkiewicz et al. [63] developed a UBEM that could investigate key decisions related to energy efficiency early on in the design process and as part of retrofit programmes and reported CVRMSE values of up to 46% for individual buildings, while Yang et al. [78] reported CVRMSE values of 31% for their model, intended to be used for energy efficiency suggestions, by city planners and local authorities. Osterbring et al. [47] calculated percentage errors of up to 100% when addressing the impacts of energy efficiency measures in a local building portfolio. Even for more critical applications such as energy systems integration, CVRMSE values of 25.4% and 53.7% were reported for load profile and demand response operations respectively [57,72]. Finally, for the rapidly expanding application of climate resilience, researchers reported large percentage errors for individual buildings of up to 100% when exploring the resilience of buildings against a three-day power outage due to snowstorms [66], and even up to 262% when attempting to test power outages due to energy peak demand [67].

It is therefore clear, from the evidence brought forward by this review, that currently the ranges of accuracy in the presented UBEM studies do not differ substantial between different applications and thus examining the use cases of UBEM through the attribute of application solely does not provide fruitful insights in terms of accuracy.

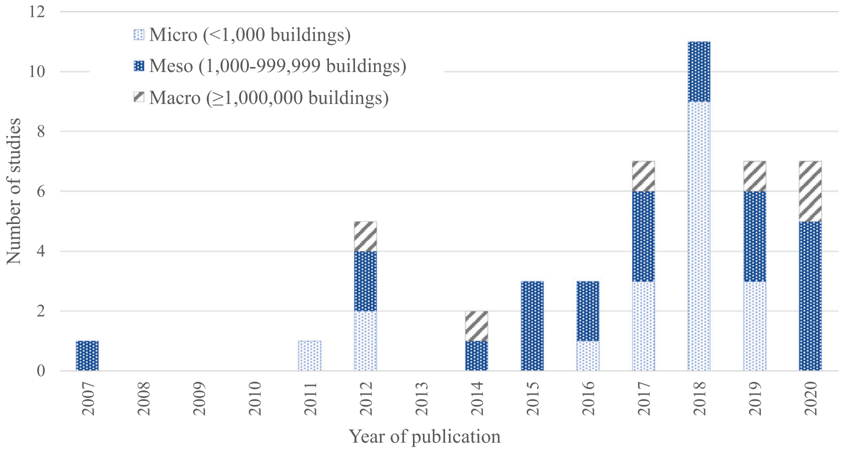


Fig. 2. Number of validated UBEM studies per year in the micro/meso/macro scale in recent years (excluding [4]).

3.2.2. Scale

The categorisation of studies into quantifiable scales allows for insights in relation to the number of buildings modelled. Evidently, validated macro scale models, in studies modelling over 1 million buildings, have rarely been achieved, most likely due to the difficulty in capturing energy data for validation at that scale. However, at the micro and meso scales, several studies have been validated contributing to a wide set of applications and contextual environments.

Micro scale studies (involving less than 1000 buildings) have seen a considerable surge in recent years, (see Fig. 2). The micro scale has the most relatively diverse range in terms of number of buildings, offering a very wide range of applications to be explored. From the energy efficiency retrofit analysis of a small number of university campus buildings [63] or hundreds of residential and commercial buildings in mixed developments [35,47], to energy demand quantification by providing assistance in evaluating energy efficiency policies [57], energy systems integration by thermal load forecasting [64] and demand side management and response [72]. An increased number of emerging applications, such as climate resilience, can be applied in this scale, allowing critical insights, such as forecasting the cooling demand of homes during excessive heat conditions [70]. The results in terms of accuracy in the micro scale show that the error ranges remain large (up to 1000% [61]) when assessing individual buildings but can be as low as 1% [55] when evaluating models at the aggregate spatial resolution.

The meso scale has been the standard magnitude for UBEM research over the years, as this scale captures the interest of many stakeholders and policy makers, and also possibly due to the nature of data available by local authorities and utility companies. The meso scale has captured most of the city-wide studies, featuring research from Japan, USA, Canada, Australia, Lebanon, EU and UK, mainly for the applications of energy demand quantification in urban energy infrastructure planning [48] and energy efficiency retrofit analysis looking at energy savings from energy efficiency measures in building envelopes [34] and strategies for upgrading electrical appliances [43] and HVAC systems [68]. The accuracy reported in this scale exhibited ranges between 1% and 262% for aggregate and individual buildings respectively [43,67].

The macro scale is the one containing the fewest studies, as it is of increased difficulty to acquire data for model validation at this scale. The city of New York in the USA however, has had a large archive of extensive land use and geographic data at the tax lot level, allowing the modelling of its building stock [39] for energy demand quantification. In the UK, the use of data from geographic areas, designed to improve the reporting of small area statistics, resulted in modelling a minimum 4,000,000 buildings [71]. The small number of studies in this scale (only 6) does not allow for clear outcomes, thus the accuracy findings remain largely similar to the previous two scales with ranges between 2%–75% for aggregate and individual buildings respectively [53,73].

Overall, the results with respect to accuracy, show annual percent errors of 1%–2% have been achieved at aggregate spatial resolution for micro, meso and macro scale studies. When comparing results for individual buildings, large spreads of error are found in all scales, with no clear indication that either scale enables more accurate modelling of individual buildings. However, individual studies that have evaluated accuracy at different scales, do show that the accuracy of the modelling method decreases as the number of buildings increases [63].

3.2.3. Input data and computational methods

The input data and the computational methods are inherently interlinked, hence these two attributes will be presented in one section. Firstly, the analysis of the input data and the relationship between input data and different computational methods was explored and finally the impact of the input data and the computational methods on the accuracy of UBEM was examined.

A plethora of available datasets in the public domain, such as energy consumption data, but also built environment geospatial data, from digital online maps (e.g. OpenStreetMap) and detailed semantic information, from online registers (e.g. Energy Performance Certificate databases), can be acquired and processed to create urban building energy models. Ideally, to achieve the highest possible accuracy of a UBEM, all inputs should be carefully selected and measured. However, in reality this is rarely the case. An overview of the input data in all 51 models analysed in this review is presented in Fig. 3, where is it shown that the majority of inputs are at best estimated but often assumed, and in many cases even omitted from the text in the presented studies.

Geometry of buildings is a key input variable in physics-based UBEM as it can be significantly related to the energy demand, but it is also an integral part of all spatial visualisation techniques. As it can be observed in Fig. 3 very few studies have fully measured the geometry of numerous buildings. The two studies that managed to acquire measured data for the geometry of buildings, were also able to provide the accuracy of their models in individual buildings, reporting errors of between 3%–26% and 27%–40% [35,36]. Furthermore, the use of LiDAR (Light Detection and Ranging) method, has been applied in some cases to create or improve digital elevation models [49,50]. These models are often held by local authorities, in digitised platforms, such as geographical information systems (GIS), where one can acquire the buildings' footprints and heights [46,47]. Nonetheless, most frequently, the geometry of buildings is estimated based either on archetypes, that have been developed using data from large surveys (i.e. census data) or national registers, providing estimations for the building stock [38,60], or on aerial imagery using services such as OpenStreetMaps, Google

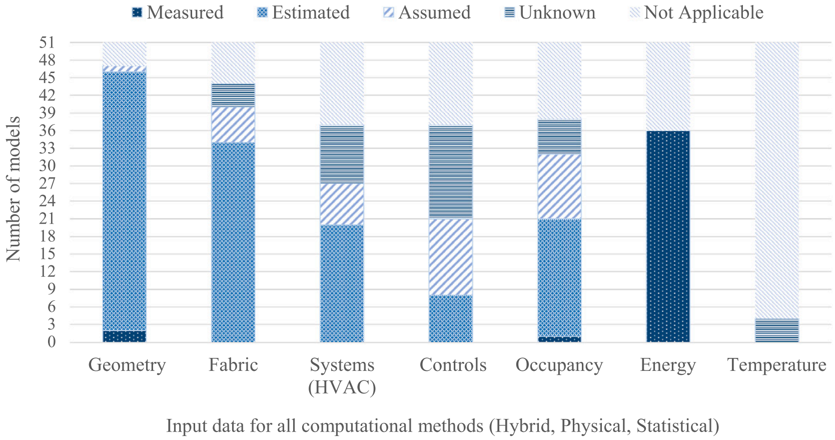


Fig. 3. Overview of input data for all UBEMs in the reviewed studies.

Maps or Google Earth. This has been the case for the majority of the models, where the aforementioned techniques were employed [55,81].

Fabric, or the construction materials comprising a building's external walls, is an essential parameter in modelling energy demand, as this is largely responsible for the thermal properties of the buildings. Obtaining detailed and accurate data about the fabric of buildings can be extremely challenging even for single buildings, let alone for larger scales, hence none of the studies presented here include measured data in relation to the building fabric (see Fig. 3). The use of archetypes has assisted in applying assumptions for this required input [43,66,73], with many of the reviewed studies having made inferences with regards to the fabric based on construction periods [41,46,65] or national codes, standards and Energy Performance Certificate (EPC) databases [47,81].

Energy systems in buildings can vary from an open fire in a domestic setting, to very complex HVAC configurations in commercial buildings. Despite the fact that computations of building energy demand are highly dependent on system related information, obtaining measurements from energy systems in a large scale requires detailed surveys that are both lengthy and costly. Therefore, frequently, the information related to the energy systems, especially in the residential sector, is estimated based on the year of construction [57,65] or the existence of district heating, or estimated based on the guidelines of professional bodies [34,68]. However, this input is commonly excluded from the text, with very few studies even stating explicit assumptions. Yet, it is worth noting that most studies which included some information about the systems, also reported the accuracy of the model for individual buildings, as shown in Table 2, reaching higher level of detail and transparency in their approach.

Information on HVAC controls is mainly used in detailed, small scale studies, usually comprising of single building analysis. Still, as the applications of the UBEM field expand to include more operational aspects, it is inevitable that information regarding the controls of HVAC systems will be regarded as vital. In the reviewed studies, as it can clearly be observed in Fig. 3, this field of input data was majorly neglected, with the main information that could be linked to the control of an HVAC system, being the assumption related to the threshold temperatures of the indoor environment for triggering control actions, often set at 27 °C for cooling and 20 °C for residential heating [59,82].

Occupancy related data is an area of extensive research the past years. Studies on individual buildings have explored a number of techniques to measure occupancy. In UBEM however, most related information is estimated or rather assumed, based on standard schedules [64,72] or based on profiles derived from analysis of hourly residential electricity use and relevant energy standards and guidelines [61,70].

Energy consumption data have been used as input data primarily for the calibration of models that use statistical computational methods, during the stage of model training [56,69], but also for the calibration of models using physics-based computational methods [67,73,76]. This stream of data can be used in different temporal resolutions (annual, monthly, daily, hourly), but as the data are mostly obtained from local or national energy providers, more often than ever are found in annual resolution [71,74]. The cases where energy data are not applicable in Fig. 3, are mostly those where the physics-based models have not been calibrated.

Temperature data is a set of input data that should be of interest to the model developers, since the use of internal temperatures can enhance the evaluation of the accuracy by increasing precision of the modelling parameters, especially during calibration [60]. However, the collection of such datasets is scarce and therefore their use in calibrating or validating models is very rare.

The different streams of input data (i.e. geometry, fabric, etc.) do not equally apply to all models but rather depend on the computational method employed in the modelling approach. Figs. 4 and 5 show the input data requirements for physics-based and statistical computational methods respectively. Looking at the portion of the input data that have been classified as "Not Applicable", it can be concluded that UBEM studies using statistical computational methods have fewer data requirements than those using physics-based.

The majority of validated UBEM models found in this study use a physics-based computational method (28) or a statistics based one (16), with hybrid computational methods found in 7 of the studies.

Statistical computational methods include a large set of techniques from which linear regression [42,60] and more recently data-driven techniques and machine learning methods, such as artificial neural networks (ANN) and support vector machine (SVM) [48,54,64], have been widely used in UBEM.

Physics-based computational methods include some simpler techniques based on energy balance and heat transfer equations [50,62], to simplify the description of the system under study, enabling compatibility with linear or mixed-integer linear optimisation based control schemes [47,72], and some more complex approaches, such as dynamic thermal simulation, which provides the ability to model buildings in great detail, with considerable advantage being the possibility of analysing decision making and control scenarios through the vast range of input variables [46,51]. In both cases the input data requirements are large, yet the highly detailed and sophisticated dynamic thermal simulation software require a considerably higher number of inputs, leading to increased amounts of assumptions and high computational costs [67]. From Fig. 4 it is clear that in many cases, information related to controls, systems and occupants is omitted from the text, deeming them as unknown. This imposes large uncertainties with regards to the inputs that form part of these models, allowing them to be characterised largely as non-transparent.

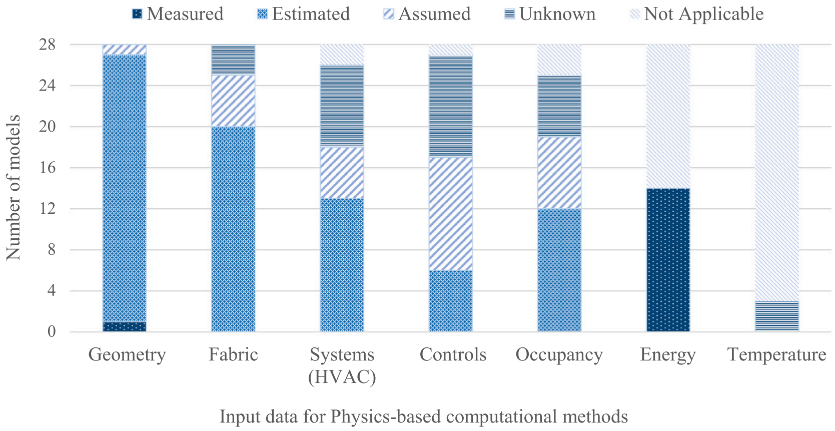


Fig. 4. Overview of input data in UBEMs with physics-based computational methods.

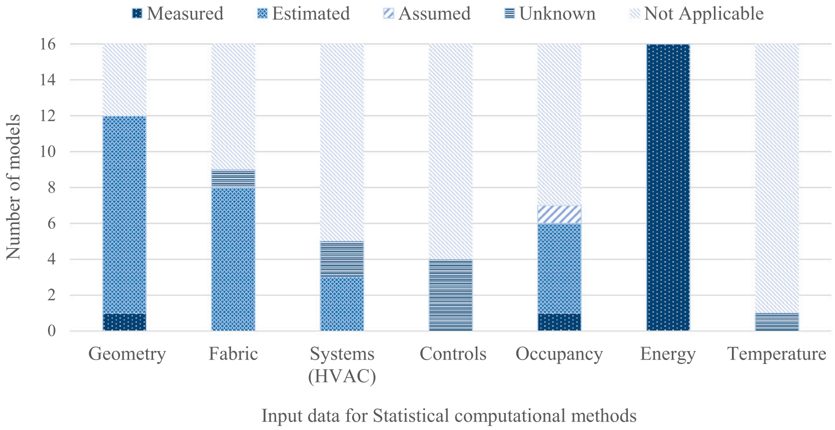


Fig. 5. Overview of input data in UBEMs with statistical computational methods.

The lack of measured data for essential parameters for building energy demand modelling, such as the geometry and the fabric of buildings, together with the regular absence of information related to the HVAC systems, controls and occupancy, from the majority of the studies, has led researchers to attempt bridging the gap of missing data by making the maximum use out of available data, via the development of hybrid computational methods.

Hybrid computational methods are a combination of statistical and physics-based methods. This synergy of different scientific fields (statistics and physics) has, in many cases, resulted in methodologies that benefit from the increased availability in data, such as energy data from utility companies. The idea of a hybrid model was presented in one of the early UBEM studies by Reiter [4], but lately, these have been increasingly explored. Hybrid models are platforms that allow the exchange of information between statistical and physics-based models. The output variables resulting from strong statistical relationships or machine learning techniques can form the input variables to physics-based models [68,75,77], and equally the output variables of detailed physics-based models based on laws of physics (i.e. heat transfer, energy balance equations) can be used to generate extensive datasets to train data-driven algorithms [63].

Overall, the evidence show that still at large, input data in UBEM are assumed or can even be considered as missing from the description of the studies. Studies that do not include estimations nor assumptions for any of their input data, have reported percentage errors mainly on the aggregate spatial resolution from −4% [37] to over 30% [59]. Conversely, studies that have attempted to include as much data as possible by measurements, estimations or clear assumptions, have reported accuracy in the single building spatial resolution and in more than one temporal resolutions as well, covering annual, monthly, daily and hourly, allowing them to be considered as suitable for an increased number of applications. Nonetheless, the increased detail and quality of the inputs for these studies does not necessarily result in improved accuracy. In Table 2 it can be seen that the latter studies report percentage errors in annual and aggregate resolutions ranging up to 10% [79] and in annual and single building resolutions up to 75% [49] or even almost 1000% [61], while values of CVRMSE have been reported to reach 14.4% for daily aggregate resolutions and 31.3% for daily single building temporal and spatial resolutions [63].

Consequently, it can be observed that the quantity and quality of input data have not had a significant impact on the accuracy of UBEM until now, but have rather enabled an increased level of analysis, allowing the investigation of up to hourly recorded variables at building level, somewhat positively exposing the inability of some models to capture the demand variability at these temporal and spatial resolutions.

Examining the computational methods, the studies with statistical approaches had the R^2 value at the aggregate spatial resolution as the most consistently reported measure of error, while among those that used physics-based, the most common measure or error was the percentage error both at aggregate and single building spatial resolutions. At the annual aggregate resolutions both approaches were able to deliver percent errors as low as 1%. Shimoda et al. [34] developed a UBEM employing a physics-based computational method for the region of Osaka in Japan and reported a 1% accuracy. Whereas Lorimer [37] employed linear regression in their model for England in the UK and reported a 1.5% accuracy. Models reporting the outcomes of statistical computational methods, are most often validated at the aggregate spatial resolution, with few reporting metrics at the individual building resolution. However, those that do, report similar ranges compared to models using physics-based computational methods. For example Mastrucci et al. [42] using a statistics based approach reported individual building level percent errors mostly between ±20% with Strzalka et al. [35] reporting percentage errors between 3 and 26%.

Nonetheless, there have been three studies that have performed a direct comparison between physics-based and statistical computational methods in the same paper. Reiter [4] developed a model using a physics-based approach based on heat transfer equations and a heuristic learning approach, and a separate model making use of statistical relationships, to independently estimate the energy demand of a community. The accuracy of both models was similar, with the physics-based presenting a percentage error of 6.26% and the statistical 5.94%. Nouvel et al. [45] compared a quasi-steady state heat balance model with a linear regression approach. At the zipcode level, they found that the physics-based and the statistical approaches had comparable performance with the model using statistical computational method performing slightly better (MAPE = 26%). Nageler et al. [57] compared the performance of a sigmoid energy signature statistical approach with the dynamic building simulation tool IDA ICE, a model using physics-based computational method. They too found similar performances between the two approaches with the statistical performing slightly better (R^2 = 0.97) than the physics-based (R^2 = 0.92). However, they note that the physics-based approach has the additional ability to perform what-if scenarios to evaluate different desired situations for example with energy efficiency retrofit analysis.

From the literature review and the three studies described above, it appears that physics-based and statistical computational methods can achieve similar levels of accuracy at the annual aggregate and individual building level scales. The choice of the approach is most likely dependent on the intended application, as in the literature the majority of studies with energy efficiency retrofit analysis as the primary intended application used physics-based approaches, while those whose intended application was energy demand quantification, more often used statistical computational methods.

In practice, a wide variety of input data and computational methods have been explored in UBEM. However, there is no evidence to show that physics-based, statistical, or hybrid methods have led to more accurate models. Further, within each computational method, there is no evidence that the amount or type of input data leads to more accurate models. This can be attributed to the limited amount of studies that have undertaken validation in UBEM, in addition to the inconsistency of accuracy reporting practices in UBEM as well as the nature of the application these models are aimed for.

3.2.4. Calibration and validation methods

The increased interest of stakeholders now pointing at UBEM for a variety of applications, has enhanced the need for rigour in methodologies. Calibration and validation are both key parts of any modelling process. The scope of this work was to explore the accuracy in UBEM and therefore all the reviewed studies have undergone some form of validation. However, remarkably, almost half of the studies did not present any evidence of calibration. This was particularly profound among UBEM studies that used physics-based computational methods (and especially those employing dynamic thermal simulation), where less than 50% of the models examined included evidence of having been calibrated.

In the UBEM studies that used physics-based computational methods and provided evidence of calibration, a common approach was the Bayesian calibration. This method allows capturing the diversity of the input data by characterising each parameter undergoing calibration as a probability distribution instead of a single value. In one of the first studies to apply the Bayesian calibration in UBEM, Booth et al. [40] demonstrated a significant reduction in the percentage error of their model, from 0.176 to 0.005%, thus already low due to the aggregate spatial resolution it was measured in. The results published by Sokol et al. [51] also exhibit a significant reduction in the model's percentage error after applying the Bayesian calibration, from 69% for the non-calibrated model to 44% after calibration, using data on a monthly temporal resolution. The Bayesian calibration has become increasingly popular in recent years as researchers become more familiar with the technique [60,72,76]. In calibrated models employing physics-based approaches, where Bayesian calibration was not used, researchers developed custom approaches, often adjusting a small number of model parameters (i.e. infiltration rates, occupancy profiles) for a certain number of buildings in their datasets [57,67,79]. The validation process, in UBEMs using physics-based computational methods, is, in most cases, the final stage of the study, where the predicted outputs of the developed model are compared to different datasets than those used for calibrating the model, hence the external validation is applied to assess their accuracy. Here, the validation can be exercised on either a single dataset [44] or multiple [56].

In UBEMs making use of statistical computational methods, calibration is considered the training of the model, which is often an inherent part of the validation process that employs a training and a testing procedure. Among studies that used statistical computational methods, the internal validation was the most commonly used. In the internal validation method, the dataset is divided into training and testing parts and there are three subcategories: split, cross and bootstrap [33]. Split internal validation was introduced early on by Reiter [4], where the dataset was divided to an "evaluation" period of four months (model training) and a "prediction" period of three months (model testing). More recently, Gassar et al. [71] and Xu et al. [69] applied the commonly used ratios of 80/20 and 70/30 for model training/testing, respectively. Cross internal validation is regarded as more robust, since the model is tested several times using different parts of the data, with the ten-fold cross validation method being regularly preferred [48,65]. Bootstrap internal validation, despite being considered highly efficient and the most vigorous from a statistical point of view [83], has yet to be applied widely in statistical computational methods in UBEM with Mastrucci et al. [42] introducing it to the field.

Overall, with respect to calibration, the evidence in Table 2 show that a large portion of the studies that did not specify any form of calibration and analysed the accuracy of the models at the single building spatial resolution were morel likely to present large percentage errors (Quan et al. reported ±800% [44], Osterbring et al. found up to 100% [47], Wang et al. noted almost 1000% [61], Katal et al. recorded up to more than 100% [66]). For the studies that did report some form of calibration the data are not conclusive as to which method consistently provides higher accuracy. However, Bayesian calibration in physics-based approaches, has been reported to improve the accuracy of the models up to 40% [51] or even 90% [40], with two further studies having reported the lowest CVRSME values at the aggregate and hourly spatial and temporal resolutions, with mean values of 5.6% and 7.8% [60,72]. As for validation, it has been at the core of this work, therefore all the reviewed models are validated, with the majority applying external (34 models) and internal (14 models) validation. Internal validation is mainly utilised for statistical computational methods where the accuracy is not reported in the single building spatial resolution but rather in aggregate, with varying R^2 values at daily (R^2 = 0.7) and annual temporal resolutions (R^2 = 0.99) [65,71]. External validation is prevalent mainly across physics-based computational methods, hence the reported accuracy is also found at the single building spatial resolution regularly, evidently with exceedingly large variations in values of R^2, from as low as 0.1 [50] to as high as 0.96 [58].

Overall, this work has showed that there is no single key attribute that governs the accuracy of UEBM but rather the collective deficiencies across all attributes. All contextual attributes (Application, Scale, Input data) calculation techniques (Computational methods) and assessment techniques (Calibration, Validation approaches) can have significant impact in the accuracy of UBEM.

Table 3
ASHRAE Guideline 14-2002 Values [80] in comparison to values achieved in UBEM practice at aggregate, single building, monthly and hourly spatial and temporal resolutions.

	Aggregate		Single Building	
	Monthly	Hourly	Monthly	Hourly
ASHRAE Guideline 22 values				
CVRMSE (%)	–	–	<15	<30
NMBE (%)	–	–	<±5	<±10
UBEM achieved values				
CVRMSE (%)	10–20	5–40	3–50	30–120
NMBE (%)	–	−3 to +9	−15 to +4	−40 to +120

3.3. Summary of UBEM best practice

When analysing accuracy, it is essential to apply suitable quantitative approaches to investigate the highest possible standards, so that researchers and all interested stakeholders have the required confidence in UBEM that will allow it to become suitable for the vast range of applications it can cover. This review however, shows that the majority of UBEMs are not validated against measured data (<10% of identified papers) but as more data become available, more researchers are attempting to validate their models, as evidenced by the majority of validated models being reported in the last six years. Those that have reported some metric of accuracy, have used a wide variety of measures of error, making cross model comparisons difficult.

The ASHRAE Guideline 14-2002 [80], although designed for a specific use case at the building scale, has been increasingly referenced in UBEM. It is worth noting that a number of studies exploring the calibration of individual buildings [84–87] have demonstrated the capability of models to produce much lower values of CVRMSE (monthly: 1%–6%, hourly: 2%–9%) and NMBE (monthly: 1%, hourly: 0%–3%) than those recommended in the ASHRAE Guideline 14-2002. In contrast, when evaluating UBEM accuracy at the individual building scale, the ranges of hourly CVRMSE reported for a large portion of buildings are far larger than the 30% recommended in the ASHRAE guideline. This discrepancy is to be expected as there has not been any consideration of appropriate metrics yet for the spatial element inherent in UBEM. This spatial element, often represented by the number of buildings modelled, introduces more variability than is experienced at the individual building scale, due to the varied nature of building uses across a building stock.

In this work, Table 2 reports the measures of error CVRMSE and NMBE, in the aggregate and individual building spatial resolutions (in both monthly and hourly temporal resolutions), based on the findings of the UBEM studies that comprise this review. Table 3 outlines the CVRMSE and NMBE values given in the ASHRAE Guideline 14-2002 and the range of values reported in the UBEM studies that have consistently reported these measures of error at the aggregate and single building spatial resolutions, hence allowing comparison between the ASHRAE recommendations and what is found in practice. It has to be noted that the ranges presented at the individual building spatial resolution in Table 3 are not from a single study but rather represent the lowest and highest individual building errors reported in the reviewed literature.

Whilst we have attempted to distill the learnings on UBEM accuracy in Table 3, the accuracy of UBEM is rather complex and multidimensional and the enumerate spatial and temporal resolutions possible are still not fully captured. Here we have reported on the modelling practice but the required accuracy of any model depends on what is required in each specific application. In UBEM, there exists a range of applications and therefore potentially a range of acceptable accuracies. For energy efficiency retrofit analysis, reporting accurate values at the annual aggregate resolutions may be sufficient, as the aim is to determine the effect of a large set of energy efficiency measures. Although a single percent accuracy value hardly provides confidence that the underlying phenomena of energy efficiency measures are being captured. Energy efficiency retrofit analysis, however, contrasts greatly from an application in energy systems integration, where one may attempt to determine how much energy demand could be offset by photovoltaic production. In this case, being confident in the time series behaviour in aggregate and potentially at intermediate local regions is important as well. Further still is the application of climate resilience, where the aim is to understand the concurrent change in energy consumption and internal temperatures. In this context, a main metric could be the number of summertime overheating hours and/or days with peak energy demand over a certain amount. These types of metrics are still not yet reported in validated UBEM studies.

4. Future research recommendations

This work has identified clear gaps in knowledge when considering the accuracy in the field of UBEM. The reflection upon those can inform the changes needed in the modelling practice to drive the field forward. For example, it is essential to be able to compare calibration methods to further assess the impact of Bayesian approaches on accuracy, or compare different computational methods in relation to specific applications. Currently, the inconsistencies in reporting practice with respect to spatial and temporal resolutions and to the measures of error used, are creating difficulties in comparing the accuracy between studies. Therefore, it is essential for future research to strive for a number of changes to the modelling practice.

First, modellers should establish the transparent consideration of the key attributes identified in this review and systematically report the accuracy using a variety of common metrics at several of the spatial and temporal resolutions identified in this paper, to allow for clear comparison between studies. Furthermore, it is essential to formulate a stronger connection between the accuracy needed for the application and the modelling approach, as often methodologies are rather disconnected from the application, making it uncertain if the accuracy achieved is sufficient. The accuracy of UBEM should be evaluated in context with the application and therefore the use of more dynamic metrics such as the magnitude and timing of peak load that explicitly evaluate aspects of the dynamic response, need to be introduced. Developing UBEMs requires a significant amount of work that can take a substantial amount of time to complete and addressing all these issues can be challenging for individual research teams. Therefore, it would be best for the UBEM community to work together to move the field forward. To that end it is recommended to have common data sets that all researchers could use to evaluate their approaches, making comparisons considerably less challenging. This would require the field to adopt an open data approach and further make all underlying data publicly available.

5. Conclusion

Over the past years, there has been a considerable increase in the research interest in the field of UBEM. The need to mitigate the impact of the climate crisis and the continuous advances in computing technology have driven the field forward, resulting in a plethora of studies. This review examined studies that present validated models in UBEM. Of primary interest in this work, was the investigation of the accuracy in UBEM, through the systematic analysis of what are considered to be key attributes in UBEM: application, scale, input data, computational method, calibration method and validation method. For this purpose, 47 studies that reported the error of models against measured data were analysed using a singular taxonomy comprising of these key attributes.

The findings showed that the accuracy in UBEM is complex and multidimensional, considered at a variety of temporal resolutions, spatial resolutions and measures of error. Hence, careful synthesis and consideration is required when drawing conclusions, as there is no single key attribute that governs accuracy, rather the collective deficiencies across all attributes. At the annual and aggregate resolutions, percentage errors of 1% are achievable, however the spread of errors at the individual building resolution can get exceedingly increased, even up to 1000%. This review has presented the best practice of what has been achievable to date in terms of accuracy (NMBE values for single buildings in the hourly temporal resolution of −40% to +120%) and compared it to widely used recommendations (<±10) from a recognised professional body, such as the ASHRAE. With respect to factors that influence the accuracy of UBEM, it has shown that the use of statistical or physics-based computational methods can lead to similar ranges in terms of accuracy and therefore the choice of the approach to be used should depend principally on the application (e.g. energy efficiency retrofit analysis, energy systems integration, etc.). The results further showed that models using non-calibrated physics-based computational methods were more likely to report overly large errors, while those employing Bayesian calibration consistently reported lower errors, demonstrating the positive impact of calibration and in particular the Bayesian approach, on the accuracy of UBEM.

Overall, it is clear that the field of UBEM is not yet established, as there is no standardised way of conducting research and this can have a profound impact on the reported accuracy. There is no consistent reporting practice with respect to spatial and temporal resolutions and also with regards to reported measures of error, making comparison between studies difficult.

It is recommended for the field to establish the transparent consideration of the key attributes identified in this review, and systematically report the accuracy of UBEM concurrently at multiple spatial and temporal resolutions, while applying a variety of suitable measures of error, in order to avoid masking any inabilities of the developed models in relation to their application. Likewise, it is recommended that researchers in the UBEM field develop and consider additional measures of error, beyond those textbook statistics reported throughout the reviewed studies, able to capture dynamic behaviour, such as peak hourly demand or maximum daily temperature, as this is vital in understanding the dynamics of building and model performance. This is essential especially for the emerging applications of energy systems integration and climate resilience, so as to enable UBEM to be used as a tool for developing healthy, energy efficient and decarbonised cities.

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.

Acknowledgement

This research was made possible by the support from the Engineering and Physical Sciences Research Council (EPSRC), United Kingdom for the FlexTECC: Flexible Timing of Energy Consumption in Communities Innovation Fellowship (grant EP/S001670/1).

References

[1] Edelman DJ. Recent energy models: a review of methodologies. J Environ Syst 1977;7(3). http://dx.doi.org/10.2190/3J5N-U7TN-LN61-E2E6.

[2] Charpentier JP. A review of energy models. Tech. rep. 2, Laxenburg, Austria: IIASA Research Report; 1975.

[3] Sinden FW. A two-thirds reduction in the space heat requirement of a twin rivers townhouse. Energy Build 1977;1:243–60.

[4] Reiter ER. Energy consumption modelling. In *National oceanic and atmospheric administration (noaa) workshop*, Columbia, MO, 1980.

[5] Clarke JA. Environmental systems performance (Ph.D. thesis), University of Strathclyde; 1977.

[6] Lomas KJ, Eppel H, Martin CJ, Bloomfield DP. Empirical validation of building energy simulation programs. Energy Build 1997;26(3):253–76.

[7] Crawley DB, Lawrie LK, Winkelmann FC, Buhl WF, Huang YJ, Pedersen CO, Strand RK, Liesen RJ, Fisher DE, Witte MJ, Glazer J. EnergyPlus: Creating a new-generation building energy simulation program. Energy Build 2001;33(4):319–31.

[8] Swan LG, Ugursal VI. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. Renew Sustain Energy Rev 2009;13(8):1819–35. http://dx.doi.org/10.1016/j.rser.2008.09.033.

[9] Kavgic M, Mavrogianni A, Mumovic D, Summerfield A, Stevanovic Z, Djurovic-Petrovic M. A review of bottom-up building stock models for energy consumption in the residential sector. Build Environ 2010;45(7):1683–97. http://dx.doi.org/10.1016/j.buildenv.2010.01.021.

[10] Allegrini J, Orehounig K, Mavromatidis G, Ruesch F, Dorer V, Evins R. A review of modelling approaches and tools for the simulation of district-scale energy systems. Renew Sustain Energy Rev 2015;52:1391–404. http://dx.doi.org/10.1016/j.rser.2015.07.123.

[11] Frayssinet L, Merlier L, Kuznik F, Hubert JL, Milliez M, Roux JJ. Modeling the heating and cooling energy demand of urban buildings at city scale. Renew Sustain Energy Rev 2018;81:2318–27. http://dx.doi.org/10.1016/j.rser.2017.06.040.

[12] Brøgger M, Wittchen KB. Estimating the energy-saving potential in national building stocks – a methodology review. Renew Sustain Energy Rev 2018;82(July 2017):1489–96. http://dx.doi.org/10.1016/j.rser.2017.05.239.

[13] Ferrando M, Causone F, Hong T, Chen Y. Urban building energy modeling (UBEM) tools : A state-of-the-art review of bottom-up physics-based approaches. Sustainable Cities Soc 2020;62(March):102408. http://dx.doi.org/10.1016/j.scs.2020.102408.

[14] Ahmad T, Chen H, Guo Y, Wang J. A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. Energy Build 2018;165:301–20. http://dx.doi.org/10.1016/j.enbuild.2018.01.017.

[15] Ferrari S, Zagarella F, Caputo P, Bonomolo M. Assessment of tools for urban energy planning. Energy 2019;176:544–51. http://dx.doi.org/10.1016/j.energy.2019.04.054.

[16] Abdullah A, Gassar A, Cha SH. Energy prediction techniques for large-scale buildings towards a sustainable built environment : A review. Energy Build 2020;224:110238. http://dx.doi.org/10.1016/j.enbuild.2020.110238.

[17] Fathi S, Srinivasan R, Fenner A, Fathi S. Machine learning applications in urban building energy performance forecasting : A systematic review. Renew Sustain Energy Rev 2020;133(August):110287. http://dx.doi.org/10.1016/j.rser.2020.110287.

[18] Keirstead J, Jennings M, Sivakumar A. A review of urban energy system models: Approaches, challenges and opportunities. Renew Sustain Energy Rev 2012;16(6):3847–66. http://dx.doi.org/10.1016/j.rser.2012.02.047.

[19] Hong T, Chen Y, Luo X, Luo N, Lee SH. Ten questions on urban building energy modeling. Build Environ 2020;168(October 2019):106508. http://dx.doi.org/10.1016/j.buildenv.2019.106508.

[20] Johari F, Peronato G, Sadeghian P, Zhao X, Wid J. Urban building energy modeling : State of the art and future prospects. In: 128.September 2019, 2020. http://dx.doi.org/10.1016/j.rser.2020.109902.

[21] Li W, Zhou Y, Cetin K, Eom J, Wang Y, Chen G, Zhang X. Modeling urban building energy use: A review of modeling approaches and procedures. Energy 2017;141:2445–57. http://dx.doi.org/10.1016/j.energy.2017.11.071.

[22] Abbasabadi N, Mehdi Ashayeri JK. Urban energy use modeling methods and tools: A review and an outlook. Build Environ 2019;161(July):106270. http://dx.doi.org/10.1016/j.buildenv.2019.106270.

[23] Larbi Chalal M, Benachir M, White M, Shrahily R. Energy planning and forecasting approaches for supporting physical improvement strategies in the building sector: A review. Renew Sustain Energy Rev 2016;64:761–76. http://dx.doi.org/10.1016/j.rser.2016.06.040.

[24] Sousa G, Jones BM, Mirzaei PA, Robinson D. A review and critique of UK housing stock energy models, modelling approaches and data sources. Energy Build 2017;151:66–80. http://dx.doi.org/10.1016/j.enbuild.2017.06.043.

[25] Reinhart CF, Cerezo Davila C. Urban building energy modeling - a review of a nascent field. Build Environ 2016;97:196–202. http://dx.doi.org/10.1016/j.buildenv.2015.12.001.

[26] Ang YQ, Berzolla ZM, Reinhart CF. From concept to application : A review of use cases in urban building energy modeling. Appl Energy 2020;279:115738. http://dx.doi.org/10.1016/j.apenergy.2020.115738.

[27] Scopus. 2020, URL: https://www.scopus.com/ (visited on 09/01/2020).

[28] Google scholar. 2020, URL: https://scholar.google.com/ (visited on 09/01/2020).

[29] Science direct. 2020, URL: https://www.sciencedirect.com/ (visited on 09/01/2020).

[30] Blalock HM. Social statistics. International edition, New York: McGraw-Hill; 1979, URL: https://books.google.gr/books?id=JjiCAAAAIAAJ.

[31] Reddy TA. Literature review on calibration of building energy simulation programs: Uses, problems, procedures, uncertainty, and tools. ASHRAE Trans 2006;112(4844):226–40.

[32] Efron B. How biased is the apparent error rate of a prediction rule ? J Amer Statist Assoc 1986;81(394):461–70.

[33] Steyerberg EW, Harrell, Jr., FE, Borsboom GJJM, Eijkemans MJCR, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J. Clin Epidemiol 2001;54:774–81.

[34] Shimoda Y, Asahi T, Taniguchi A, Mizuno M. Evaluation of city-scale impact of residential energy conservation measures using the detailed end-use simulation model. Energy 2007;32(9):1617–33. http://dx.doi.org/10.1016/j.energy.2007.01.007.

[35] Strzalka A, Bogdahn J, Coors V, Eicker U. 3D city modeling for urban scale heating energy demand forecasting. HVAC R Res 2011;17(4):526–39. http://dx.doi.org/10.1080/10789669.2011.582920.

[36] Aranda A, Ferreira G, Mainar-Toledo MD, Scarpellini S, Sastresa EL. Multiple regression models to predict the annual energy consumption in the spanish banking sector. Energy Build 2012;49:380–7. http://dx.doi.org/10.1016/j.enbuild.2012.02.040.

[37] Lorimer S. A housing stock model of non-heating end-use energy in England verified by aggregate energy use data. Energy Policy 2012;50:419–27. http://dx.doi.org/10.1016/j.enpol.2012.07.037.

[38] Ren Z, Paevere P, Mcnamara C. A local-community-level, physically-based model of end-use energy consumption by Australian housing stock. Energy Policy 2012;49:586–96. http://dx.doi.org/10.1016/j.enpol.2012.06.065.

[39] Howard B, Parshall L, Thompson J, Hammer S, Dickinson J, Modi V. Spatial distribution of urban building energy consumption by end use. Energy Build 2012;45:141–51. http://dx.doi.org/10.1016/j.enbuild.2011.10.061.

[40] Booth AT, Choudhary R, Spiegelhalter DJ. Handling uncertainty in housing stock models. In: 48.2012 (2020). http://dx.doi.org/10.1016/j.buildenv.2011.08.016.

[41] Filogamo L, Peri G, Rizzo G, Giaccone A. On the classification of large residential buildings stocks by sample typologies for energy planning purposes. Appl Energy 2014;135:825–35. http://dx.doi.org/10.1016/j.apenergy.2014.04.002.

[42] Mastrucci A, Baume O, Stazi F, Leopold U. Estimating energy savings for the residential building stock of an entire city: A GIS-based statistical downscaling approach applied to rotterdam. Energy Build 2014;75:358–67. http://dx.doi.org/10.1016/j.enbuild.2014.02.032.

[43] Fonseca JA, Schlueter A. Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. Appl Energy 2015;142:247–65. http://dx.doi.org/10.1016/j.aqpro.2013.07.003, URL: http://www.sciencedirect.com/science/article/pii/S0306261914013257 arXiv:arXiv:1011.1669v3.

[44] Quan SJ, Li Q, Augenbroe G, Brown J. Urban data and building energy modeling : A GIS-based urban building energy modeling system using the urban-EPC engine. In: et Al SG, editor. Planning support systems and smart cities. Switzerland: Springer International Publishing; 2015, http://dx.doi.org/10.1007/978-3-319-18368-8, September.

[45] Nouvel R, Mastrucci A, Leopold U, Baume O, Coors V, Eicker U. Combining GIS-based statistical and engineering urban heat consumption models: Towards a new framework for multi-scale policy support. Energy Build 2015;107:204–12. http://dx.doi.org/10.1016/j.enbuild.2015.08.021.

[46] Cerezo Davila C, Reinhart CF, Bemis JL. Modeling boston: A workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets. Energy 2016;117:237–50. http://dx.doi.org/10.1016/j.energy.2016.10.057.

[47] Österbring M, Mata E, Thuvander L, Mangold M, Johnsson F, Wallbaum H. A differentiated description of building-stocks for a georeferenced urban bottom-up building-stock model. Energy Build 2016;120:78–84. http://dx.doi.org/10.1016/j.enbuild.2016.03.060.

[48] Ma J, Cheng JCP. Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology. Appl Energy 2016;183:182–92. http://dx.doi.org/10.1016/j.apenergy.2016.08.079.

[49] Nageler P, Zahrer G, Heimrath R, Mach T, Mauthner F, Leusbrock I, Schranzhofer H, Hochenauer C. Novel validated method for GIS based automated dynamic urban building energy simulations. Energy 2017;139:142–54. http://dx.doi.org/10.1016/j.energy.2017.07.151.

[50] Buffat R, Froemelt A, Heeren N, Raubal M, Hellweg S. Big data GIS analysis for novel approaches in building stock modelling. Appl Energy 2017;208(November):277–90. http://dx.doi.org/10.1016/j.apenergy.2017.10.041.

[51] Sokol J, Cerezo Davilla C, Reinhart CF. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. Energy Build 2017;134:11–24. http://dx.doi.org/10.1016/j.enbuild.2016.10.050.

[52] Nouvel R, Zirak M, Coors V, Eicker U. The influence of data quality on urban heating demand modeling using 3D city models. Comput Environ Urban Syst 2017;64:68–80. http://dx.doi.org/10.1016/j.compenvurbsys.2016.12.005.

[53] Olivo Y, Hamidi A, Ramamurthy P. Spatiotemporal variability in building energy use in new york city. Energy 2017;141:1393–401. http://dx.doi.org/10.1016/j.energy.2017.11.066.

[54] Kontokosta CE, Tull C. A data-driven predictive model of city-scale energy use in buildings. Appl Energy 2017;197:303–17. http://dx.doi.org/10.1016/j.apenergy.2017.04.005.

[55] Alhamwi A, Medjroubi W, Vogt T, Agert C. Modelling urban energy requirements using open source data and models. Appl Energy 2018;231(October):1100–8. http://dx.doi.org/10.1016/j.apenergy.2018.09.164.

[56] Kristensen MH, Brun A, Petersen S. Predicting danish residential heating energy use from publicly available building characteristics. Energy Build 2018;173:28–37. http://dx.doi.org/10.1016/j.enbuild.2018.05.011.

[57] Nageler P, Koch A, Mauthner F, Leusbrock I, Mach T, Hochenauer C, Heimrath R. Energy & buildings comparison of dynamic urban building energy models (UBEM): Sigmoid energy signature and physical modelling approach. Energy Build 2018;179:333–43. http://dx.doi.org/10.1016/j.enbuild.2018.09.034.

[58] Nagpal S, Reinhart C. A comparison of two modeling approaches for establishing and implementing energy use reduction targets for a university campus. Energy Build 2018;173:103–16. http://dx.doi.org/10.1016/j.enbuild.2018.05.035.

[59] Zhang R, Mirzaei P, Jones B. Development of a dynamic external CFD and BES coupling framework for application of urban neighbourhoods energy modelling. Build Environ 2018;146(July):37–49. http://dx.doi.org/10.1016/j.buildenv.2018.09.006.

[60] Kristensen MH, Hedegaard RE, Petersen S. Hierarchical calibration of archetypes for urban building energy modeling. Energy Build 2018;175:219–34. http://dx.doi.org/10.1016/j.enbuild.2018.07.030.

[61] Wang D, Landolt J, Mavromatidis G, Orehounig K, Carmeliet J. CESAR: A Bottom-up building stock modelling tool for Switzerland to address sustainable energy transformation strategies. Energy Build 2018;169:9–26. http://dx.doi.org/10.1016/j.enbuild.2018.03.020.

[62] Oliveira Panão MJN, Brito MC. Modelling aggregate hourly electricity consumption based on bottom-up building stock. Energy Build 2018;170:170–82. http://dx.doi.org/10.1016/j.enbuild.2018.04.010.

[63] Nutkiewicz A, Yang Z, Jain RK. Data-driven urban energy simulation (DUEs): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow. Appl Energy 2018;225(June):1176–89. http://dx.doi.org/10.1016/j.apenergy.2018.05.023.

[64] Koschwitz D, Frisch J, Treeck CV. Data-driven heating and cooling load predictions for non-residential buildings based on support vector machine regression and NARX recurrent neural network: A comparative study on district scale. Energy 2018;165:134–42. http://dx.doi.org/10.1016/j.energy.2018.09.068.

[65] Torabi Moghadam S, Toniolo J, Mutani G, Lombardi P. A GIS-statistical approach for assessing built environment energy use at urban scale. Sustainable Cities Soc 2018;37(September 2017):70–84. http://dx.doi.org/10.1016/j.scs.2017.10.002.

[66] Katal A, Mortezazadeh M, Wang LL. Modeling building resilience against extreme weather by integrated cityffd and citybem simulations. Appl Energy 2019;250(March):1402–17. http://dx.doi.org/10.1016/j.apenergy.2019.04.192.

[67] Krayem A, Al Bitar A, Ahmad A, Faour G, Gastellu-etchegorry JP, Lakkis I, Gerard J, Zaraket H, Yeretzian A, Najem S. Urban energy modeling and calibration of a coastal mediterranean city: The case of beirut. Energy Build 2019;199:223–34. http://dx.doi.org/10.1016/j.enbuild.2019.06.050.

[68] Kim B, Yamaguchi Y, Kimura S, Ko Y, Ikeda K, Shimoda Y. Urban building energy modeling considering the heterogeneity of hvac system stock: A case study on Japanese office building stock. Energy Build 2019;199:547–61. http://dx.doi.org/10.1016/j.enbuild.2019.07.022.

[69] Xu X, Wang W, Hong T, Chen J. Incorporating machine learning with building network analysis to predict multi-building energy use. Energy Build 2019;186:80–97. http://dx.doi.org/10.1016/j.enbuild.2019.01.002.

[70] Yi CY, Peng C. An archetype-in-neighbourhood framework for modelling cooling energy demand of a city's housing stock. Energy Build 2019;196:30–45. http://dx.doi.org/10.1016/j.enbuild.2019.05.015.

[71] Gassar AAA, Yun YG, Kim S. Data-driven approach to prediction of residential energy consumption at urban scales in London. Energy 2019;187:115973. http://dx.doi.org/10.1016/j.energy.2019.115973.

[72] Hedegaard RE, Kristensen MH, Pedersen TH, Brun A, Petersen S. Bottom-up modelling methodology for urban-scale analysis of residential space heating demand response. Appl Energy 2019;242(March):181–204. http://dx.doi.org/10.1016/j.apenergy.2019.03.063.

[73] Krarti M, Aldubyan M, Williams E. Residential building stock model for evaluating energy retrofit programs in Saudi Arabia. Energy 2020;195:116980. http://dx.doi.org/10.1016/j.energy.2020.116980.

[74] Streltsov A, Malof JM, Huang B, Bradbury K. Estimating residential building energy consumption using overhead imagery. Appl Energy 2020;280(October):116018. http://dx.doi.org/10.1016/j.apenergy.2020.116018.

[75] Jahani E, Cetin K, Ho I. City-scale single family residential building energy consumption prediction using genetic algorithm-based numerical moment matching technique. Build Environ 2020;172(October 2019):106667. http://dx.doi.org/10.1016/j.buildenv.2020.106667.

[76] Tardioli G, Narayan A, Kerrigan R, Oates M, Donnell JO, Finn DP. A methodology for calibration of building energy models at district scale using clustering and surrogate techniques. Energy & Buildings 2020;226:110309. http://dx.doi.org/10.1016/j.enbuild.2020.110309.

[77] Roth J, Martin A, Miller C, Jain RK. SynCity : USing open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods nomenclature :. Appl Energy 2020;280(October):115981. http://dx.doi.org/10.1016/j.apenergy.2020.115981.

[78] Yang X, Hu M, Heeren N, Zhang C, Verhagen T, Tukker A, Steubing B. A combined GIS-archetype approach to model residential space heating energy : A case study for the netherlands including validation. Appl Energy 2020;280(October):115953. http://dx.doi.org/10.1016/j.apenergy.2020.115953.

[79] Fernandez J, Portillo L, Flores I. A novel residential heating consumption characterisation approach at city level from available public data : Description and case study. Energy & Buildings 2020;221:110082. http://dx.doi.org/10.1016/j.enbuild.2020.110082.

[80] ASHRAE. Guideline 14-2002, measurement of energy and demand savings. Tech. rep., Atlanta, GA: American Society of Heating, Ventilating, and Air Conditioning Engineers (ASHRAE); 2002.

[81] Zucker G, Judex F, Blochle M, Kostl M, Widl E, Hauer S, Bres A, Zeilinger J. A new method for optimizing operation of large neighborhoods of buildings using thermal simulation. Energy Build 2016;125:153–60. http://dx.doi.org/10.1016/j.enbuild.2016.04.081.

[82] Schiefelbein J, Rudnick J, Scholl A, Remmen P, Fuchs M, Müller D. Automated urban energy system modeling and thermal building simulation based on OpenStreetMap data sets. Build Environ 2019;149(July 2018):630–9. http://dx.doi.org/10.1016/j.buildenv.2018.12.025.

[83] Efron B, Tibshirani RJ. An introduction to the bootstrap: monographs on statistics and applied probability. New York and London: Chapman and Hall/CRC; 1993.

[84] Raftery P, Keane M, Costa A. Calibrating whole building energy models: Detailed case study using hourly measured data. Energy Build 2011;43(12):3666–79. http://dx.doi.org/10.1016/j.enbuild.2011.09.039.

[85] Parker J, Cropper P, Shao L. A calibrated whole building simulation approach to assessing retrofit options for Birmingham airport. In *First building simulation and optimization conference*, September, Loughborough, 2012, p. 49–56.

[86] Chong A, Poh Lam K, Pozzi M, Yang J. BayesIan calibration of building energy models with large datasets. Energy Build 2017;154:343–55. http://dx.doi.org/10.1016/j.enbuild.2017.08.069.

[87] Tüysüz F, Sözer H. Calibrating the building energy model with the short term monitored data a case study of a large-scale residential building. Energy Build 2020;224:1–13. http://dx.doi.org/10.1016/j.enbuild.2020.110207.