Calibrating building energy simulation models: A review of the basics to guide future work

Article *in* Energy and Buildings · October 2021

DOI: 10.1016/j.enbuild.2021.111533

3 authors:

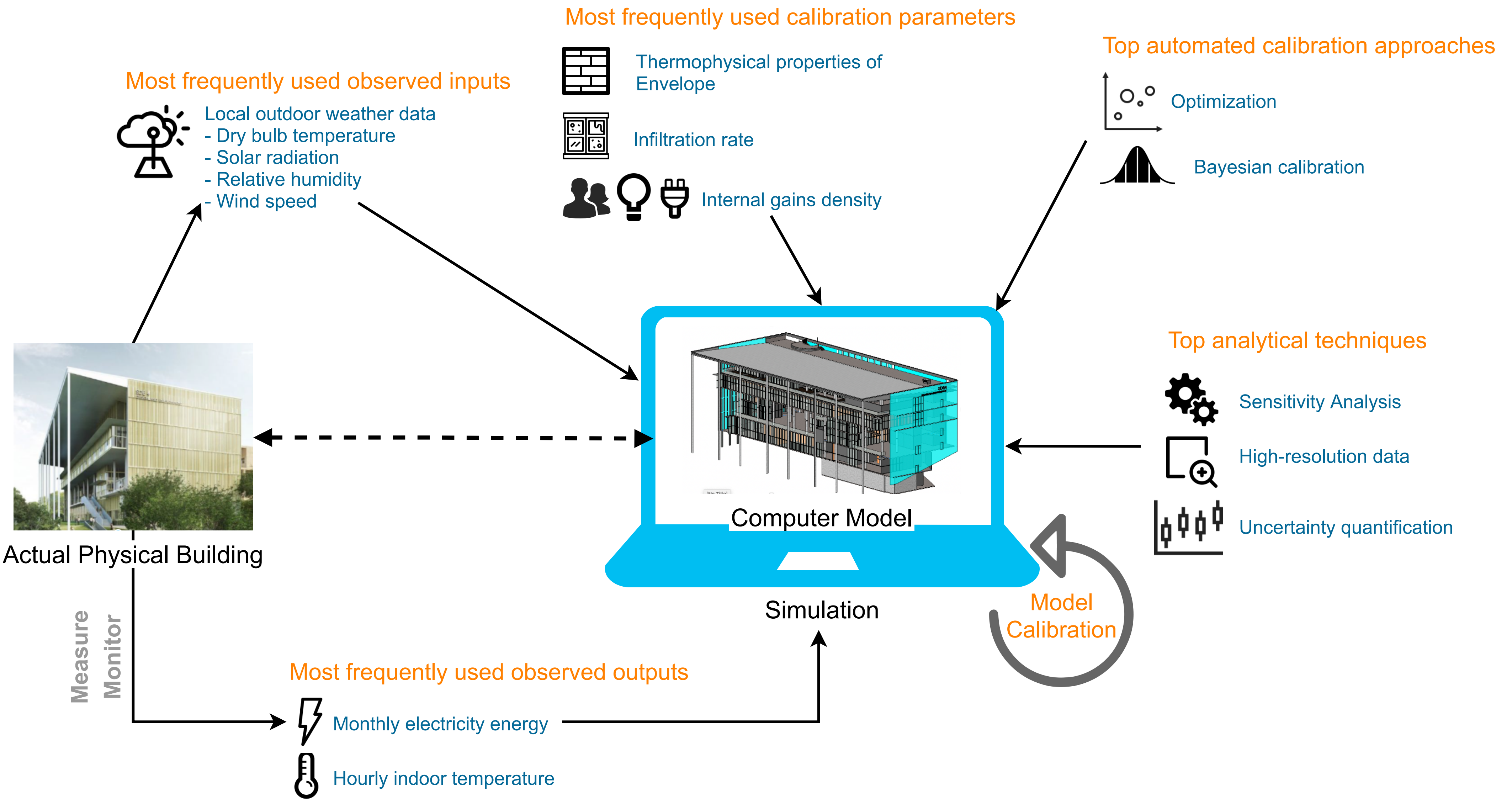
Adrian Chong
National University of Singapore

Yaonan Gu
National University of Singapore

Hongyuan Jia
Berkeley Education Alliance for Research in Singapore

Graphical Abstract

A review of the basics of model calibration to guide future work



Chong, A., Gu, Y., & Jia, H. (2021). Calibrating building energy simulation models: A review of the basics to guide future work. Energy and Buildings, 111533.
doi: https://doi.org/10.1016/j.enbuild.2021.111533

Calibrating building energy simulation models: A review of the basics to guide future work

Adrian Chong[a,*], Yaonan Gu[a], Hongyuan Jia[b,c]

[a] *Department of the Built Environment, School of Design and Environment, National University of Singapore, 4 Architecture Drive, Singapore 117566, Singapore*
[b] *School of Civil Engineering and Architecture, Chongqing University of Science and Technology, Chongqing 401331, China*
[c] *SinBerBEST Program, Berkeley Education Alliance for Research in Singapore, Singapore, 138602, Singapore*

Abstract

Building energy simulation (BES) plays a significant role in buildings with applications such as architectural design, retrofit analysis, and optimizing building operation and controls. There is a recognized need for model calibration to improve the simulations' credibility, especially with building data becoming increasingly available and the promises that a digital twin brings. However, BES calibration remains challenging due to the lack of clear guidelines and best practices. This study aims to provide the foundation for future research through a detailed systematic review of the vital aspects of BES calibration. Specifically, we conducted a meta-analysis and categorization of the simulation inputs and outputs, data type and resolution, key calibration methods, and calibration performance evaluation. This study also identified reproducible simulations as a critical issue and proposes an incremental approach to encourage future research's reproducibility.

Keywords: Building performance simulation, Building simulation, calibration, reproducibility, Optimization, Uncertainty

---

*Corresponding authors
Email address: adrian.chong@nus.edu.sg (Adrian Chong)

1. Introduction

1.1. Calibration in building energy simulation

Building energy simulation (BES) can broadly be defined as a physics-based mathematical model that allows the detailed calculation of a building's energy performance and occupant thermal comfort under the influence of various inputs such as weather, building geometry, internal loads, HVAC systems, operational schedules, and simulation specific parameters. Originally intended for use during the design phase, BES is increasingly being used throughout a building's lifecycle [1]. However, there are increasing concerns about the model's credibility within the building industry as significant discrepancies between simulated and measured energy use become more apparent with the rapid deployment of smart energy meters and the internet of things (IoT) [2]. For instance, Turner and Frankel [3] analyzed 121 LEED buildings and found that measured energy use can be between 0.5 to 2.75 times the predicted energy use. Mantesi et al. [4] showed that default settings and methods of modeling thermal mass can result in up to 26% divergence in the simulation predictions.

Previous studies [5, 6] observed that the main causes of discrepancies between predicted and actual energy performance stem from: (a) specification uncertainty arising from assumptions due to a lack of information; (b) model inadequacy arising from simplifications and abstractions of actual physical building systems; (c) operational uncertainty arising from a lack of feedback regarding actual use and operation of buildings; and (d) scenario uncertainty arising from specifying model conditions such as weather conditions and building occupancy. Consequently, model calibration is often undertaken to match simulation predictions to actual observations better and increase the model's credibility for making predictions. Although calibration is not essential for BES research, it is becoming increasingly important for establishing model credibility. The International Energy Agency's Energy in Buildings and Communities (IEA-EBC) Annex 53 also reported the significance of the development and application of model calibration and uncertainty analysis for BES [7].

The typical reason for BES calibration is to more confidently predict using simulation. Although predictions may be wrong, they can still be useful. Trucano et al. [8] list examples of such predictions that are also relevant for BES, and they include:

- Simulating an experiment without knowledge of its results or prior to its execution. E.g., retrofit analysis that compares the cost-effectiveness of different energy conservation measures (ECMs).
- Making scientific pronouncements about phenomena that cannot currently be studied experimentally. E.g., Observing changes in building energy performance considering different climate change scenarios or a scenario where occupancy and building usage can become sporadic in the event of a pandemic.
- Using computation to extrapolate existing understanding into experimentally unexplored regimes. E.g., using BES to create a baseline for quantifying energy savings for buildings with multiple interactive ECMs.

1.2. Calibration, validation, and verification

Calibration, validation, and verification are commonly used in the existing literature to indicate consistency between model predictions and actual observations, which can be misleading since they are not synonymous. The semantics of the words calibration, validation, and verification have been subject to philosophical debates because of the paradoxical view that models are a representation of reality and thus by definition not true [9]. As such, there have been arguments that simulation models can never be validated but can only be improved through invalidation [10]. Nevertheless, BES models are typically created for practical purposes such as architectural design, HVAC design and operation, retrofit analysis, building operational optimization, urban-scale energy efficiency analysis, etc. Therefore, from a computational simulation and engineering perspective, the idea of validation is not to establish the truth of a scientific theory but to evaluate and quantify if the model is acceptable for its intended purpose.

Within this context, BES calibration, validation, and verification can be formally defined following the guide by the American Institute of Aeronautics and Astronautics (AIAA) [11], which is also compatible with the US Department of Energy's Advanced Simulation and Computing's (ASC) definitions [8]:

Calibration : The process of adjusting numerical or physical modeling parameters in the computational model for the purpose of improving agreement with experimental data.

Validation : The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.

Verification : The process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model.

1.3. Related work

Over the past two decades, three literature review articles [12–14] have been published regarding BES calibration. In 2005 as part of an ASHRAE initiated research project (RP-1051), Reddy [12] classified calibration methodologies from existing literature into four classes: (1) calibration based on manual, iterative, and pragmatic intervention; (2) calibration based on a suite of informative graphical comparative displays; (3) calibration based on specific tests and analytical procedures; and (4) analytical/mathematical methods of calibration. In 2014, Coakley et al. [13] extended these classifications to include advancements in optimization techniques, Bayesian calibration, and alternative modeling techniques such as meta-modeling. Additionally, a broader definition of calibration approaches as either manual or automated was proposed. In the following year, Fabrizio and Monetti [14] built upon the study by Coakley et al. [13] by discussing in further detail the issues affecting BES calibration.

1.4. Aim and objectives

Although there have been numerous BES calibration studies over the past decade, most studies focused on applying a specific calibration methodology to specific case study buildings. Combined with the lack of open code and data, BES calibration remains challenging to replicate. Additionally, as described in the preceding paragraph, existing review articles focus on providing an overview of current calibration methodologies. However, proper specification of model inputs and outputs is equally important. To date, there has been little quantitative analysis about model inputs and outputs, calibration methods, and the criteria for evaluating calibration performance. The determination of all these details is highly subjective, often requiring a high level of expertise, experience, and domain knowledge. Despite its importance, there is little guidance on best practices to facilitate BES calibration.

With the aim of enhancing reproducibility and enabling others to build upon published work more easily, the objectives of this review article are to:

- Synthesize relevant BES literature and the relationship between various model inputs and outputs.
- Perform a detailed meta-analysis of the calibration methods and measures of calibration performance currently utilized in the existing literature.
- Provide recommendations to facilitate reproducibility in BES.

We believe that meeting these objectives will provide a solid foundation and platform for future research to advance the current state of BES calibration. This is also the first systematic review on the subject.

In Section 2 we describe the methodology of our systematic review. Section 3 contextualizes the review with an overview of the simulation engine used, and the location of case studies. Section 4 describes the state-of-the-art calibration approaches, including a comparison against the 2014 review by Coakley et al. [13]. Section 5 analyzes the relationship between the inputs and outputs used for calibration. Section arizes the metrics commonly used to evaluate calibration performance. Section 7 discusses the significant findings and identifies areas for future research. Section 8 concludes the paper.

2. Method

A systematic review was adopted to provide a comprehensive and unbiased summary of evidence on calibration methods and techniques, model inputs/outputs, and calibration performance metrics.

2.1. Search and eligibility criteria

Table 1 presents the search strategy used to identify relevant publications from the Scopus database. The keywords “model calibration” and “building energy simulation” were used to identify an initial list of publications. To capture as many relevant publications as possible, synonyms that are interchangeable with “model calibration” and “building energy simulation” in the BES literature were included in the search string.

The initial search returned 2,762 publications. Limiting the search to English journal articles published between 2015 and 2020 resulted in 781 publications. We limit the review to the immediate past six years to reflect recent trends and state-of-the-art in BES. Additionally, the most recent review paper for BES calibration was in the year 2014 [13] and 2015 [14].

Further refinements to the search criteria were made by including relevant subject areas (Engineering, Environmental Science, and Energy) and explicitly excluding irrelevant subject areas (Material science, Social Sciences, and Chemical engineering). These criteria excluded 338 studies and left 443 for the review. The titles and abstracts of the 443 studies were subsequently screened to identify relevant publications and their corresponding source journal, yielding 186 publications.

2.2. Study selection

The full papers of the 186 publications were subsequently screened for relevance to this review based on the following criteria: (1) the study involved

Table 1: Search criteria for systematic literature review

Criteria	Description
Keywords	[“calibration” OR “model calibration”] AND [“building performance simulation” OR “building energy model” OR “building energy modeling” OR “building energy simulation” OR“building simulation” OR “energy simulation” OR “whole building energy model”]
Database	Scopus
Search date	16 January 2021
Limit to	Year: 2015-2020 Language: English Document type: Article Source type: Journal Subject areas: Engineering, Environmental Science, Energy Journals: Energy and Buildings, Applied Energy, Automation in Construction, Building and Environment, Solar Energy, Applied thermal energy, Journal of Building Performance Simulation, Journal of Building Engineering, Building Simulation, HVAC and R Research
Exclude	Subject areas: Material science, Social Sciences, Chemical engineering
Total number of publications returned: 186	

the use of building energy simulation; (2) the study contains the application of calibration methods or techniques; and (3) the study is not vague on the proposed calibration approach and is not ambiguous on the model input(s) and output(s). Of the 186 studies, 107 were selected for the review.

3. Study context

This section provides the context to this review by summarizing the 107 selected papers regarding the simulation engine used and the location of the calibration case studies.

3.1. Simulation engine

A majority of the papers reviewed (60%) used EnergyPlus for a variety of reasons (Table 2). EnergyPlus is an open-source whole-building energy simulation engine that has been and continues to be supported by the U.S. Department of Energy (DOE) [15]. Moreover, EnergyPlus supports many application software [16]. 15% and 7% of the papers reviewed use TRNSYS and DOE-2 respectively. TRNSYS [17] is a transient systems simulation program that is designed to provide flexibility in conducting energy simulations through a modular structure and extensive add-on component libraries. DOE-2 [18] is a building energy simulation program that performs hourly simulation given descriptions of the building layout, constructions, operating schedules, HVAC systems, and utility rates.

What stands out in Table 2 is that Resistance-Capacitance (RC) networks (also known as lumped parameter models) were used in 10% of the studies reviewed. Unlike the other three commonly used white box simulation engines (EnergyPlus, TRNSYS, and DOE-2), an RC network is a gray-box model that combines simplified physical representations of the building with operation data that are used to identify the model's coefficients [19]. The benefits of an RC network lies in having physical descriptions of the building while being computationally more efficient than white-box models. Further easing implementation, the development of RC models may also follow the well-established international standard ISO 13790:2008 [20] that was subsequently revised by ISO 52016-1:2017 [21].

Table 2: Simulation engines used in the reviewed calibration applications ($N = 107$).

Simulation Engine	Type	Percentage of Papers Reviewed
EnergyPlus	white-box	60%
TRNSYS		15%
DOE-2		7%
Resistance-Capacitance (RC)	Gray-box	10%
Others	NA	8%

3.2. Location

Fig. 1 shows the geographic distribution of the case study buildings extracted from the papers reviewed across various simulation scales (top map plot) and within each köppen climate zones (bottom bar-plot). From the figure, it is apparent that a majority of case study applications are located in the U.S. or Europe, with several in China and South Korea. These applications are situated between latitude 30°N and 65°N. Consequently, 96% of the studies belong to an arid (dry), temperate (mild mid-latitude), or continental (cold mid-latitude) climate group. Only 4% of the applications were in the equatorial region characterized by a warm and humid climate all year round.

An inspection of the data in Fig. 1 reveals that over three-quarters of the case studies are at the building scale (77%). 15% at the urban-scale and the remaining 8%[1] for the calibration of a single building component or system. A further observation that emerged from the data was that urban-scale case studies are located only in the U.S. (54%), Europe (34%), and the Middle East (12%). None of the urban-scale studies were located in the tropics.

[1] 8% does not include whole building calibration studies that employ multi-stage or sequential calibration approaches that utilize the calibration of a single building component or system as part of the process.

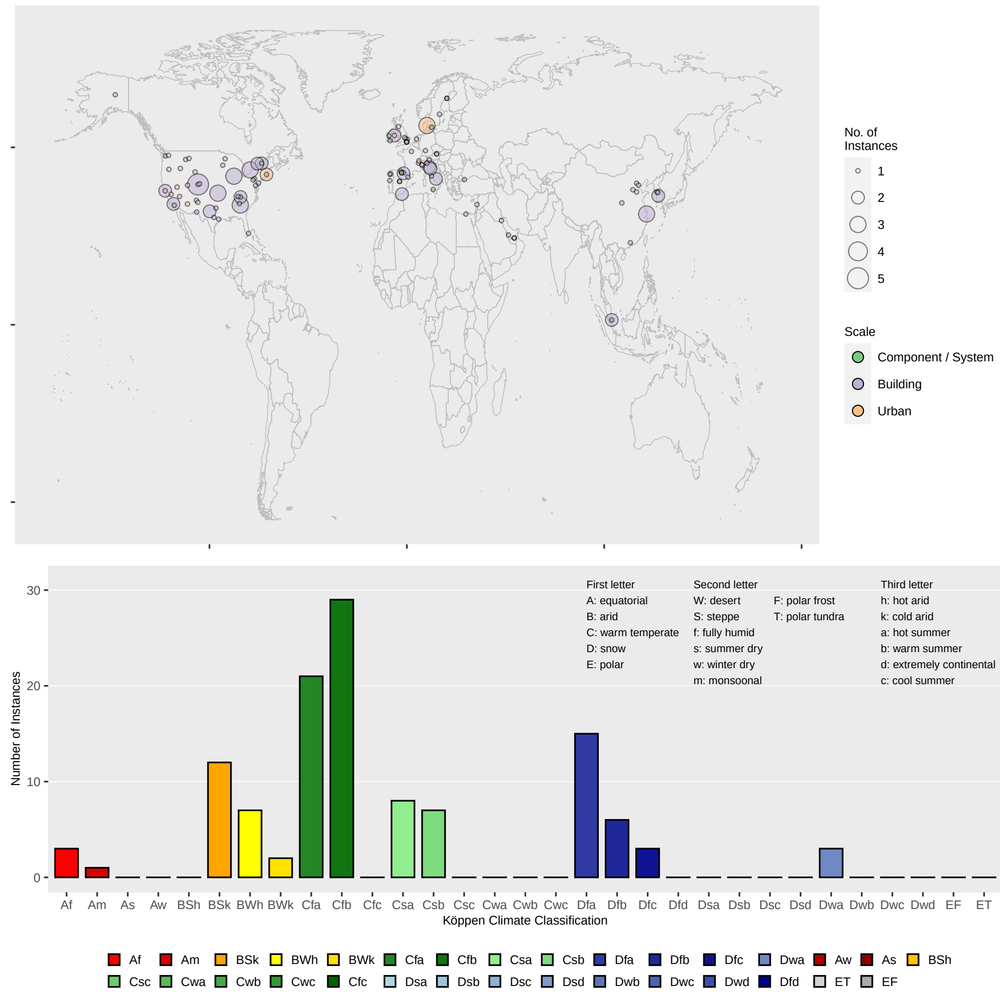


Figure 1: Geographic distribution of case study buildings and the corresponding scale of the simulation (component/system, building, or urban) (top plot) and the distribution of the case study buildings based on the köppen climate zones (bottom plot.

4. Key calibration approaches

In general, calibration approaches can be classified as either manual or automated [13]. Automated approaches employ some form of computerized processes to tune model parameters by maximizing the model's fit to observations. In contrast, manual approaches rely on iterative pragmatic intervention by the modeler. The number of papers utilizing an automated calibration approach has approximately tripled when comparing this review to the review by Coakley et al. [13] in 2014 (Fig. 2).

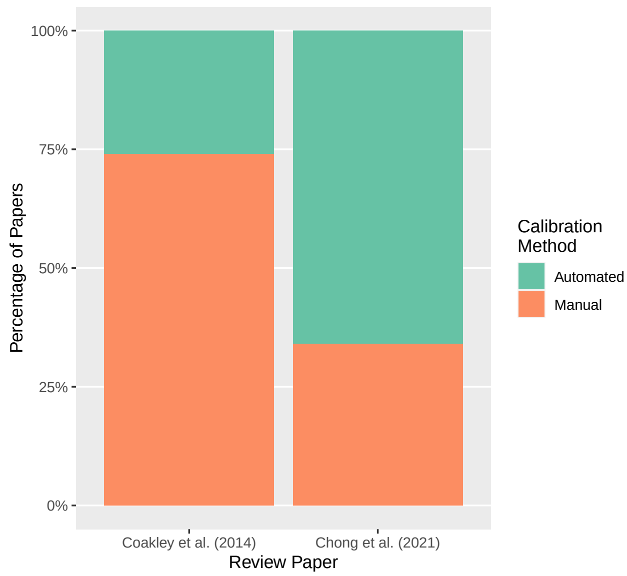


Figure 2: Comparison of the type of calibration approach (Automated or Manual) in this review with that by Coakley et al. (2014) [13]

In this review, a majority of the automated approaches employ either mathematical optimization (58.5%) or Bayesian calibration (33%), with several using sampling methods to select a subset of models with the best fit (8.5%). To aid future applications, Table 3 provides a list of packages, libraries, code repositories, and applications for sensitivity analysis, optimization and Bayesian calibration.

4.1. Optimization-based calibration

Genetic algorithm (GA) [34–42], particle swarm optimization (PSO) [43–51], and the Hooke-Jeeves (HJ) algorithm [51–55] are the most widely used algorithms for optimization-based calibration. Both GA and PSO belong to the class of evolutionary algorithms that are population-based with a metaheuristic characteristic. Specifically, the non-dominated sorting genetic algorithm II

Table 3: Applications, R packages, Python libraries and code repositories for performing sensitivity analysis, optimization, and Bayesian inference algorithms on building energy simulation models.

Name	Type	Language	Method(s)	Ref.
Sensitivity Analysis				
sensitivity	CRAN	R	SRC, SRRC, PRCC, Morris, FAST, Sobol	[22]
SALib	PyPI/GitHub	Python	Morris, FAST, Sobol	[23]
Optimization-based calibration				
GenOpt	Application	Java	GPSHJ, PSO	[24]
jEPlus	Application	Java	NSGA-II	[25]
DEAP	PyPI/GitHub	Python	NSGA-II, PSO	[26]
ecr	CRAN/GitHub	R	NSGA-II, PSO	[27]
Bayesian calibration				
SAVE	CRAN/GitHub	R	Bayesian emulation, calibration, and validation following Bayarri et al. [28] with roots in Kennedy and O'Hagan [29] and Higdon et al. [30]	[31]
bc-stan	GitHub	R	Bayesian emulation and calibration following Kennedy and O'Hagan [29] and Higdon et al. [30]	[32]
pySIP	GitHub	Python	Bayesian emulation and calibration for continuous time stochastic state-space (e.g. RC networks)	[33]

Abbreviations: PyPI (Python Package Index); CRAN (Comprehensive R Archive Network); SRC (Standardized Regression Coefficients); PRCC (Partial Rank Correlation Coefficient); SRRC (Standardized Rank Regression Correlation); FAST (Fourier Amplitude Sensitivity Testing); GPSHJ (Generalized Pattern Search Hooke Jeeves); PSO (Particle Swarm Optimization); NSGA-II (Non-dominated sorting genetic algorithm II);

(NSGA-II) algorithm has been extensively applied for the optimization-based calibration of BES models [34–39] because of its ability to obtain a better spread of solutions and convergence than other multi-objective evolutionary algorithms

[56]. Proposed by Kennedy and Eberhart [57], PSO optimizes via swarm intelligence and is inspired by the social behavior of organisms in groups such as a bird flock or a fish school. Lastly, the HJ algorithm [58] belongs to the family of generalized pattern search (GPS) algorithms and has gained popularity in BES because the number of function evaluations increases only linearly with the number of design parameters [24].

A common feature of the GA, PSO, and HJ algorithm is that they are all gradient-free. Therefore, they are suitable for optimization frameworks that minimize a cost function that needs to be evaluated by an external BES program. Additionally, population-based metaheuristic algorithms such as PSO and GA initialize the optimization with a population of randomly distributed points to reduce the risk of converging to local minima. However, situations of falling far from the pareto-optimal front can be hard to detect, and therefore defining a stopping criterion is difficult. Although guidelines [59–61] specifying thresholds for accuracy metrics such as CV(RMSE) and NMBE are often used, it has been shown that these are not proper stopping criteria for optimization-based calibration [38]. Nonetheless, minimizing CV(RMSE) was also found to be the most robust cost function under different combinations of error metrics, calibration output, and calibration dataset time resolution [38]. Constraints in the form of a specified range for the modeling parameters are often added to prevent unreasonable values [62].

Optimization-based calibration has been widely applied in BES (Fig. 2). A prominent example is the Autotune project that aims to replace manual calibration with a calibration method that leverages supercomputing, large databases of simulation results, and an evolutionary algorithm to automate the calibration process [63, 64]. Sun et al. [65] proposed a pattern-based optimization approach that determines the parameters to tune based on the identified bias in monthly utility bills. Yang and Becerik-Gerber [66] performed independent single objective optimization at the component, zone, and building level. The union of the independent solution sets is then used for the subsequent multi-objective optimization.

Optimization has also been used for the calibration of building components and sub-systems such as models of BIPV [48], absorption thermal energy storage [67], and components of the air-handling unit [49]. Likewise, optimization has been used to calibrate urban-scale building energy models (UBEMs). Santos et al. [68] calibrated 56 buildings in a district using GA while considering the urban heat island effect. Zekar and Khatib [69] applied optimization for the calibration of an urban-scale RC model. Since numerical optimization can be computationally impractical at the urban-scale, optimization-based calibration of UBEMs often involves first parameter reduction in the form of day typing, zone typing, and the use of archetypes.

4.2. Calibration under uncertainty

Uncertainty is an inevitable characteristic of BES models because of the complexity and interactions between different building systems. In many building energy applications, uncertainty management is an important aspect when accounting for risk in the decision-making process. It is somewhat surprising that only 32 of the 107 (∼30%) papers reviewed involved some form of uncertainty quantification during model calibration.

4.2.1. Types and sources of uncertainty

In general, uncertainty can be classified as either aleatory or epistemic [70, 71]. Aleatory uncertainty (or irreducible uncertainty) is the uncertainty caused by inherent variations or randomness of the building system or subsystem under investigation that cannot be explained by the data collected. In contrast, epistemic uncertainty (or reducible uncertainty) is the uncertainty that arises from a lack of knowledge (or data). The distinction between aleatory and epistemic uncertainty has merit in guiding the uncertainties that have the potential of being reduced [72]. However, developing BES models involves a significant degree of subjectivity that depends on the data available that may also evolve throughout a building's lifecycle. As a result, most uncertainties are often a combination of both aleatory and epistemic uncertainty, making it difficult to distinguish between the two.

Related to the types of uncertainty is the identification and classification of uncertainty by their sources, which forms an important part of a comprehensive uncertainty quantification framework [71, 73]. In BES calibration, the sources of uncertainty can be classified as follows:

- Parameter uncertainty: Uncertainty associated with influential model inputs that are not known with certainty.
- Model form uncertainty: Model discrepancy (also called model inadequacy) that results from all assumptions, conceptualizations, abstractions, and approximations of the real-world physical processes.
- Observation uncertainty: Uncertainties that result from observation errors.

4.2.2. Uncertainty quantification

Uncertainty quantification in BES can be broadly categorized as either forward or inverse. Forward approaches quantify uncertainty in the model output(s) by propagating them from uncertainties in the model parameters (Fig. 3). Statistical sampling techniques such as Monte Carlo simulation or Latin Hypercube Sampling are easy to apply and the most common in the field of BES [74–78].

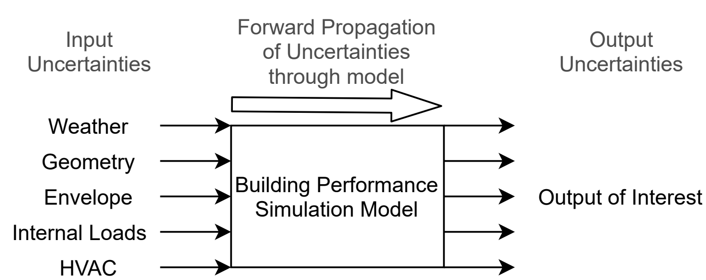


Figure 3: Forward approach to uncertainty quantification by propagating input uncertainties to obtain uncertainties in the output of interest.

Inverse approaches involve quantifying various sources of uncertainties given a set of observations from the building system being modeled. In particular, the calibration paradigm known as Bayesian calibration has gained popularity in BES due to its ability to naturally incorporate uncertainty and combine prior information with measured data to derive posterior estimates of the model parameters (Eq. 1).

$$p(t \mid y) \propto p(y \mid t) \times p(t) \tag{1}$$

A notable approach within the Bayesian calibration paradigm is the formulation proposed by Kennedy and O'Hagan (KOH) [29]. KOH's approach differs from traditional approaches by allowing for various sources of uncertainty and attempting to correct for any model inadequacy or bias (Eq. 2). There have been several applications of KOH's approach in the field of BES calibration [79–87], including a detailed guideline for its application in the field [32].

$$y(x) = \eta(x,t) + \delta(x) + \epsilon(x) \tag{2}$$

where, $y(x)$ is the observed field measurement, $\eta(x,t)$ is the output of the BES given observable inputs x and calibration parameters t, $\delta(x)$ is the model inadequacy, and $\epsilon(x)$ is the observation errors.

Other noteworthy approaches include Bayesian hierarchical modeling for the calibrating of urban-scale building energy models [88, 89] and sequential updating taking advantage of Bayes theorem to keep the model up to date without losing past knowledge [85, 90].

Given the complexity of BES models, posterior distributions often cannot be derived analytically. Consequently, Markov chain Monte Carlo (MCMC) is often used in Bayesian calibration to sample from the posterior distributions because of its flexibility and straightforward application to complex problems. However, it is well known that performing Bayesian inference via MCMC is computationally expensive, especially when likelihood evaluations involve computationally expensive models such as in the case of BES. The Gelman-Rubin statistic ($\hat{R}$), also known as the potential scale reduction factor, is often used to determine if convergence to a stationary distribution has been achieved [32, 81, 88, 91, 92].

To alleviate the high computation cost of Bayesian inference, metamodels have been proposed as surrogates of the energy model. Gaussian processes (GP) [32, 79–82, 85, 86, 91, 93] and linear regression [83, 87, 94–96] are the most popular with competing trade-offs between computation cost and accuracy. Additionally, more efficient MCMC sampling strategies such as Hamiltonian Monte Carlo (HMC) [81, 97] and Approximate Bayesian Computation (ABC) methods [98] have also been proposed to reduce computation cost.

4.3. Analytical Tools and Techniques

Analytical tools and techniques are often applied to both manual or automated calibration approaches. Coakley et al. [13] list these techniques with detailed explanations. Table 4 presents a subset of the techniques [13] that is relevant to this review. As can be seen from Table 4, We do not extend the classifications proposed in [13] but augment their descriptions so that it encompasses the publications reviewed.

4.3.1. Analytical techniques by approach and application

Fig. 4 provides an overview of the number of papers employing a certain analytical technique to assist or complete the calibration process. What stands out in the figure is that the application of sensitivity analysis (SA) and the use of high-resolution data (HIGH) have the highest frequency. By contrast, in the review by Coakley et al. [13], SA and the use of high-resolution data are not common analytical techniques that form a part of the calibration process.

The increase in the use of high-resolution data could be attributed to the proliferation of IoT devices and sensor networks in buildings making hourly and sub-metered data more readily available for calibration. The increase in the use of SA could be associated with the growth in the utilization of automated approaches (Fig. 2). This is further corroborated by Fig. 5, which shows the breakdown of analytical techniques according to the calibration approach

Table 4: Analytical tools and techniques that were used to support the calibration process applied in the papers reviewed. Adapted from [13].

Acronym	Name	Description
SA	Sensitivity analysis	Used to provide insights on how variations in uncertain inputs map onto the outputs. Can be used to identify non-influential parameters that are ignored during calibration or to help set priorities for future efforts (e.g. identify important parameters for measurements or detailed investigation).
HIGH	High-res data	Utilizes data at hourly (or sub-hourly) resolutions as opposed to daily or monthly temporal resolution data
AUDIT	Detailed audit	Conducting detailed audits to gain a better knowledge of the building systems or sub-systems.
UQ	Uncertainty quantification	The assessment of parameter uncertainty as part of the calibration process.
EXPERT	Expert knowledge / templates / model database	Approach which utilize expert knowledge or judgment as a key element of the process. Often involves the use of databases or templates of typical building parameters and components to reduce user input requirements.
PARRED	Parameter reduction	Involves reducing the number of model parameters. Examples include day-typing (reducing detailed schedules into typical day-type schedules), zone-typing (aggregating spaces with similar thermal zones) or building archetypes (grouping buildings into representative archetypes for urban-scale model calibration)
BASE	Base-case modeling	The use of measured base-loads to calibrate the building model. Calibration is carried out during the base-case when (a) heating and cooling loads are minimal and the building is dominated by internal loads, thus minimizing the impact of weather-dependent variables, or (b) internal loads are minimal and the HVAC system is not operating to better characterize weather dependent variables such as the building envelope when internal temperatures are free-floating.
EVIDENCE	Evidence-based model development	Approaches that implement a procedural approach to model development, making changes according to source evidence rather than ad-hoc intervention. Often requires model development version controls to keep track of the changes.
[99]		
SIG	Signature analysis	The use of graphical analysis techniques to identify the impacts of different model parameters on the output of interest.
STEM	Short-term energy monitoring	On-site measurements for a short period of time. Typically used to identify typical energy end-use profiles and/or base-loads.
INT	Intrusive testing	Intrusive techniques involving interventions in the operation of the actual building.

(manual or automated) and the spatial scale (component/system, building, or urban). The results, as shown in Fig. 5 indicates that sensitivity analysis (SA), high-resolution data, uncertainty quantification (UQ), and building audits are

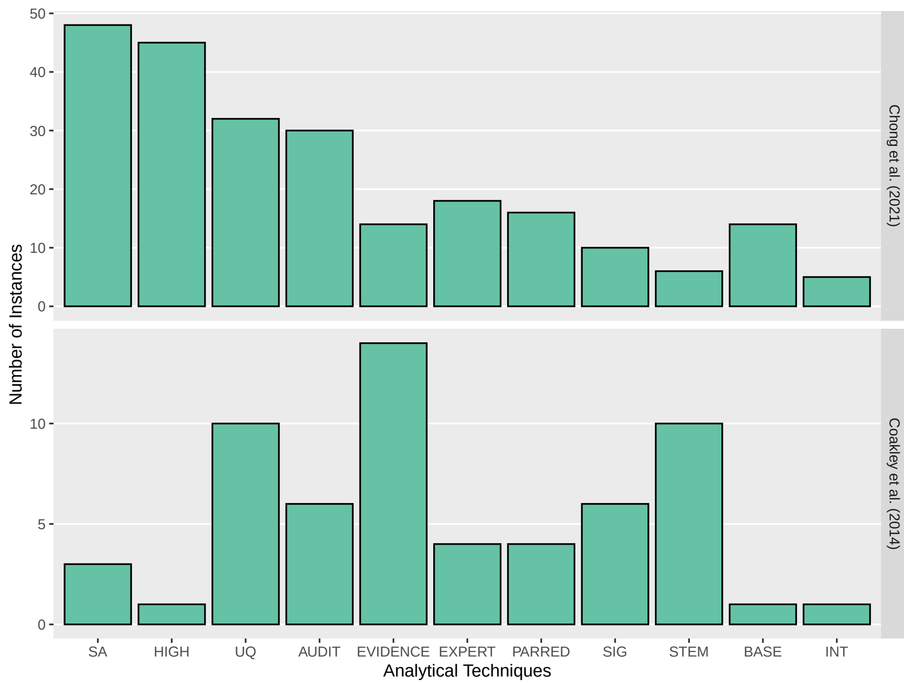


Figure 4: Analytical techniques (for both manual and automated calibration approaches) used in this review (top plot) and the review by Coakley et al. (2014) [13].

the most commonly applied in automated approaches. Comparatively, SA is not as widely used in manual calibration approaches.

Moving on to model calibration at different spatial scales, it can be observed that SA, high-resolution data, UQ, and building audits are prevalent at the building-scale. On the contrary, parameter reduction and expert knowledge are the predominant analytical techniques for urban scale model calibration efforts. Parameter reduction aims to reduce the number of model inputs by characterizing and grouping similar inputs to reduce the complexity of the model while preserving the final decision based on the full set of parameters. Well-known examples of parameter reduction techniques in BES are day-typing (grouping schedules with similar profile) and zone-typing (grouping similar thermal zones) [13]. At the urban-scale, archetypes are commonly used to reduce the number of model inputs and therefore the effort and cost of modeling distinct buildings

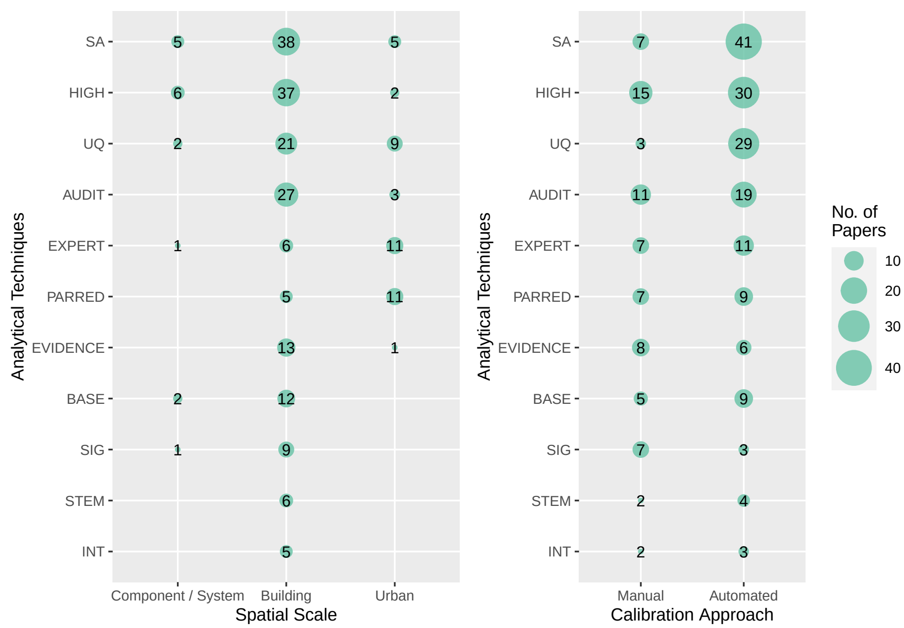


Figure 5: Analytical techniques used in the papers reviewed grouped by the corresponding spatial scale (left plot) and calibration approach (right plot). Calibration processes can involve more than one analytical technique. Therefore, the values do not add up to the total number of papers reviewed (N=107).

[87, 89, 92, 95, 100–102].

Archetype generation involves two steps, segmentation followed by characterization [103]. Segmentation divides buildings with similar characteristics based on key parameters such as building type, construction year or period, floor area, building height, and/or shape (if geometry data is not available) [87, 89, 95, 100]. What follows is the characterization of building construction and operation properties based on expert knowledge that involves deriving input values from existing databases, building codes and standards, and representations of national building typologies (also referred to as reference buildings). For example, several studies [89, 92, 101] modeled construction properties (inferred from construction year) using information from the TABULA (2009-2012) and EPISCOPE (2013-2016) projects [104, 105] that were aimed at providing national residential building typologies for various European countries. Another example is the use of the U.S. Commercial Buildings Energy Consumption Survey (CBECS) and the U.S. Residential Energy Consumption Survey (RECS) databases to derive detailed information on the construction and operation of the buildings (e.g. insulation levels, internal loads and schedules, mechanical systems, and hot water consumption) [95, 106]. Likewise, Chen, Deng and Hong [102] derived input values based on the minimum energy efficiency requirements in the California's building energy efficiency standards Title 24 while Krayem et al. [107] defined internal loads and schedules following the ASHRAE 90.1 Standard.

4.3.2. Sensitivity Analysis

The results of this review confirm the close association between sensitivity analysis (SA) and automated model calibration processes (Fig. 6). Only about 20% of the papers utilizing manual approaches employ SA in contrast with 65% of the papers for automated approaches. A possible explanation is that equifinality issues are especially challenging for automated approaches since objective functions are normally designed to minimize discrepancies between simulated and observed responses. This might produce a model with a higher prediction accuracy, but it might not inform the modeler about the true parameter values [108]. In contrast, manual approaches adjust the calibration parameters based on heuristics that are based on the expertise of an experienced modeler. Of the studies that employed a manual approach without sensitivity analysis, the dominant (86%) analytical techniques employed include conducting detailed audits [109–115], utilizing expert knowledge or judgment [100, 106, 107, 116–119], implementing an evidence-based approach[100, 112, 120–124], or using high-resolution data [100, 112, 120–122, 124].

It is evident from Fig. 6 that global sensitivity analyses are more commonly used in automated approaches. A possible explanation is the ability of global methods to provide an overall view of the importance of different inputs while considering their interactions [125]. Specifically, screening methods (46%) are the most popular followed by perturbation (23%), regression (13%), metamodel (10%), variance (4%), and regional sensitivity analysis (RSA) (4%) (Fig. 6).

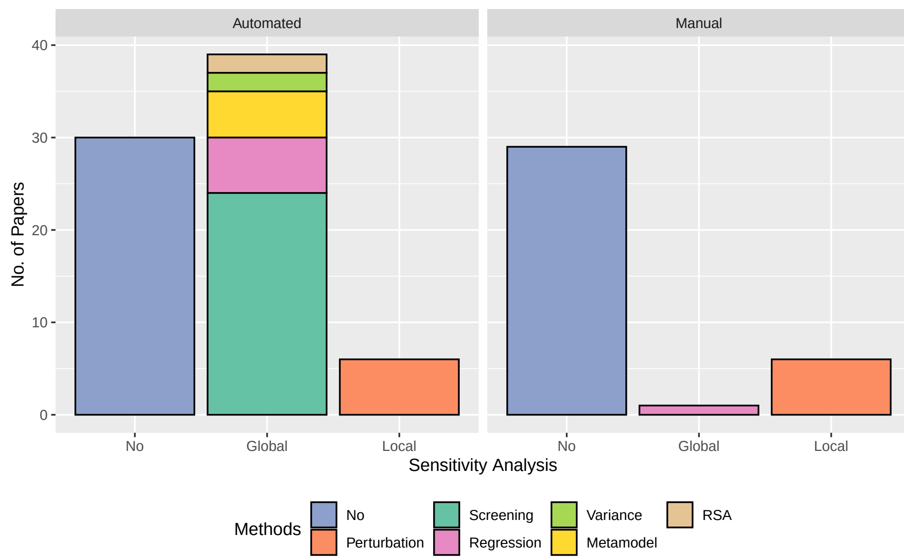


Figure 6: Types of sensitivity analysis used in building energy simulation split by automated vs manual calibration approaches.

In this review, variance-based SA methods are not common because they are computationally demanding requiring large sample sizes for accurate approximations of the sensitivity indices. Suppose that there are t uncertain parameters, the approximate number of model evaluations required is approximately t for perturbation local SA methods; $10 - 100t$ for screening methods; $100 - 1000t$ for regression and RSA methods; and $> 1000t$ for variance-based methods [125]. Consequently, metamodels, surrogates or emulators are typically used in place of computationally expensive simulation runs required by computationally demanding SA methods [125]. Specific choices of metamodels may also provide sensitivity measures that can be used to rank model parameters according to their influence on the output of interest. Examples include the use of random forest variable importance [91, 94, 96] and estimates of marginal posterior using Gaussian processes [82].

Screening methods are popular due to their low computation cost compared to other global SA methods, making it suitable for BES models that are typically non-linear with high-dimensional parameter space. The method of Morris [126] is the most established and widely used screening method. Sampling for the Morris method is carried out by randomly selecting r starting points that are perturbed One-at-A-Time (OAT). The computation cost is therefore $r(t+1)$ for t model parameters. A measure of global sensitivity is commonly obtained using the r trajectories to compute the mean μ [126] or the modified mean μ^* proposed by [127]. In general, most studies rely on graphical plots of μ (or μ^*) and σ for better interpretability when screening out non-influential parameters [32, 36, 37, 39, 49, 50, 52, 66, 79, 81, 85, 88, 128]. Others consider only μ (or μ^*) to rank and identify dominant parameters [45, 47, 78, 84, 86, 101].

Perturbation methods are the simplest type of SA and involve varying (perturbing) the model inputs from their base or nominal values One-At-a-Time (OAT). Compared with global SA methods, the advantages of perturbation methods include (1) its ease of application and interpretation [129], and (2) requiring the least number of model evaluations [125]. The order of influence is often used to identify a subset of parameters to be calibrated [65, 128, 130–135]. Sun et al. [65] illustrate this clearly by using parametric perturbation to identify a priority list of 17 calibration parameters that would be adjusted within a pattern-based automated calibration framework. On the other hand, studies have also applied perturbation methods after the calibration to investigate possible causes for remaining discrepancies between simulated and measured data [136] or to determine if the calibrated model is robust to uncertainties in particular input factors [137].

Regression methods refer to the use of regression or correlation coefficients to derive information about output sensitivity to variations in input uncertainty. The input/output dataset is often generated using Monte Carlo simulation or Latin hypercube sampling. Several types of regression or correlation coefficients have been used as sensitivity measures in building energy analysis [129]. The choice depends on linearity and monotonicity assumptions between the inputs and output [125]. In this review, we found that Standardized Regression Coefficients (SRC) was most commonly applied [62, 91, 94, 96] with some studies employing partial rank correlation coefficient (PRCC) [46] and standardized rank regression correlation (SRCC) [138].

4.4. Multi-stage calibration

Supporting multi-stage calibration through a combination of data from building information models (BIM), as-built documents, on-site audits, occupancy sensors, indoor environmental quality (IEQ) sensors, the building management system (BMS), and metered HVAC component energy consumption may not be a far-fetched reality that is only possible for state-of-the-art buildings as building data becomes more easily available and accessible. This review found that more than 90% of the papers reviewed calibrated the model against a maximum of one (62%) or two (29%) outputs. About 8% of the studies calibrated the model against three outputs, while only 1% performed the calibration using three outputs.

Multi-stage calibration is of interest because it is often proposed to more accurately represent the building being modeled. Since calibration is an underdetermined problem [32], it is possible for a model that is calibrated at a coarser spatial or temporal level to meet the most stringent error thresholds without accurately representing the building at finer spatial or temporal levels [37, 99, 139]. It has been argued that it is crucial to achieve simultaneous accuracy at multiple levels of the simulation to correctly provide insights at the respective levels. For instance, Yang and Becerik-Gerber [66] asserts that building-level accuracy is needed to provide insights on overall energy performance but an ECM level accuracy would be needed for estimating the energy savings potential of different ECMs. Similarly, Li et al. [140] showed using statistical hypothesis testing that an energy model calibrated for one ECM cannot be used to accurately cross-estimate the energy consumption of another ECM.

Our review found that calibrating the model with data of the building under free-floating conditions is a dominant feature of multi-stage approaches

[50, 51, 76, 141]. Such a base-case method is often employed because the number of uncertain parameters is substantial reduced when there are little or no internal loads (in particular occupancy) and the HVAC system is not operating. Additionally, it is well-known that occupancy is a highly uncertain model input [142] with significant influence on the predictive accuracy of a calibrated energy model [140, 143, 144]. Consequently, parameters like envelope material thermal properties and infiltration rates are more straightforward to identify during periods when the building is free-floating (See Section 5.4).

5. Data requirements

5.1. Inputs and outputs

As illustrated in Fig. 7, BES models represent aspects of reality that are manipulated and experimented with using a variety of simulations [145, 146]. Therefore, the goal is to have models adequately represent actual building performance over a sufficiently wide range of inputs that encompasses the simulation aim and thus application. Since BES models are complex computer models with many inputs and outputs, a comparison of the input and output mapping is carried out through a meta-analysis of the existing literature. To avoid confusion, we refer to the following input classification scheme (Fig. 8) for model calibration.

First, model inputs can be observed or unobserved. We use the verb observed/unobserved instead of the adjectives observable/non-observable because what is observable may differ across different calibration cases. We also make a distinction between the model's variables and parameters. Variables refer to inputs to the model that varies over time and are not always observable. By contrast, parameters do not relate to time-varying values but to quantities that influence the output or behavior of the model. In some contexts, a quantity could be either a variable or a parameter depending on how it is modeled. Window opening, for example, could be modeled as a variable using a schedule to define when the window is open/close or parameterized to be based on occupant comfort [147]. Out of the unobserved parameters, those assumed to be responsible for the discrepancy between simulation and measured output(s) are then calibrated. Fig. 7 illustrates the typical calibration process with reference to Fig. 8. In this diagram, observed input(s) refers to both input parameters and variables whose value could be determined directly or derived/estimated from available evidence or data. The model is just a representation of reality and the selected calibration parameters are tuned to match simulated to observed output(s).

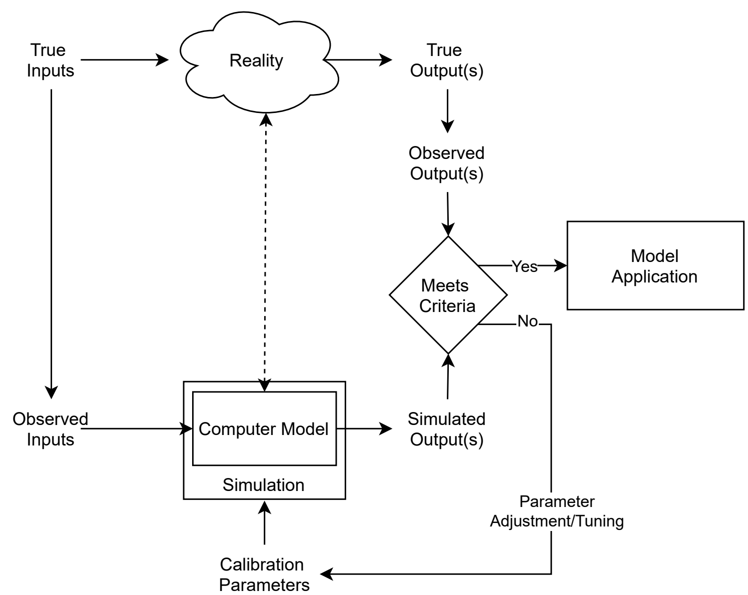


Figure 7: Schematic illustrating the calibration process given that the simulation is a representation of reality.

5.2. Most common observed outputs

It is apparent from Fig 9 that most of the studies were conducted at the building scale. A clear trend of decreasing temporal data resolution as we move from component/system to building to urban scale building energy simulations can also be observed. A closer inspection of the figure shows that models of building components and sub-systems are often calibrated using sub-hourly or

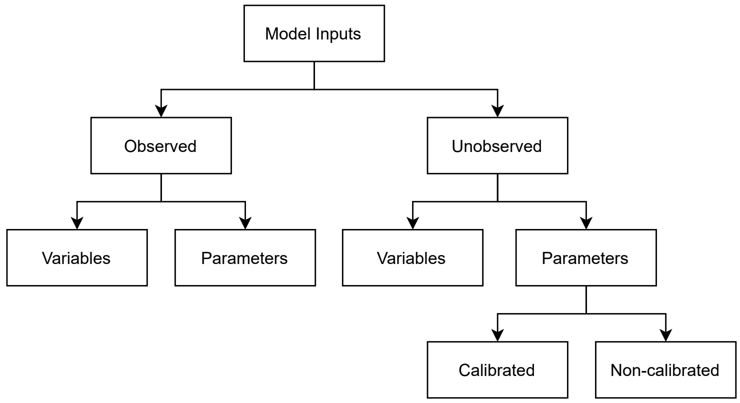


Figure 8: Classification of model inputs when calibrating an energy model.

hourly data regardless of the type of outputs used for the calibration. Additionally, the use of HVAC energy was the most common [49, 51, 81, 148].

Compared with components and sub-systems, electricity and dry bulb temperature are most frequently used for the calibration of energy models at the building-scale. Specifically, the adjustment of parameters using indoor air temperature is often carried out during free-floating periods when the indoor temperatures are allowed to float without any HVAC system intervening. Consequently, outdoor air temperature is also frequently monitored concurrently to provide the boundary conditions of the simulation [36, 37, 43, 45, 50, 76, 79, 97, 110, 120, 134, 141, 149–152]. However, what stands out in building scale studies is that calibration against indoor dry bulb temperature [35, 37, 38, 43, 45, 50, 54, 55, 76–78, 90, 97, 110, 120, 122, 124, 131, 134, 137, 149, 150], HVAC energy [48, 66, 121, 135, 139, 153], and equipment electricity consumption [53, 134, 139, 143] is almost always carried out at an hourly resolution. In contrast, calibration against electricity and gas/steam energy is generally carried out with monthly resolution data.

Turning now to urban-scale building energy models (UBEM), about two-thirds of the studies used monthly or annual electricity, gas/steam, total load/energy, and/or cooling load/energy for the calibration. The use of monthly or annual measurements for the calibration is not surprising because using higher resolution data might be computationally intractable at the urban-scale. Additionally, UBEM studies often utilize utility data that are only available at a monthly resolution [40, 95, 107].

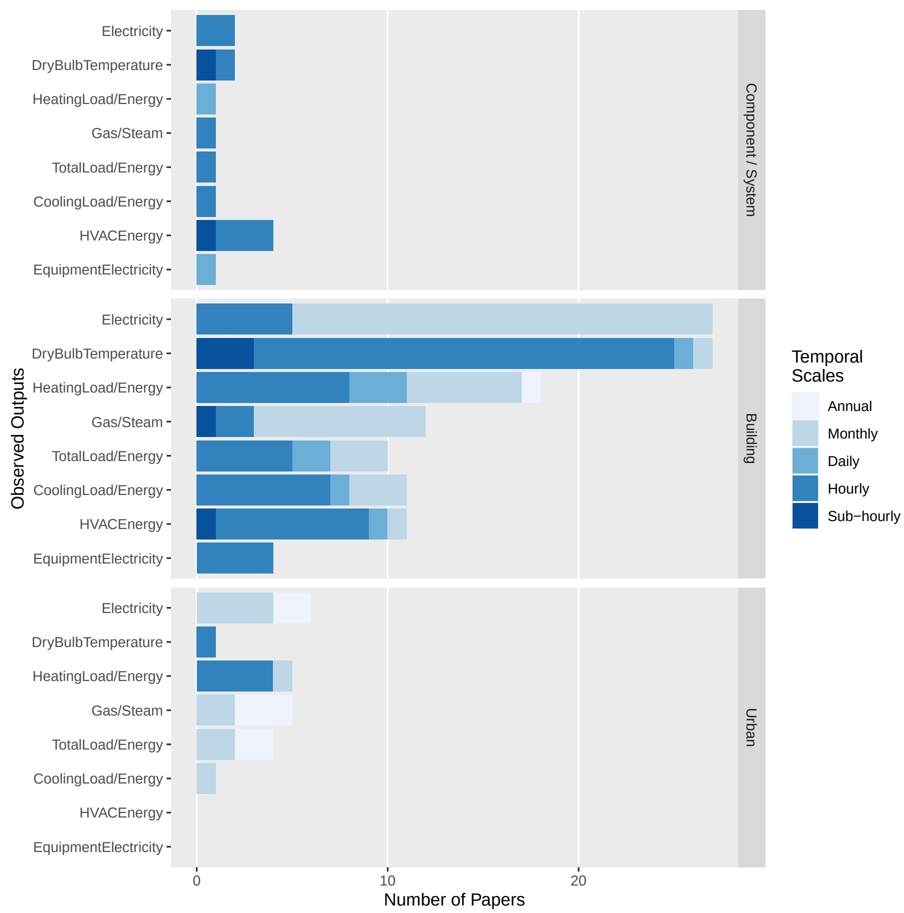


Figure 9: Most common observed output used for calibrating building energy simulation models split by the temporal resolution used for the calibration and the scale of the energy model (i.e., calibration of Component / System, Building, or Urban scale building energy models).

5.3. Most common observed inputs

Fig. 10 provides an overview of the most commonly used observed inputs. What stands out from Fig. 10 is the obvious use of local weather data (dry bulb temperature, solar radiation, relative humidity, wind speed and direction) as observed inputs to the model. If local site measurements are not available, an annual meteorological year (AMY) weather file from the nearest weather station is used. These observations indicate the importance of using actual weather data for the calibration since the weather file forms the energy simulation's boundary conditions. The weather file's importance was also demonstrated in previous research that showed that the annual building energy consumption and the monthly building loads could vary by $\pm 7\%$ and $\pm 40\%$, respectively, based on the provided weather data [154].

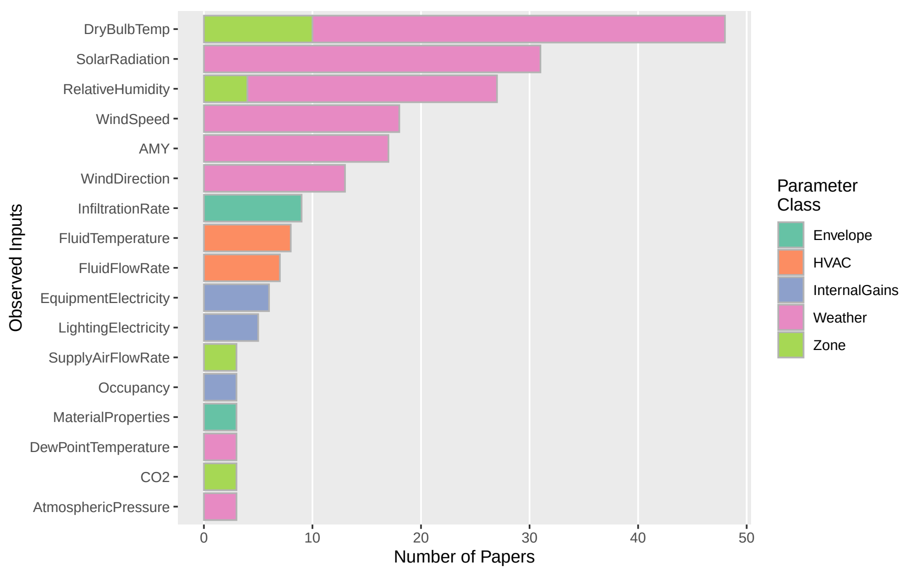


Figure 10: Most common observed inputs used for calibrating building energy simulation models. The color indicates the class of the model parameter.

Interestingly, several studies used measured indoor environmental conditions as inputs to the model to obtain a model that is better calibrated at the zone level. For instance, Mihai and Zmeureanu [137] showed that using measured indoor air temperatures in place of those from the technical specification led to more accurate predictions of zone airflow rates. Yin, Kiliccote, and Piette [139] used air temperature and airflow measurements to derive the zone thermostat setpoint and the VAV box minimum/maximum airflow respectively. Infiltration rate is sometimes derived from measurements because it is highly uncertain and can have a significant influence on a building's energy use [155]. In this review, infiltration rates are typically derived from airtightness values that are obtained using the blower door test [34, 36, 37, 45, 100, 111, 113, 156].

5.4. Mapping calibration parameters to outputs

Fig. 11 lists the model parameters most commonly adjusted to match simulation output to the measurements faceted by the type of sensitivity analysis conducted and whether the calibration utilized an automated or manual approach. The figure reveals two interesting observations. First, SA, especially global SA, is less likely to be used when the calibration involves using schedules (occupant, equipment, lighting, and HVAC operation). Second, automated calibration approaches tend to calibrate parameters such as material properties, infiltration rate, and internal load densities compared to schedules. By contrast, manual approaches are equally likely to calibrate material properties and schedules.

Fig. 12 shows the magnitude of the relationship between the most commonly used calibration parameters and their corresponding observed outputs. The mapping reveals that parameters concerning the building envelope (material properties and infiltration rate), internal gains (occupant, lighting, and equipment power density), and zone cooling and heating setpoints are often adjusted when calibrating the energy model to building electricity energy consumption. Closer inspection of the first column of Fig. 12 shows that HVAC component efficiency and zone outdoor air levels were also calibrated in a considerable number of papers. Not surprisingly, hot water usage was also adjusted in several studies that were calibrating models of residential typologies [95, 130, 157, 158]. About six to seven articles calibrated the internal loads' schedule [40, 65, 102, 107, 143, 157, 159].

Fig. 12 reveals several other interesting observations. First, the same calibration parameters were used when calibrating against both building electricity

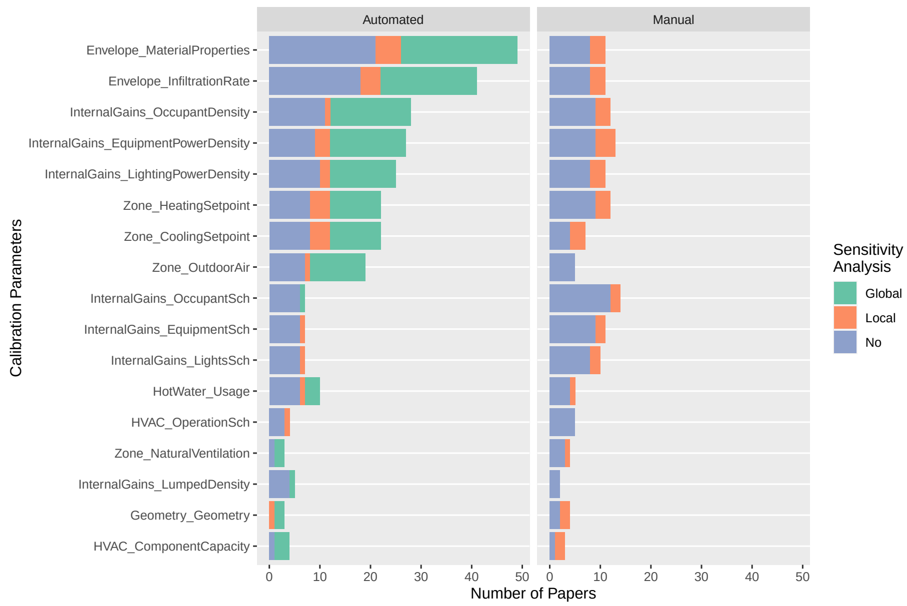


Figure 11: Most common calibration parameters used for calibrating building energy simulation models split by sensitivity analysis and calibration approach.

and gas/steam energy consumption. Second, the parameters calibrated when matching simulation predictions to total building load/energy are somewhat similar, except that equipment and lighting schedules are less likely to be adjusted.

Turning to zone dry-bulb temperature as the observed output, parameters concerning envelope material properties are the most commonly adjusted, followed by infiltration rate. A similar observation can be made when the model is calibrated to the building's heating or cooling load/energy. Material properties and infiltration rate are selected because indoor temperature measurements are often used to investigate the relative changes in the building envelope performance with varying boundary conditions [51, 131, 149, 150]. Additionally, studies have found parameters affiliated with infiltration rate and material properties to influence indoor air temperature [45, 50, 76, 79, 131].

Finally and intuitively, Fig. 12 shows that HVAC component capacity and efficiency are typically adjusted when the observed output is the HVAC component's energy consumption. Likewise, EPD and equipment schedules are adjusted when the model is calibrated against equipment energy consumption.

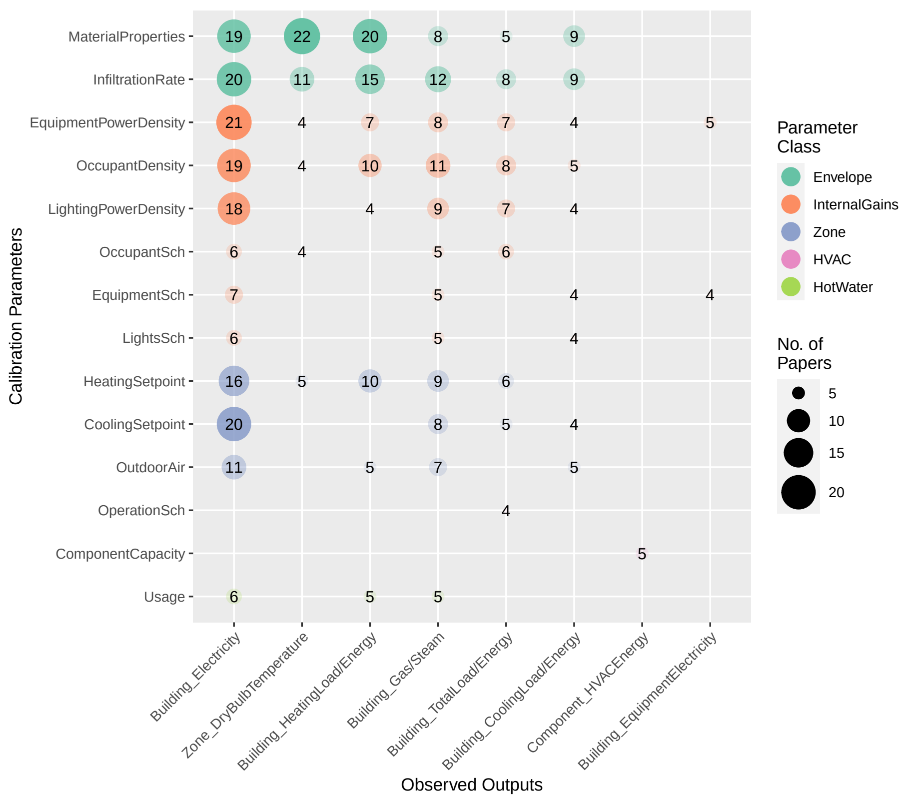


Figure 12: The magnitude of the relationship between the calibration parameters and their corresponding observed outputs for calibrating building energy simulation models.

6. Calibration performance evaluation

6.1. Current approaches

Table 5 ranks the metrics used to assess calibration performance based on the number of occurrences in the papers reviewed. A large proportion of the papers use CV(RMSE) or NMBE to determine if a BES model was calibrated. This result is not unexpected since BES models are often deemed "calibrated" if they meet the CV(RMSE) and NMBE limits (Table 6) specified by ASHRAE Guideline 14 [59], the International Performance Measurement and Verification Protocol (IPMVP) [60], or the Federal Energy Management Program (FEMP) [61]. Interestingly, approximately half of the papers reviewed (51%) used two metrics in their evaluation. 24% utilized one metric while 19% used three metrics simultaneously. The remaining 7% of the papers reviewed used either four or five metrics for the assessment.

Table 5: Metrics used for the evaluation of calibration performance. Each paper may employ more than one metric when assessing calibration performance. Therefore, the cumulative sum for the column "No. of Papers" is greater than the total number of papers reviewed (N=107).

Metric	Acronym	No. of Papers
Coefficient of Variation of the Root Mean Square Error	$CV(RMSE)$	72
Normalized Mean Bias Error	$NMBE$	59
Root Mean Square Error	$RMSE$	20
Coefficient of Determination	R^2	12
Goodness of Fit	GOF	11
Annual Percentage Error	APE	8
Coefficient of Variation	CV	4
Mean Absolute Percentage Error	$MAPE$	4
Mean Absolute Error	MAE	3
Gelman-Rubin statistic	$\hat{R}$	3
Others†	–	18

† Metrics with $\leq$ 2 counts.

CV(RMSE) (eq. 3) provides an indication of how close the simulation predictions are to measured data while NMBE (eq. 4) serves as an indicator of overall bias in the simulation predictions. However, NMBE suffers from cancellation between positive and negative bias which can lead to misleading interpretations of predictive performance [160]. This review also confirms the findings of Ruiz and Bandera [161] that the NMBE acronymn is often erroneously referred to as MBE even though the formula is correct (i.e., MBE (%) = formula for NMBE).

Table 6: Error limits specified by various guidelines and protocols for a building energy simulation model to be deemed calibrated.

Guideline / Protocol	Monthly Criteria (%)		Hourly Criteria (%)	
	NMBE	CV(RMSE)	NMBE	CV(RMSE)
ASHRAE Guideline 14 [59]	±5	15	±10	30
IPMVP [60]	–	–	±5	20
FEMP [61]	±5	15	±10	30

NMBE is MBE normalized by the mean of the observed values so that they are comparable. Several papers also utilize RMSE, which provides a measure of the variability of the residuals and is the non-normalized form of CV(RMSE).

$$CV(RMSE) = \frac{1}{\bar{m}} \cdot \sqrt{\frac{\sum_{i=1}^{n}(m_i - s_i)^2}{n-p}} \times 100 \tag{3}$$

$$NMBE(\%) = \frac{1}{\bar{m}} \cdot \frac{\sum_{i=1}^{n}(m_i - s_i)}{n-p} \times 100 \tag{4}$$

where m_i and s_i are the measured and simulated values respectively, $\bar{m}$ is the mean of the measured values, n is the number of data points, and p is the number of adjustable model parameters.

Around 10% of the papers use GOF and R^2 to assess calibration performance. GOF (eq. 5) which was proposed by ASHRAE RP-1051 [162] incorporates both variance and bias errors through a formulation that considers both CV(RMSE) and NMBE. Since GOF combines CV(RMSE) and NMBE into a single composite function, it has the advantage of being able to identify a single optimal solution and to some extent solve multi-objective optimization problems more effectively. Therefore, it has been used to define the cost function in several optimization-based calibrations [35, 36, 40, 40, 49, 149]. For similar reasons, studies that utilize sampling methods (such as Monte Carlo sampling [102] and Latin Hypercube sampling [75–77]) have also used GOF to rank and identify suitable solutions.

$$GOF = \frac{\sqrt{2}}{2} \cdot \sqrt{CV(RMSE)^2 + NMBE^2} \tag{5}$$

R^2 (eq. 6) provides an indication of the variability in the dependent variable from the mean values that are explained by the regression model. ASHRAE Guideline 14 [59] recommends the use of $CV(RMSE)$ and R^2 to select the best whole-building energy use regression models such as the algorithms of the ASHRAE Inverse Model Toolkit (IMT), which was developed from RP-1050 [163, 164]. Although there is currently no prescribed minimum value for R, IPMVP [165] advised that an R^2 value of 0.75 provides a reasonably good causal relationship between energy use and the independent variables. Using EnergyPlus simulations, Chakraborty and Elzarka [160] demonstrated that R^2 used in tandem with a range normalized RMSE (RN(RMSE)) (eq. 7) would provide a better representation of the predictive performance of system-level energy models.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(m_i - s_i)^2}{\sum_{i=1}^{n}(m_i - \bar{m})^2} \tag{6}$$

where m_i and s_i are the measured and simulated values respectively, $\bar{m}$ is the mean of the measured values, and n is the number of data points.

$$RN(RMSE) = \frac{1}{range(m)} \cdot \sqrt{\frac{\sum_{i=1}^{n}(m_i - s_i)^2}{n-p}} \times 100 \tag{7}$$

where m_i and s_i are the measured and simulated values respectively, m is the difference between the maximum and minimum of the measured values, n is the number of data points, and p is the number of adjustable model parameters.

6.2. Evaluating probabilistic predictions

Calibration methods that involve uncertainty quantification often provide probabilistic predictions to support risk-conscious decision-making. However, almost all of the evaluation methods in the literature evaluate probabilistic predictions in a deterministic manner. Specifically, central tendency measures such as the mean or median are used to compute accuracy metrics, some of which are CV(RMSE) [32, 81, 82, 84, 89, 91, 93–95, 101], NMBE [32, 81, 89, 93], APE [82, 95, 101, 166], RMSE [84, 87, 91], and MAPE [87] (Table 5). However, it has been shown that using a single value such as the mean to represent the entire distribution may result in an optimistic bias of the model's prediction accuracy [85]. Therefore, these metrics are often accompanied by graphical plots comparing probabilistic predictions (e.g. using box-plots or error bars) to the observed values [32, 82, 84, 88, 89].

Alternative assessment methods have also been proposed to more precisely evaluate probabilistic predictions. For example, assessing performance by comparing CV(RMSE) and NMBE median or mean values with their 95% confidence intervals [40, 88, 98]. The Kolmogorov-Smirnov (KS) test has also been used to assess calibration performance by comparing the predicted and measured EUI distributions [95, 166].

In order to facilitate comparison between probabilistic predictions and deterministic observations, Chong, Augenbroe, and Da [144] proposed using the coverage width-based criterion (CWC). Likewise, the continuous rank probability score (CRPS) was proposed to measure the distance between the probabilistic predictions and their corresponding observations. [83]. Both the CWC and the CRPS are the only metrics proposed reviewed that consider both correctness and informativeness of the probabilistic predictions. For detailed explanation and formulation of the CWC and the CRPS, the reader is referred to [144] and [167] respectively.

6.3. Validation using out-of-sample data

63% of the studies reviewed did not evaluate the calibrated model on an out-of-sample test dataset. The remaining 37% advocated the use of an out-of-sample test dataset to avoid bias in the evaluation process. For instance, out-of-sample test buildings have been used to evaluate the robustness and homogeneity of urban-scale archetype predictive performance [88, 89, 95]. In contrast to out-of-sample buildings, Hedegarrd et al. [92] calibrated 159 BES models using one month of hourly data, and evaluated their predictions using the subsequent month. Wang et al. [101] calibrated 84 residential buildings using five years of monthly data and evaluated their predictions using the subsequent two years of data.

At the building-scale, the out-of-sample dataset typically comprises either a randomly sampled subset of the time-series data that was not used for the calibration [32, 81], data from a period after the model was calibrated [42, 66, 82, 83, 139, 158], or a selected period based on occupancy levels and season [124].

7. Discussion

7.1. Inputs and outputs

The most prominent finding from the meta-analysis is that monthly building electricity and hourly indoor dry bulb temperature measurements are most commonly used to calibrate BES models, especially at the building scale. A possible explanation is that electricity and gas/steam data are often obtained from utility providers who typically provide monthly data. In comparison, measurements of the other outputs such as HVAC energy, equipment electricity, and indoor dry bulb temperature would involve installing sub-meters and/or accessing the building automation system where data is usually available at sub-hourly resolution.

Another finding is that material thermophysical properties, infiltration rate, and internal load densities are frequently selected for calibration, especially in automated calibration frameworks. It is well known amongst researchers in BES calibration that these parameters are the main model parameters used to describe a building and often represent a significant source of uncertainty when estimating building energy performance. One well-known early study that is often cited for uncertainties in infiltration rates is that of Persily [168]. Likewise, material properties have also been shown to be uncertain due to various reasons such as poor detailing/workmanship and thermal bridges [2, 169, 170].

Previous studies have demonstrated the importance of schedule adjustment in model calibration [143, 144]. Therefore, it is somewhat surprising that schedules are typically not considered in automated calibration frameworks. This inconsistency might be due to the sharp increase in computation cost if every schedule parameter were considered in the calibration. Another possible explanation for not considering schedules is that it could result in identifiability issues if a comprehensive dataset is not available to avoid overparameterization [32, 171]. Consequently, schedule adjustment typically involves simplification, such as selecting from a list of predefined discrete schedules that best fit the measured data [40, 102, 107, 157]. As data in the built environment becomes more available and accessible, developing scalable calibration algorithms that can consider multiple data sources might prove important in future research.

7.2. Calibrating urban-scale models

The result of this review indicates that approximately 15% of the papers are at the urban-scale. Of the 15%, most are located in the U.S. (54%) and Europe (34%), with none in a tropical climate. Since the urban context (inter-building effects and urban microclimate) is an important aspect that should be considered in UBEMs [172, 173], it would be interesting to evaluate the performance of the UBEM calibration methodologies in the tropics and cities outside of the U.S. and Europe.

This review also found that UBEMs are typically calibrated using monthly or annual data (Fig. 9) and rely on expert knowledge and parameter reduction techniques (Fig. 5). This finding is consistent with previous studies, which found that the UBEM generation process typically relies on assumptions due to a lack of data quality and accessibility [103, 174]. Consequently, the credibility of UBEMs is often questionable due to the widespread use of default or reference values, and the fact that UBEM calibration remains a significantly overparameterized problem. However, as discussed in the preceding section 7.3, predictive accuracy is not synonymous to model credibility. The need for parsimonious models was also recently asserted in a review of UBEM use cases [175].

With IoT and the proliferation of sensors in the built environment, wide-ranging data streams at increasing spatial and temporal scales may be more easily accessible in the future. Having access to large amounts of data entails other challenges such as ensuring data quality and consistency [174, 176], and selecting only the necessary information needed for energy modeling [177]. However, it also brings the opportunity for future research that investigates the minimum level of complexity and data required to achieve a UBEM that is commensurate with its intended purpose. Data convergence across multiple spatial and temporal scales is needed to support such work. Additionally, an interdisciplinary team of modelers, urban planners, policymakers, and decision-makers more generally is necessary to address these challenges.

7.3. Credibility or absolute predictive accuracy

It is apparent from this review that predictive accuracy is widely used to evaluate BES models. However, a model with low absolute predictive accuracy might still be reasonable for its intended use. For example, a relative comparison between different design options only requires relative accuracy, which is typically easier to achieve than absolute predictive accuracy. Likewise, it is also possible for a BES model to exhibit a good fit to observation data but not accurately represent building systems or sub-systems due to many modeling parameters and uncertainties. Chong and Menberg [32] demonstrated that a low CV(RMSE) and NMBE is not indicative of good estimates of the true values of the calibration parameters.

The Cambridge English dictionary defines credibility as “can be believed or trusted”. From a modeling standpoint, credibility would not require the model to be accurate or have high fidelity. For example, Yin, Kiliccote, and Piette [139] evaluated an EnergyPlus model’s prediction of dynamic response by field-testing a set of demand response control strategies. As pointed by Rykiel [178], the crux of the issue is therefore determining (1) if a model is acceptable for its intended purpose; and (2) how confident should we be about the model’s inference about the actual building system. Similar concepts of fit-for-purpose modeling strategies were also discussed in BES, emphasizing the need to consider the aim of the simulation when selecting different modeling approaches [144, 179–182].

Likewise, the choice of model and calibration approach should not be decoupled from the intended purpose of the simulation. Simple models are more transparent and require less data for parameter estimation and calibration but could increase model bias or inadequacy. In contrast, complex models are designed to represent actual physical systems better but tend to be more data and computationally intensive. The challenge then is in being able to abstract a reasonable simplification of reality to meet the simulation objectives while considering the available data. Consequently, it is imperative that the purpose of the simulation and the corresponding performance criteria be specified before any calibration is carried out. However, current calibration studies rely solely on measures of accuracy such as CV(RMSE) and NMBE to determine whether a model is "calibrated". There is currently no guidance on how the credibility of BES models for various applications can be qualified. The association between model complexity, simulation objectives, and data informativeness is also poorly understood. Further research on this topic is therefore recommended.

7.4. Reproducible research in BES

In this review, we found that BES simulations and existing calibration approaches are difficult to reproduce from the publications alone because of the complexity of BES models, and the absence of clarity concerning the reporting of (a) calibration parameters, observed inputs, and observed outputs and (b) assumptions made during data pre- and post-processing.

BES models and the associated code and data should be made openly available to improve the quality of scientific research, reduce duplicated efforts, and facilitate collaborations [183]. Furthermore, reproducibility will become increasingly difficult with increasing data sources and more complex calibration methodologies as we attempt to bridge the gap between simulation and reality. Without access to the code and the data, it would be almost impossible to implement the fundamentals of scientific research that include transparency, rigor, and independent verification [183, 184].

While full reproducibility requires complete openness and familiarity with open-source toolkits, it is still valuable to open parts of the code and/or data. Therefore, we recommend an incremental approach to encourage reproducible research in BES. For a start, publications should include a checklist to ensure clear reporting of the context and processes involved in the calibration (Table 7). Next, all code should be published. The code only needs to be available and does not need to be structured or clean. Even poorly written code informs a lot about the calibration approach. Additionally, a subset of the data or synthetic data can be used where data privacy is of concern.

Table 7: Checklist for improving reproducibility of publications that involves the calibration of building energy simulation models.

Checklist to enhance reproducibility
General Information
Aim of the simulation
Building location (Longitude and Latitude)
Building typology
Weatherfile
Total, conditioned and unconditioned floor area
Simulation engine
Measured Data
Observed output(s) and data source
Observed input(s) and data source
Pre-processing
Calibration Method
Calibration parameters and their corresponding ranges
Calibration approach (Automated or Manual) and algorithm if an automated approach was used
Analytical tools and techniques (see section 4.3)
Calibration sequence if a multi-stage sequential approach is involved
Results and Conclusion
Post-processing
Recommendation

The technical challenges impeding reproducible publications can be summarized as poor documentation of the dependencies necessary for the code to run, imprecise documentation on how to install and run the associated code, and a lack of a robust way to run exact versions of all software involved [185]. Additionally, the choice of either a permissive or copyleft license (i.e., legal terms) should be considered [186]. Docker is a popular platform that can provide the infrastructure to facilitate reproducible research in BES [187, 188]. Specifically, Docker images and Dockerfiles are Docker concepts that help resolve the technical challenges to reproducibility mentioned at the start of this paragraph [185]. To facilitate reproducibility, we created a GitHub repository (section 8) to demonstrate a Docker based approach for reproducing BES research.

8. Conclusion

Calibration remains a challenging task because there are no clear guidance and best practices on calibration procedures such as model inputs and outputs, calibration methods, calibration performance evaluation, simulation reproducibility. As a result, BES calibration has remained highly subjective, and perhaps even elusive, and almost impossible to reproduce. Therefore, this study contributes to existing knowledge of BES calibration by providing a coherent and detailed summary of the calibration methodology, data requirements, performance evaluation criteria, and the current state of knowledge.

The findings indicate a significant increase in the use of automated calibration approaches. Amongst the automated calibration approaches, optimization and Bayesian calibration were the most popular. In general, global sensitivity analysis is often applied within automated approaches. In contrast, the dominant techniques used in manual approaches include using detailed audits, expert knowledge, and/or evidence-based procedures. High-resolution data is prevalent in both automated and manual approaches possibly due to increasing sensing capabilities and data availability in the built environment.

BES models are usually calibrated against one or two observed outputs. The two most commonly used data sources for BES calibration were monthly electricity consumption and hourly indoor dry bulb temperature. Monthly electricity often stems from utility bills and is often used to calibrate the building envelope's thermophysical parameters, infiltration rate, various internal gains densities, and indoor setpoint temperatures. Hourly measurements of indoor dry-bulb temperature during free-floating periods when the indoor temperatures are allowed to float during non-operating hours are often used to calibrate thermophysical parameters of the building envelope and infiltration rate.

The review indicates a lack of reproducibility due to the absence of clarity in reporting the modeling and data assumptions, calibration parameters, observed inputs, and observed outputs. Therefore, an incremental approach to encourage reproducibility in BES research was proposed in this study, along with a fully reproducible example on GitHub (section 8).

Taken together, the present study lays the groundwork that future calibration studies can build on. While it is clear that there is a significant body of work available, the precise mechanism of BES calibration and the evaluation of model credibility remains to be elucidated. Incorporating multiple data sources within automated calibration algorithms would also be exciting for future work with increasing data availability. We also believe that a culture of reproducibility will significantly aid efforts in establishing a standardized calibration methodology.

Data availability

The research compendium for this article can be found at `https://github.com/ideas-lab-nus/calibrating-building-simulation-review`, hosted at GitHub.

The simple example of reproducible building energy simulation (section 7.4) can be found at `https://github.com/ideas-lab-nus/reproducing-building-simulation`, hosted at GitHub.

CRediT authorship contribution statement

Adrian Chong: Conceptualization, Methodology, Formal analysis, Investigation, Writing – Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration. **Yaonan Gu:** Investigation, Data curation. **Hongyuan Jia:** Software, Data curation, Writing - Review & Editing.

Acknowledgements

This research is supported by the National University of Singapore, Singapore under its start-up grant (Project No. R-296-000-190-133); the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore.

References

[1] J. L. Hensen, R. Lamberts, Building performance simulation for design and operation, 2nd Edition, Routledge, 2019.

[2] P. De Wilde, The gap between predicted and measured energy performance of buildings: A framework for investigation, Automation in construction 41 (2014) 40–49. `doi:http://dx.doi.org/10.1016/j.autcon.2014.02.009`.

[3] C. Turner, M. Frankel, et al., Energy performance of leed for new construction buildings, New Buildings Institute (2008) 1–42.

[4] E. Mantesi, C. J. Hopfe, M. J. Cook, J. Glass, P. Strachan, The modelling gap: Quantifying the discrepancy in the representation of thermal mass in building simulation, Building and Environment 131 (2018) 74–98. `doi:https://doi.org/10.1016/j.buildenv.2017.12.017`.

[5] A. C. Menezes, A. Cripps, D. Bouchlaghem, R. Buswell, Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap, Applied energy 97 (2012) 355–364. doi:https://doi.org/10.1016/j.apenergy.2011.11.075.

[6] S. De Wit, G. Augenbroe, Analysis of uncertainty in building design evaluations and its implications, Energy and buildings 34 (9) (2002) 951–958. doi:https://doi.org/10.1016/S0378-7788(02)00070-1.

[7] H. Yoshino, T. Hong, N. Nord, Iea ebc annex 53: Total energy use in buildings—analysis and evaluation methods, Energy and Buildings 152 (2017) 124–136. doi:https://doi.org/10.1016/j.enbuild.2017.07.038.

[8] T. G. Trucano, L. P. Swiler, T. Igusa, W. L. Oberkampf, M. Pilch, Calibration, validation, and sensitivity analysis: What's what, Reliability Engineering & System Safety 91 (10-11) (2006) 1331–1357. doi:https://doi.org/10.1016/j.ress.2005.11.031.

[9] N. Oreskes, K. Shrader-Frechette, K. Belitz, Verification, validation, and confirmation of numerical models in the earth sciences, Science 263 (5147) (1994) 641–646. doi:10.1126/science.263.5147.641.

[10] L. F. Konikow, J. D. Bredehoeft, Ground-water models cannot be validated, Advances in water resources 15 (1) (1992) 75–83. doi:https://doi.org/10.1016/0309-1708(92)90033-X.

[11] A. I. of Aeronautics, Astronautics, AIAA guide for the verification and validation of computational fluid dynamics simulations, American Institute of aeronautics and astronautics., 1998. doi:https://doi.org/10.2514/4.472855.001.

[12] T. A. Reddy, Literature review on calibration of building energy simulation programs: uses, problems, procedures, uncertainty, and tools, ASHRAE transactions 112 (2006) 226.

[13] D. Coakley, P. Raftery, M. Keane, A review of methods to match building energy simulation models to measured data, Renewable and sustainable energy reviews 37 (2014) 123–141. doi:https://doi.org/10.1016/j.rser.2014.05.007.

[14] E. Fabrizio, V. Monetti, Methodologies and advancements in the calibration of building energy models, Energies 8 (4) (2015) 2548–2574. doi:https://doi.org/10.3390/en8042548.

[15] US Department of Energy (DOE) Office of Energy Efficiency Renewable Energy, EnergyPlus.
URL https://www.energy.gov/eere/buildings/downloads/energyplus-0

[16] IBPSA-USA, Building Energy Software Tools (BEST) directory.
URL https://www.buildingenergysoftwaretools.com/software-listing?keywords=EnergyPlus

[17] S. A. Klein, TRNSYS 18: A transient simulation program (2017).
URL https://sel.me.wisc.edu/trnsys

[18] DOE-2. [link].
URL https://www.doe2.com

[19] X. Li, J. Wen, Review of building energy modeling for control and operation, Renewable and Sustainable Energy Reviews 37 (2014) 517–537. doi:http://dx.doi.org/10.1016/j.rser.2014.05.056.

[20] Energy performance of buildings — Calculation of energy use for space heating and cooling, Standard, International Organization for Standardization, Geneva, CH (Mar. 2008).

[21] Energy performance of buildings — Energy needs for heating and cooling, internal temperatures and sensible and latent heat loads — Part 1: Calculation procedures, Standard, International Organization for Standardization, Geneva, CH (Jun. 2017).

[22] B. Iooss, S. D. Veiga, A. Janon, G. Pujol, with contributions from Baptiste Broto, K. Boumhaout, T. Delage, R. E. Amri, J. Fruth, L. Gilquin, J. Guillaume, M. I. Idrissi, L. Le Gratiet, P. Lemaitre, A. Marrel, A. Meynaoui, B. L. Nelson, F. Monari, R. Oomen, O. Rakovec, B. Ramos, O. Roustant, E. Song, J. Staum, R. Sueur, T. Touati, F. Weber, sensitivity: Global Sensitivity Analysis of Model Outputs, r package version 1.24.0 (2021).
URL `https://CRAN.R-project.org/package=sensitivity`

[23] J. Herman, W. Usher, Salib: an open-source python library for sensitivity analysis, Journal of Open Source Software 2 (9) (2017) 97. `doi:10.21105/joss.00097`.

[24] M. Wetter, et al., Genopt-a generic optimization program, in: Seventh International IBPSA Conference, Rio de Janeiro, 2001, pp. 601–608.

[25] Zhang, Yi, JEPlus - An parametric tool for EnergyPlus and TRNSYS.
URL `http://www.jeplus.org/wiki/doku.php`

[26] F.-A. Fortin, F.-M. De Rainville, M.-A. G. Gardner, M. Parizeau, C. Gagné, Deap: Evolutionary algorithms made easy, The Journal of Machine Learning Research 13 (1) (2012) 2171–2175.

[27] J. Bossek, ecr: Evolutionary Computation in R, r package version 2.1.0 (2017).
URL `https://CRAN.R-project.org/package=ecr`

[28] M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, J. Tu, A framework for validation of computer models, Technometrics 49 (2) (2007) 138–154. `doi:https://doi.org/10.1198/004017007000000092`.

[29] M. C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical Methodol-

ogy) 63 (3) (2001) 425–464. doi:https://doi.org/10.1111/1467-9868.00294.

[30] D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, R. D. Ryne, Combining field data and computer simulations for calibration and prediction, SIAM Journal on Scientific Computing 26 (2) (2004) 448–466. doi:https://doi.org/10.1137/S1064827503426693.

[31] J. Palomo, R. Paulo, G. García-Donato, SAVE: An R package for the statistical analysis of computer models, Journal of Statistical Software 64 (13) (2015) 1–23.
URL http://www.jstatsoft.org/v64/i13/

[32] A. Chong, K. Menberg, Guidelines for the bayesian calibration of building energy models, Energy and Buildings 174 (2018) 527–547. doi:10.1016/j.enbuild.2018.06.028.

[33] L. Raillon, S. Rouchier, S. Juricic, pysip: an open-source tool for bayesian inference and prediction of heat transfer in buildings, in: Congres français de thermique, Nantes, 2019.

[34] C. Bandera, G. Ruiz, Towards a new generation of building envelope calibration, Energies 10 (12) (2017). doi:10.3390/en10122102.

[35] G. Ramos Ruiz, C. Fernández Bandera, Analysis of uncertainty indices used for building envelope calibration, Applied Energy 185 (2017) 82–94. doi:10.1016/j.apenergy.2016.10.054.

[36] G. Ramos Ruiz, C. Fernández Bandera, T. Gómez-Acebo Temes, A. Sánchez-Ostiz Gutierrez, Genetic algorithm for building envelope calibration, Applied Energy 168 (2016) 691–705. doi:10.1016/j.apenergy.2016.01.075.

[37] S. Zuhaib, M. Hajdukiewicz, J. Goggins, Application of a staged automated calibration methodology to a partially-retrofitted university build-

ing energy model, Journal of Building Engineering 26 (2019). doi:10.1016/j.jobe.2019.100866.

[38] S. Martínez, P. Eguía, E. Granada, A. Moazami, M. Hamdy, A performance comparison of multi-objective optimization-based approaches for calibrating white-box building energy models., Energy and Buildings 216 (2020). doi:10.1016/j.enbuild.2020.109942.

[39] S. Martínez, E. Pérez, P. Eguía, A. Erkoreka, E. Granada, Model calibration and exergoeconomic optimization with nsga-ii applied to a residential cogeneration, Applied Thermal Engineering 169 (2020). doi:10.1016/j.applthermaleng.2020.114916.

[40] S. Nagpal, J. Hanson, C. Reinhart, A framework for using calibrated campus-wide building energy models for continuous planning and greenhouse gas emissions reduction tracking, Applied Energy 241 (2019) 82–97. doi:10.1016/j.apenergy.2019.03.010.

[41] S. Tian, S. Shao, B. Liu, Investigation on transient energy consumption of cold storages: Modeling and a case study, Energy 180 (2019) 1–9. doi:10.1016/j.energy.2019.04.217.

[42] J. Chen, X. Gao, Y. Hu, Z. Zeng, Y. Liu, A meta-model-based optimization approach for fast and reliable calibration of building energy models, Energy 188 (2019). doi:10.1016/j.energy.2019.116046.

[43] C. Andrade-Cabrera, D. Burke, W. Turner, D. Finn, Ensemble calibration of lumped parameter retrofit building models using particle swarm optimization, Energy and Buildings 155 (2017) 513–532. doi:10.1016/j.enbuild.2017.09.035.

[44] T. Yang, Y. Pan, J. Mao, Y. Wang, Z. Huang, An automated optimization method for calibrating building energy simulation models with measured data: Orientation and a case study, Applied Energy 179 (2016) 1220–1231. doi:10.1016/j.apenergy.2016.07.084.

[45] F. Roberti, U. Oberegger, A. Gasparella, Calibrating historic building energy models to hourly indoor air and surface temperatures: Methodology and case study, Energy and Buildings 108 (2015) 236–243. doi:10.1016/j.enbuild.2015.09.010.

[46] C. Andrade-Cabrera, W. Turner, D. Finn, Augmented ensemble calibration of lumped-parameter building models, Building Simulation 12 (2) (2019) 207–230. doi:10.1007/s12273-018-0473-5.

[47] Q. Zhang, Z. Tian, Z. Ma, G. Li, Y. Lu, J. Niu, Development of the heating load prediction model for the residential building of district heating based on model calibration, Energy 205 (2020). doi:10.1016/j.energy.2020.117949.

[48] S.-W. Ha, S.-H. Park, J.-Y. Eom, M.-S. Oh, G.-Y. Cho, E.-J. Kim, Parameter calibration for a trnsys bipv model using in situ test data, Energies 13 (18) (2020). doi:10.3390/en13184935.

[49] G. Larochelle Martin, D. Monfet, H. Nouanegue, K. Lavigne, S. Sansregret, Energy calibration of hvac sub-system model using sensitivity analysis and meta-heuristic optimization, Energy and Buildings 202 (2019). doi:10.1016/j.enbuild.2019.109382.

[50] M. Ferrara, C. Lisciandrello, A. Messina, M. Berta, Y. Zhang, E. Fabrizio, Optimizing the transition between design and operation of zebs: Lessons learnt from the solar decathlon china 2018 scutxpolito prototype, Energy and Buildings 213 (2020). doi:10.1016/j.enbuild.2020.109824.

[51] A. Cacabelos, P. Eguía, L. Febrero, E. Granada, Development of a new multi-stage building energy model calibration methodology and validation in a public library, Energy and Buildings 146 (2017) 182–199. doi:10.1016/j.enbuild.2017.04.071.

[52] W. Li, Z. Tian, Y. Lu, F. Fu, Stepwise calibration for residential building

thermal performance model using hourly heat consumption data, Energy and Buildings 181 (2018) 10–25. doi:10.1016/j.enbuild.2018.10.001.

[53] A. Abdelalim, W. O'Brien, Z. Shi, Data visualization and analysis of energy flow on a multi-zone building scale, Automation in Construction 84 (2017) 258–273. doi:10.1016/j.autcon.2017.09.012.

[54] A. Ogando, N. Cid, M. Fernández, Energy modelling and automated calibrations of ancient building simulations: A case study of a school in the northwest of spain, Energies 10 (6) (2017). doi:10.3390/en10060807.

[55] E. Carlon, M. Schwarz, A. Prada, L. Golicza, V. Verma, M. Baratieri, A. Gasparella, W. Haslinger, C. Schmidl, On-site monitoring and dynamic simulation of a low energy house heated by a pellet boiler, Energy and Buildings 116 (2016) 296–306. doi:10.1016/j.enbuild.2016.01.001.

[56] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multi-objective genetic algorithm: Nsga-ii, IEEE transactions on evolutionary computation 6 (2) (2002) 182–197. doi:10.1109/4235.996017.

[57] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of ICNN'95-international conference on neural networks, Vol. 4, IEEE, 1995, pp. 1942–1948. doi:10.1109/ICNN.1995.488968.

[58] R. Hooke, T. A. Jeeves, "direct search"solution of numerical and statistical problems, Journal of the ACM (JACM) 8 (2) (1961) 212–229. doi:https://doi.org/10.1145/321062.321069.

[59] ASHRAE, Guideline 14, measurement of energy and demand savings, American Society of Heating, Ventilating, and Air Conditioning Engineers, Atlanta, Georgia (2014).

[60] EVO, International performance measurement and verification protocol: Concepts and options for determining energy and water savings volume 1, Efficiency Valuation Organization (2012).

[61] US DOE FEMP, M&V guidelines: Measurement and verification for performance-based contracts, version 4.0, Energy Efficiency and Renewable Energy (2015).

[62] S. Qiu, Z. Li, Z. Pang, W. Zhang, Z. Li, A quick auto-calibration approach based on normative energy models, Energy and Buildings 172 (2018) 35–46. doi:10.1016/j.enbuild.2018.04.053.

[63] G. Chaudhary, J. New, J. Sanyal, P. Im, Z. O'Neill, V. Garg, Evaluation of "autotune" calibration against manual calibration of building energy models, Applied Energy 182 (2016) 115–134. doi:10.1016/j.apenergy.2016.08.073.

[64] A. Garrett, J. New, Scalable tuning of building models to hourly data, Energy 84 (2015) 493–502. doi:10.1016/j.energy.2015.03.014.

[65] K. Sun, T. Hong, S. Taylor-Lange, M. Piette, A pattern-based automated approach to building energy model calibration, Applied Energy 165 (2016) 214–224. doi:10.1016/j.apenergy.2015.12.026.

[66] Z. Yang, B. Becerik-Gerber, A model calibration framework for simultaneous multi-level building energy simulation, Applied Energy 149 (2015) 415–431. doi:10.1016/j.apenergy.2015.03.048.

[67] H. Schreiber, F. Lanzerath, A. Bardow, Predicting performance of adsorption thermal energy storage: From experiments to validated dynamic models, Applied Thermal Engineering 141 (2018) 548–557. doi:10.1016/j.applthermaleng.2018.05.094.

[68] L. Santos, A. Afshari, L. Norford, J. Mao, Evaluating approaches for district-wide energy model calibration considering the urban heat island effect, Applied Energy 215 (2018) 31–40. doi:10.1016/j.apenergy.2018.01.089.

[69] A. Zekar, S. Khatib, Development and assessment of simplified building representations under the context of an urban energy model: Application

to arid climate environment, Energy and Buildings 173 (2018) 461–469. doi:10.1016/j.enbuild.2018.04.030.

[70] W. Tian, Y. Heo, P. De Wilde, Z. Li, D. Yan, C. S. Park, X. Feng, G. Augenbroe, A review of uncertainty analysis in building energy assessment, Renewable and Sustainable Energy Reviews 93 (2018) 285–301. doi:https://doi.org/10.1016/j.rser.2018.05.029.

[71] C. J. Roy, W. L. Oberkampf, A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing, Computer methods in applied mechanics and engineering 200 (25-28) (2011) 2131–2144. doi:10.1016/j.cma.2011.03.016.

[72] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? does it matter?, Structural safety 31 (2) (2009) 105–112. doi:10.1016/j.strusafe.2008.06.020.

[73] Y. Sun, Closing the building energy performance gap by improving our predictions, Ph.D. thesis, Georgia Institute of Technology (2014).

[74] G. Yun, K. Song, Development of an automatic calibration method of a vrf energy model for the design of energy efficient buildings, Energy and Buildings 135 (2017) 156–165. doi:10.1016/j.enbuild.2016.11.060.

[75] L. Harmer, G. Henze, Using calibrated energy models for building commissioning and load prediction, Energy and Buildings 92 (2015) 204–215. doi:10.1016/j.enbuild.2014.10.078.

[76] J. Cipriano, G. Mor, D. Chemisana, D. Pérez, G. Gamboa, X. Cipriano, Evaluation of a multi-stage guided search approach for the calibration of building energy simulation models, Energy and Buildings 87 (2015) 370–385. doi:10.1016/j.enbuild.2014.08.052.

[77] N. Sakiyama, L. Mazzaferro, J. Carlo, T. Bejat, H. Garrecht, Natural ventilation potential from weather analyses and building simulation, Energy and Buildings (2020). doi:10.1016/j.enbuild.2020.110596.

[78] M. Giuliani, G. Henze, A. Florita, Modelling and calibration of a high-mass historic building for reducing the prebound effect in energy assessment, Energy and Buildings 116 (2016) 434–448. doi:10.1016/j.enbuild.2016.01.034.

[79] S. Martínez, A. Erkoreka, P. Eguía, E. Granada, L. Febrero, Energy characterization of a paslink test cell with a gravel covered roof using a novel methodology: Sensitivity analysis and bayesian calibration, Journal of Building Engineering 22 (2019) 1–11. doi:10.1016/j.jobe.2018.11.010.

[80] K. Menberg, Y. Heo, R. Choudhary, Influence of error terms in bayesian calibration of energy system models, Journal of Building Performance Simulation 12 (1) (2019) 82–96. doi:10.1080/19401493.2018.1475506.

[81] A. Chong, K. Lam, M. Pozzi, J. Yang, Bayesian calibration of building energy models with large datasets, Energy and Buildings 154 (2017) 343–355. doi:10.1016/j.enbuild.2017.08.069.

[82] J. Yuan, V. Nian, B. Su, Q. Meng, A simultaneous calibration and parameter ranking method for building energy models, Applied Energy 206 (2017) 657–666. doi:10.1016/j.apenergy.2017.08.220.

[83] Q. Li, G. Augenbroe, J. Brown, Assessment of linear emulators in lightweight bayesian calibration of dynamic building energy models for parameter estimation and performance prediction, Energy and Buildings 124 (2016) 194–202. doi:10.1016/j.enbuild.2016.04.025.

[84] Y. Heo, D. Graziano, L. Guzowski, R. Muehleisen, Evaluation of calibration efficacy under different levels of uncertainty, Journal of Building Performance Simulation 8 (3) (2015) 135–144. doi:10.1080/19401493.2014.896947.

[85] A. Chong, W. Xu, S. Chao, N.-T. Ngo, Continuous-time bayesian calibra-

tion of energy models using bim and energy data, Energy and Buildings 194 (2019) 177–190. doi:10.1016/j.enbuild.2019.04.017.

[86] S. Chen, D. Friedrich, Z. Yu, J. Yu, District heating network demand prediction using a physics-based energy model with a bayesian approach for parameter calibration, Energies 12 (18) (2019). doi:10.3390/en12183408.

[87] G. Tardioli, A. Narayan, R. Kerrigan, M. Oates, J. O'Donnell, D. Finn, A methodology for calibration of building energy models at district scale using clustering and surrogate techniques, Energy and Buildings 226 (2020). doi:10.1016/j.enbuild.2020.110309.

[88] M. Kristensen, R. Hedegaard, S. Petersen, Hierarchical calibration of archetypes for urban building energy modeling, Energy and Buildings 175 (2018) 219–234. doi:10.1016/j.enbuild.2018.07.030.

[89] M. Kristensen, R. Hedegaard, S. Petersen, Long-term forecasting of hourly district heating loads in urban areas using hierarchical archetype modeling, Energy 201 (2020). doi:10.1016/j.energy.2020.117687.

[90] S. Rouchier, M. Jiménez, S. Castaño, Sequential monte carlo for on-line parameter estimation of a lumped building energy model, Energy and Buildings 187 (2019) 86–94. doi:10.1016/j.enbuild.2019.01.045.

[91] H. Lim, Z. Zhai, Comprehensive evaluation of the influence of meta-models on bayesian calibration, Energy and Buildings 155 (2017) 66–75. doi:10.1016/j.enbuild.2017.09.009.

[92] R. Hedegaard, M. Kristensen, T. Pedersen, A. Brun, S. Petersen, Bottom-up modelling methodology for urban-scale analysis of residential space heating demand response, Applied Energy 242 (2019) 181–204. doi:10.1016/j.apenergy.2019.03.063.

[93] Y.-J. Kim, C.-S. Park, Stepwise deterministic and stochastic calibration of an energy simulation model for an existing building, Energy and Buildings 133 (2016) 455–468. doi:10.1016/j.enbuild.2016.10.009.

[94] H. Lim, Z. Zhai, Influences of energy data on bayesian calibration of building energy model, Applied Energy 231 (2018) 686–698. doi:10.1016/j.apenergy.2018.09.156.

[95] J. Sokol, C. Cerezo Davila, C. Reinhart, Validation of a bayesian-based method for defining residential archetypes in urban building energy models, Energy and Buildings 134 (2017) 11–24. doi:10.1016/j.enbuild.2016.10.050.

[96] W. Tian, S. Yang, Z. Li, S. Wei, W. Pan, Y. Liu, Identifying informative energy data in bayesian calibration of building energy models, Energy and Buildings 119 (2016) 363–376. doi:10.1016/j.enbuild.2016.03.042.

[97] L. Lundström, J. Akander, Bayesian calibration with augmented stochastic state-space models of district-heated multifamily buildings, Energies 13 (1) (2019). doi:10.3390/en13010076.

[98] C. Zhu, W. Tian, B. Yin, Z. Li, J. Shi, Uncertainty calibration of building energy models by combining approximate bayesian computation and machine learning algorithms, Applied Energy 268 (2020). doi:10.1016/j.apenergy.2020.115025.

[99] P. Raftery, M. Keane, A. Costa, Calibrating whole building energy models: Detailed case study using hourly measured data, Energy and buildings 43 (12) (2011) 3666–3679. doi:https://doi.org/10.1016/j.enbuild.2011.09.039.

[100] J. Fernandez, L. del Portillo, I. Flores, A novel residential heating consumption characterisation approach at city level from available public data: Description and case study, Energy and Buildings 221 (2020). doi:10.1016/j.enbuild.2020.110082.

[101] C.-K. Wang, S. Tindemans, C. Miller, G. Agugiaro, J. Stoter, Bayesian calibration at the urban scale: a case study on a large residential heating demand application in amsterdam, Journal of Building Performance Simulation 13 (3) (2020) 347–361. doi:10.1080/19401493.2020.1729862.

[102] Y. Chen, Z. Deng, T. Hong, Automatic and rapid calibration of urban building energy models by learning from energy performance database, Applied Energy 277 (2020). doi:10.1016/j.apenergy.2020.115584.

[103] C. F. Reinhart, C. C. Davila, Urban building energy modeling–a review of a nascent field, Building and Environment 97 (2016) 196–202. doi:http://dx.doi.org/10.1016/j.buildenv.2015.12.001.

[104] I. Ballarini, S. P. Corgnati, V. Corrado, Use of reference buildings to assess the energy saving potentials of the residential building stock: The experience of tabula project, Energy policy 68 (2014) 273–284. doi:https://doi.org/10.1016/j.enpol.2014.01.027.

[105] T. Loga, B. Stein, N. Diefenbach, Tabula building typologies in 20 european countries—making energy-related features of residential building stocks comparable, Energy and Buildings 132 (2016) 4–12. doi:https://doi.org/10.1016/j.enbuild.2016.06.094.

[106] Z. Taylor, Y. Xie, C. Burleyson, N. Voisin, I. Kraucunas, A multi-scale calibration approach for process-oriented aggregated building energy demand models, Energy and Buildings 191 (2019) 82–94. doi:10.1016/j.enbuild.2019.02.018.

[107] A. Krayem, A. Al Bitar, A. Ahmad, G. Faour, J.-P. Gastellu-Etchegorry, I. Lakkis, J. Gerard, H. Zaraket, A. Yeretzian, S. Najem, Urban energy modeling and calibration of a coastal mediterranean city: The case of beirut, Energy and Buildings 199 (2019) 223–234. doi:10.1016/j.enbuild.2019.06.050.

[108] K. Beven, A manifesto for the equifinality thesis, Journal of hydrology 320 (1-2) (2006) 18–36. doi:https://doi.org/10.1016/j.jhydrol.2005.07.007.

[109] G. Allesina, E. Mussatti, F. Ferrari, A. Muscio, A calibration methodology for building dynamic models based on data collected through survey and billings, Energy and Buildings 158 (2018) 406–416. doi:10.1016/j.enbuild.2017.09.089.

[110] Z. Mylona, M. Kolokotroni, S. Tassou, Frozen food retail: Measuring and modelling energy use and space environmental systems in an operational supermarket, Energy and Buildings 144 (2017) 129–143. doi:10.1016/j.enbuild.2017.03.049.

[111] J. Vesterberg, S. Andersson, T. Olofsson, Calibration of low-rise multi-family residential simulation models using regressed estimations of transmission losses, Journal of Building Performance Simulation 9 (3) (2016) 304–315. doi:10.1080/19401493.2015.1067257.

[112] D. Guyot, F. Giraud, F. Simon, D. Corgier, C. Marvillet, B. Tremeac, Building energy model calibration: A detailed case study using sub-hourly measured data, Energy and Buildings 223 (2020) 110189. doi:https://doi.org/10.1016/j.enbuild.2020.110189.

[113] R. Escandón, R. Suárez, J. Sendra, On the assessment of the energy performance and environmental behaviour of social housing stock for the adjustment between simulated and measured data: The case of mild winters in the mediterranean climate of southern europe, Energy and Buildings 152 (2017) 418–433. doi:10.1016/j.enbuild.2017.07.063.

[114] F. Ascione, N. Bianco, R. De Masi, F. De'Rossi, G. Vanoli, Energy retrofit of an educational building in the ancient center of benevento. feasibility study of energy savings and respect of the historical value, Energy and Buildings 95 (2015) 172–183. doi:10.1016/j.enbuild.2014.10.072.

[115] V. Monetti, E. Fabrizio, M. Filippi, Impact of low investment strategies for space heating control: Application of thermostatic radiators valves to an old residential building, Energy and Buildings 95 (2015) 202–210. doi:10.1016/j.enbuild.2015.01.001.

[116] G. Kazas, E. Fabrizio, M. Perino, Energy demand profile generation with detailed time resolution at an urban district scale: A reference building approach and case study, Applied Energy 193 (2017) 243–262. doi:10.1016/j.apenergy.2017.01.095.

[117] D. Jermyn, R. Richman, A process for developing deep energy retrofit strategies for single-family housing typologies: Three toronto case studies, Energy and Buildings 116 (2016) 522–534. doi:10.1016/j.enbuild.2016.01.022.

[118] H. Samuelson, A. Ghorayshi, C. Reinhart, Analysis of a simplified calibration procedure for 18 design-phase building energy models, Journal of Building Performance Simulation 9 (1) (2016) 17–29. doi:10.1080/19401493.2014.988752.

[119] P. Beagon, F. Boland, M. Saffari, Closing the gap between simulation and measured energy use in home archetypes, Energy and Buildings 224 (2020). doi:10.1016/j.enbuild.2020.110244.

[120] M. Tokarik, R. Richman, Life cycle cost optimization of passive energy efficiency improvements in a toronto house, Energy and Buildings 118 (2016) 160–169. doi:10.1016/j.enbuild.2016.02.015.

[121] Y. Ji, P. Xu, A bottom-up and procedural calibration method for building energy simulation models based on hourly electricity submetering data, Energy 93 (2015) 2337–2350. doi:10.1016/j.energy.2015.10.109.

[122] M. Royapoor, T. Roskilly, Building model calibration using energy and environmental data, Energy and Buildings 94 (2015) 109–120. doi:10.1016/j.enbuild.2015.02.050.

[123] N. Jain, E. Burman, S. Stamp, D. Mumovic, M. Davies, Cross-sectoral assessment of the performance gap using calibrated building energy performance simulation, Energy and Buildings 224 (2020). doi:10.1016/j.enbuild.2020.110271.

[124] A. O' Donovan, P. O' Sullivan, M. Murphy, Predicting air temperatures in a naturally ventilated nearly zero energy building: Calibration, validation, analysis and approaches, Applied Energy 250 (2019) 991–1010. doi:10.1016/j.apenergy.2019.04.082.

[125] F. Pianosi, K. Beven, J. Freer, J. W. Hall, J. Rougier, D. B. Stephenson, T. Wagener, Sensitivity analysis of environmental models: A systematic review with practical workflow, Environmental Modelling & Software 79 (2016) 214–232. doi:https://doi.org/10.1016/j.envsoft.2016.02.008.

[126] M. D. Morris, Factorial sampling plans for preliminary computational experiments, Technometrics 33 (2) (1991) 161–174.

[127] F. Campolongo, J. Cariboni, A. Saltelli, An effective screening design for sensitivity analysis of large models, Environmental modelling & software 22 (10) (2007) 1509–1518.

[128] K. Kim, J. Haberl, Development of a home energy audit methodology for determining energy and cost efficient measures using an easy-to-use simulation: Test results from single-family houses in texas, usa, Building Simulation 9 (6) (2016) 617–628. doi:10.1007/s12273-016-0299-y.

[129] W. Tian, A review of sensitivity analysis methods in building energy analysis, Renewable and Sustainable Energy Reviews 20 (2013) 411–419. doi:http://dx.doi.org/10.1016/j.rser.2012.12.014.

[130] J. Robertson, B. Polly, J. Collis, Reduced-order modeling and simulated annealing optimization for efficient residential building utility bill calibra-

tion, Applied Energy 148 (2015) 169–177. doi:10.1016/j.apenergy.2015.03.049.

[131] R. Enríquez, M. Jiménez, M. Heras, Towards non-intrusive thermal load monitoring of buildings: Bes calibration, Applied Energy 191 (2017) 44–54. doi:10.1016/j.apenergy.2017.01.050.

[132] F. Tüysüz, H. Sözer, Calibrating the building energy model with the short term monitored data: A case study of a large-scale residential building, Energy and Buildings 224 (2020). doi:10.1016/j.enbuild.2020.110207.

[133] K. Kim, J. Haberl, Development of methodology for calibrated simulation in single-family residential buildings using three-parameter change-point regression model, Energy and Buildings 99 (2015) 140–152, cited By 22. doi:10.1016/j.enbuild.2015.04.032.
URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84929469342&doi=10.1016%2fj.enbuild.2015.04.032&partnerID=40&md5=d3beecae33793e99fbc196aa014108ef

[134] A. Elharidi, P. Tuohy, M. Teamah, A. Hanafy, Energy and indoor environmental performance of typical egyptian offices: Survey, baseline model and uncertainties, Energy and Buildings 135 (2017) 367–384. doi:10.1016/j.enbuild.2016.11.011.

[135] B. Glasgo, C. Hendrickson, I. L. Azevedo, Assessing the value of information in residential building simulation: Comparing simulated and actual building loads at the circuit level, Applied Energy 203 (2017) 348–363. doi:http://dx.doi.org/10.1016/j.apenergy.2017.05.164.

[136] I. Allard, T. Olofsson, G. Nair, Energy evaluation of residential buildings: Performance gap analysis incorporating uncertainties in the evaluation methods, Building Simulation 11 (4) (2018) 725–737. doi:10.1007/s12273-018-0439-7.

[137] A. Mihai, R. Zmeureanu, Bottom-up evidence-based calibration of the hvac air-side loop of a building energy model, Journal of Building Performance Simulation 10 (1) (2017) 105–123. doi:10.1080/19401493.2016.1152302.

[138] F. Ascione, N. Bianco, T. Iovane, G. Mauro, D. Napolitano, A. Ruggiano, L. Viscido, A real industrial building: Modeling, calibration and pareto optimization of energy retrofit, Journal of Building Engineering 29 (2020). doi:10.1016/j.jobe.2020.101186.

[139] R. Yin, S. Kiliccote, M. Piette, Linking measurements and models in commercial buildings: A case study for model calibration and demand response strategy evaluation, Energy and Buildings 124 (2016) 222–235. doi:10.1016/j.enbuild.2015.10.042.

[140] N. Li, Z. Yang, B. Becerik-Gerber, C. Tang, N. Chen, Why is the reliability of building simulation limited as a tool for evaluating energy conservation measures?, Applied Energy 159 (2015) 196–205. doi:10.1016/j.apenergy.2015.09.001.

[141] C. Aparicio-Fernández, J.-L. Vivancos, P. Cosar-Jorda, R. Buswell, Energy modelling and calibration of building simulations: A case study of a domestic building with natural ventilation, Energies 12 (17) (2019). doi:10.3390/en12173360.

[142] D. Yan, W. O'Brien, T. Hong, X. Feng, H. B. Gunay, F. Tahmasebi, A. Mahdavi, Occupant behavior modeling for building performance simulation: Current state and future challenges, Energy and Buildings 107 (2015) 264–278. doi:https://doi.org/10.1016/j.enbuild.2015.08.032.

[143] Y.-S. Kim, M. Heidarinejad, M. Dahlhausen, J. Srebric, Building energy model calibration with schedules derived from electricity use data, Applied Energy 190 (2017) 997–1007. doi:10.1016/j.apenergy.2016.12.167.

[144] A. Chong, G. Augenbroe, D. Yan, Occupancy data at different spatial resolutions: Building energy performance and model calibration, Applied Energy 286 (2021) 116492. doi:https://doi.org/10.1016/j.apenergy.2021.116492.

[145] G. Augenbroe, The role of simulation in performance-based building, in: J. L. Hensen, R. Lamberts (Eds.), Building performance simulation for design and operation, Routledge, 2019, Ch. 10, p. 343.

[146] C. A. Aumann, A methodology for developing simulation models of complex systems, Ecological Modelling 202 (3-4) (2007) 385–396. doi:https://doi.org/10.1016/j.ecolmodel.2006.11.005.

[147] H. B. Gunay, W. O'Brien, I. Beausoleil-Morrison, Implementation and comparison of existing occupant behaviour models in energyplus, Journal of Building Performance Simulation 9 (6) (2016) 567–588. doi:http://dx.doi.org/10.1080/19401493.2015.1102969.

[148] S. Kanteh Sakiliba, N. Bolton, M. Sooriyabandara, The energy performance and techno-economic analysis of zero energy bill homes, Energy and Buildings 228 (2020). doi:10.1016/j.enbuild.2020.110426.

[149] A. Figueiredo, J. Kämpf, R. Vicente, R. Oliveira, T. Silva, Comparison between monitored and simulated data using evolutionary algorithms: Reducing the performance gap in dynamic building simulation, Journal of Building Engineering 17 (2018) 96–106. doi:10.1016/j.jobe.2018.02.003.

[150] J. Lee, S. Yoo, J. Kim, D. Song, H. Jeong, Improvements to the customer baseline load (cbl) using standard energy consumption considering energy efficiency and demand response, Energy 144 (2018) 1052–1063. doi:10.1016/j.energy.2017.12.044.

[151] M. De Rosa, M. Brennenstuhl, C. Cabrera, U. Eicker, D. Finn, An iter-

ative methodology for model complexity reduction in residential building simulation, Energies 12 (12) (2019). doi:10.3390/en12122448.

[152] C. Cornaro, S. Rossi, S. Cordiner, V. Mulone, L. Ramazzotti, Z. Rinaldi, Energy performance analysis of stile house at the solar decathlon 2015: Lessons learned, Journal of Building Engineering 13 (2017) 11–27. doi:10.1016/j.jobe.2017.06.015.

[153] E. Carlon, V. Verma, M. Schwarz, L. Golicza, A. Prada, M. Baratieri, W. Haslinger, C. Schmidl, Experimental validation of a thermodynamic boiler model under steady state and dynamic conditions, Applied Energy 138 (2015) 505–516. doi:10.1016/j.apenergy.2014.10.031.

[154] M. Bhandari, S. Shrestha, J. New, Evaluation of weather datasets for building energy simulation, Energy and Buildings 49 (2012) 109–118. doi:https://doi.org/10.1016/j.enbuild.2012.01.033.

[155] A. Persily, A. Musser, S. J. Emmerich, Modeled infiltration rate distributions for us housing, Indoor air 20 (6) (2010) 473–485. doi:https://doi.org/10.1111/j.1600-0668.2010.00669.x.

[156] D. Kim, S. Cox, H. Cho, P. Im, Model calibration of a variable refrigerant flow system with a dedicated outdoor air system: A case study, Energy and Buildings 158 (2018) 884–896. doi:10.1016/j.enbuild.2017.10.049.

[157] S. Nagpal, C. Mueller, A. Aijazi, C. Reinhart, A methodology for autocalibrating urban building energy models using surrogate modeling techniques, Journal of Building Performance Simulation 12 (1) (2019) 1–16. doi:10.1080/19401493.2018.1457722.

[158] M. Manfren, B. Nastasi, Parametric performance analysis and energy model calibration workflow integration - a scalable approach for buildings, Energies 13 (3) (2020). doi:10.3390/en13030621.

[159] S. Asadi, E. Mostavi, D. Boussaa, M. Indaganti, Building energy model calibration using automated optimization-based algorithm, Energy and Buildings 198 (2019) 106–114. doi:10.1016/j.enbuild.2019.06.001.

[160] D. Chakraborty, H. Elzarka, Performance testing of energy models: are we using the right statistical metrics?, Journal of Building Performance Simulation 11 (4) (2018) 433–448. doi:10.1080/19401493.2017.1387607.

[161] G. R. Ruiz, C. F. Bandera, Validation of calibrated energy models: Common errors, Energies 10 (10) (2017) 1587. doi:https://doi.org/10.3390/en10101587.

[162] T. A. Reddy, I. Maor, C. Panjapornpon, Calibrating detailed building energy simulation programs with measured data—part i: General methodology (rp-1051), Hvac&R Research 13 (2) (2007) 221–241. doi:10.1080/10789669.2007.10390952.

[163] J. K. Kissock, J. S. Haberl, D. E. Claridge, Inverse modeling toolkit: numerical algorithms, ASHRAE transactions 109 (2003) 425.

[164] J. S. Haberl, A. Sreshthaputra, D. E. Claridge, J. K. Kissock, Inverse model toolkit: Application and testing, ASHRAE transactions 109 (2003) 435.

[165] EVO, Uncertainty assessment for ipmvp, international performance measurement and verification protocol, Efficiency Valuation Organization (2019).

[166] C. Cerezo, J. Sokol, S. AlKhaled, C. Reinhart, A. Al-Mumin, A. Hajiah, Comparison of four building archetype characterization methods in urban building energy modeling (ubem): A residential case study in kuwait city, Energy and Buildings 154 (2017) 321–334. doi:10.1016/j.enbuild.2017.08.029.

[167] T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation, Journal of the American statistical Association 102 (477) (2007) 359–378. doi:https://doi.org/10.1198/016214506000001437.

[168] A. K. Persily, Airtightness of commercial and institutional buildings: blowing holes in the myth of tight buildings (1998).

[169] F. G. N. Li, A. Smith, P. Biddulph, I. G. Hamilton, R. Lowe, A. Mavrogianni, E. Oikonomou, R. Raslan, S. Stamp, A. Stone, A. Summerfield, D. Veitch, V. Gori, T. Oreszczyn, Solid-wall u-values: heat flux measurements compared with standard assumptions, Building Research & Information 43 (2) (2015) 238–252. doi:10.1080/09613218.2014.967977.

[170] V. Gori, V. Marincioni, P. Biddulph, C. A. Elwell, Inferring the thermal resistance and effective thermal mass distribution of a wall from in situ measurements to characterise heat transfer at both the interior and exterior surfaces, Energy and Buildings 135 (2017) 398–409. doi:ps://doi.org/10.1016/j.enbuild.2016.10.043.

[171] D. H. Yi, D. W. Kim, C. S. Park, Parameter identifiability in bayesian inference for building energy models, Energy and Buildings 198 (2019) 318–328. doi:https://doi.org/10.1016/j.enbuild.2019.06.012.

[172] T. Hong, Y. Chen, X. Luo, N. Luo, S. H. Lee, Ten questions on urban building energy modeling, Building and environment 168 (2020) 106508. doi:https://doi.org/10.1016/j.buildenv.2019.106508.

[173] C. Miller, D. Thomas, J. Kämpf, A. Schlueter, Urban and building multiscale co-simulation: case study implementations on two university campuses, Journal of Building Performance Simulation 11 (3) (2018) 309–321. doi:https://doi.org/10.1080/19401493.2017.1354070.

[174] Y. Chen, T. Hong, X. Luo, B. Hooper, Development of city buildings dataset for urban building energy modeling, Energy and Buildings 183

(2019) 252–265. doi:https://doi.org/10.1016/j.enbuild.2018.11.008.

[175] Y. Q. Ang, Z. M. Berzolla, C. F. Reinhart, From concept to application: A review of use cases in urban building energy modeling, Applied Energy 279 (2020) 115738. doi:https://doi.org/10.1016/j.apenergy.2020.115738.

[176] F. Noardo, L. Harrie, K. A. Ohori, F. Biljecki, C. Ellul, T. Krijnen, H. Eriksson, D. Guler, D. Hintz, M. A. Jadidi, M. Pla, S. Sanchez, V.-P. Soini, R. Stouffs, J. Tekavec, J. Stoter, Tools for BIM-GIS Integration (IFC Georeferencing and Conversions): Results from the GeoBIM Benchmark 2019, ISPRS International Journal of Geo-Information 9 (9) (2020) 502. doi:10.3390/ijgi9090502.

[177] F. Biljecki, J. Lim, J. Crawford, D. Moraru, H. Tauscher, A. Konde, K. Adouane, S. Lawrence, P. Janssen, R. Stouffs, Extending CityGML for IFC-sourced 3D city models, Automation in Construction 121 (2021) 103440. doi:10.1016/j.autcon.2020.103440.

[178] E. J. Rykiel Jr, Testing ecological models: the meaning of validation, Ecological modelling 90 (3) (1996) 229–244. doi:https://doi.org/10.1016/0304-3800(95)00152-2.

[179] M. Trčka, J. L. Hensen, Overview of hvac system simulation, Automation in Construction 19 (2) (2010) 93–99. doi:https://doi.org/10.1016/j.autcon.2009.11.019.

[180] I. Gaetani, P.-J. Hoes, J. L. Hensen, Occupant behavior in building energy simulation: Towards a fit-for-purpose modeling strategy, Energy and Buildings 121 (2016) 188–204. doi:https://doi.org/10.1016/j.enbuild.2016.03.038.

[181] I. Gaetani, P.-J. Hoes, J. L. Hensen, A stepwise approach for assessing the appropriate occupant behaviour modelling in building performance

simulation, Journal of Building Performance Simulation 13 (3) (2020) 362–377. doi:https://doi.org/10.1080/19401493.2020.1734660.

[182] S. Zhan, A. Chong, Data requirements and performance evaluation of model predictive control in buildings: A modeling perspective, Renewable and Sustainable Energy Reviews (2021) 110835doi:https://doi.org/10.1016/j.rser.2021.110835.

[183] S. Pfenninger, J. DeCarolis, L. Hirth, S. Quoilin, I. Staffell, The importance of open data and software: Is energy research lagging behind?, Energy Policy 101 (2017) 211–215. doi:http://dx.doi.org/10.1016/j.enpol.2016.11.046.

[184] M. McNutt, Journals unite for reproducibility (2014). doi:10.1126/science.aaa1724.

[185] C. Boettiger, An introduction to docker for reproducible research, ACM SIGOPS Operating Systems Review 49 (1) (2015) 71–79. doi:https://doi.org/10.1145/2723872.2723882.

[186] S. Pfenninger, L. Hirth, I. Schlecht, E. Schmid, F. Wiese, T. Brown, C. Davis, M. Gidden, H. Heinrichs, C. Heuberger, et al., Opening the black box of energy modelling: Strategies and lessons learned, Energy Strategy Reviews 19 (2018) 63–71. doi:https://doi.org/10.1016/j.esr.2017.12.002.

[187] B. L. Ball, N. Long, K. Fleming, C. Balbach, P. Lopez, An open source analysis framework for large-scale building energy modeling, Journal of Building Performance Simulation 13 (5) (2020) 487–500. doi:https://doi.org/10.1080/19401493.2020.1778788.

[188] H. Jia, A. Chong, eplusr: A framework for integrating building energy simulation and data-driven analytics, Energy and Buildings (2021) 110757doi:https://doi.org/10.1016/j.enbuild.2021.110757.