

Article

Uncertainty's Indices Assessment for Calibrated Energy Models

Vicente Gutiérrez González ^{1,*}, Lissette Álvarez Colmenares ²,
Jesús Fernando López Fidalgo ², Germán Ramos Ruiz ¹ and Carlos Fernández Bandera ¹

¹ School of Architecture, University of Navarra, 31009 Pamplona, Spain; gramrui@unav.es (G.R.R.); cfbandera@unav.es (C.F.B.)

² ICS Statistical Unit, University of Navarra, 31009 Pamplona, Spain; lisettesilva@gmail.com (L.Á.C.); fidalgo@unav.es (J.F.L.F.)

* Correspondence: vgutierrez@unav.es; Tel.: +34-948-425-600 (ext. 800000)

Received: 5 April 2019; Accepted: 27 May 2019; Published: 31 May 2019



Abstract: Building Energy Models (BEMs) are a key element of the Energy Performance of Buildings Directive (EPBD), and they are at the basis of Energy Performance Certificates (EPCs). The main goal of BEMs is to provide information for building stakeholders; they can be a powerful market tool to increase demand for energy efficiency solutions in buildings without affecting the comfort of users, as well as providing other benefits. The next generation of BEMs should value buildings in a holistic and cost-effective manner across several complementary dimensions: envelope performances, system performances, and controlling the ability of buildings to offer flexible services to the grid by optimizing energy consumption, distributed generation, and storage. SABINA is a European project that aims to look for flexibility to the grid, targeting the most economic source possible: existing thermal inertia in buildings. In doing so, SABINA works with a new generation of BEMs that tend to mimic the thermal behavior of real buildings and therefore requires an accurate methodology to choose the model that complies with the requirements of the system. This paper details our novel extensive research on which statistical indices should be chosen in order to identify the best model offered by the calibration process developed by Fernandez et al. in a previous paper and therefore is a continuation of that work.

Keywords: uncertainty index; validation of calibrated energy models; energy simulation; Zero Energy Calibration (ZEC); Building Energy Models (BEMs); law-data-driven BEMs

1. Introduction and Motivation for the Work

BEMs are key elements of the Energy Performance of Buildings Directive, and they are at the basis of Energy Performance Certificates (EPCs) and assessment. Assessment and certification processes should be user-friendly, cost-effective, and more reliable in order to instill trust in investors in the energy efficiency sector [1]. Therefore, the next generation of EPCs will need to fulfill these requirements, as well as the next generation of BEMs. Until now, EPCs have been based on two concepts [2]: standard energy rating and measured energy rating. In the former, the energy consumed by a building is calculated through an energy model (law-driven models, Option D of the International Performance Measurement and Verification Protocol (IPMVP)) [3], and in the latter, the energy is measured through meters and sensors installed in the building (data-driven models, Option C of the IPMVP).

In a previous paper written by some of the authors [4], it was explained in detail how this new generation of BEMs should be produced and that the new technique is able to merge the law-driven models [5] and the data-driven models [6–9], resulting in “law-data-driven models”. In summary, the concept uses well-known software such as EnergyPlus [10] to combine the model based on as-built

parameters with the model based on parameters estimated using measurements of the system and, through a calibration process, producing the new technique. This technique has produced very good results and is based on the use of measured temperature from the real building as part of the energy balance of the BEM, following the idea that Sonderegger postulated in 1977: *“Instead of telling the computer how the building is built and asking it for the indoor temperature, one tells the computer the measured indoor temperature and asks it for the building parameters”* [11].

SABINA is a project that is looking for services on the grid based on the “demand response” concept [12] and the idea of increasing the amount of renewable energy consumed locally by buildings. To reach the EU’s long-term objectives for reducing greenhouse gas emissions, this share should reach more than 30% in 2030, and almost 50% in some scenarios in 2050 [13]; new management systems are thus required. What is most needed is additional flexibility in the system. SABINA targets the most economic source possible: existing thermal inertia in buildings [14]. This goal requires models that capture the thermal dynamics of the building, and the Zero Energy Calibration (ZEC) methodology has been chosen to select those kinds of models [4]. The usefulness of a model depends on the accuracy and reliability of its output, but all models are imperfect abstractions of reality, because there is imprecision and uncertainty associated with any model.

Currently, there is the protocol IPMVP [3], and two guidelines: FEMP [15] and ASHRAE [16], which offer a set of error indices ($CV(RMSE)$, $NMBE$, and R^2) to evaluate the quality of the calibrated models considering the monthly and hourly energy consumption (simulated vs. real). Other methodologies use indoor air temperature (simulated vs. real) to calibrate the building models with the same indices [4,17–20]. When doing so, it is not clear if these indices, which were selected for energy evaluation, will have a good performance for temperature. In this paper, a large number of error indices have been analyzed with the aim of selecting the best ones to choose the model that represents the real building indoor air temperature. This new evaluation methodology has been tested and verified in different building models: the “Amigos” [4], “Humanities” [21], and “The School of Architecture” [22] at the Pamplona Campus of the University of Navarre. In this paper, the office building of the School of Architecture has been used, as it is explained in the following sections.

Summary of the ZEC Methodology

The Zero Energy Calibration (ZEC) is a methodology for building envelope calibration. The ZEC principle is based on the idea that when introducing the free oscillation temperature of a building in the model, as a dynamic set-point, the energy consumed by the HVAC equipment in that period should be zero. If this is not the case, the reason should be the wrong configuration of the building parameters, and the algorithm (genetic algorithm) will look for a new vector of envelope parameters that will produce a lower energy consumption (heating plus cooling). The process finishes when the energy (the objective function) cannot be reduced further and the model envelope is calibrated.

In most automatic calibration techniques [23,24], the simulation data are used at the end of the process to be compared with the measured data, and the goal is to minimize an error value in what is known as uncertainty analysis. In such cases, the statistical indices ($CV(RMSE)$, $NMBE$, and R^2) are the objective functions that will guide the algorithm in the search for the calibrated model [25–27].

The main ways of calibration do not allow entering into the calibration process as many measured data as necessary, and thus, the thermal characterization of the model will not be improved. In this methodology (ZEC), there is no restriction in the creation of thermal zones. The major simplification that ZEC offers is that there is no implementation of uncertainty analysis in coordination with the automatic calibration algorithm and the simulation program, which makes it simpler and therefore more accessible to all kinds of professionals with energy simulation skills, but without programming capabilities.

For this reason, the ZEC methodology is simple in execution. The algorithm used to perform the thermal zone energy balance in EnergyPlus is the Conduction Transfer Function (CTF), which offers a very fast and elegant solution to solve the Fourier differential equation and to find the temperature

of the thermal zone. However, as explained in the EnergyPlus Engineering Reference, “conduction transfer function series become progressively more unstable as the time step decreases. This became a problem as investigations into short time step computational methods for the zone/system interactions progressed because, eventually, this instability caused the entire simulation to diverge” [28]. This divergence is translated into extra energy consumption that affects the objective function used by ZEC, a problem that has been well documented and evaluated by Wetter et al. [29]. The result of this extra energy consumption is that some models with slightly higher energy consumption have better uncertainty temperature results than the best models selected by the energy of the objective function. From a practical point of view, this means that the best model cannot be chosen directly from the results offered by the algorithm unless an uncertainty temperature analysis is subsequently performed, in the same way as other similar works [26,27,30,31].

Taking into account the indices’ combination proposed by ASHRAE ($CV(RMSE)$, $NMBE$, and R^2) [16], the authors worked with a new statistical index that was called the ZEC_Index [4], which was the arithmetic sum of errors $CV(RMSE)$, $NMBE$, and $(1 - R^2)$. The model with the lowest ZEC_index was the one considered to have the best performance. As the indices’ combination proposed by ASHRAE ($CV(RMSE)$, $NMBE$, and R^2) is based on energy uncertainty analysis and the new proposal is based on temperature uncertainty analysis, this paper intends to confirm if there is any other statistical index or combination of indices that can improve the selection of the best model.

The uncertainty analysis should classify the best models according to the capacity to reduce the error between real temperature inside the building and simulated temperature produced by the building model. From a practical point of view, a good correlation should be found between temperature and energy with respect to the selected temperature error (uncertainty index).

In order to check if a different index can choose a better model, a list of a number of error metrics that been studied in Section 2, classified into seven groups (bias error indices, uncertainty indices based on absolute deviations, uncertainty indices based on square deviations, goodness-of-fit metrics, efficiency criteria, indices for model discrimination, and proximity measures), according to the application or structure and the statistical methodology description used to select the metrics that identify the best-adjusted calibrated energy model. Section 3 outlines the cases of study and the description of the building used to check the methodology. Section 4 presents the performance of the metrics described in Section 2 over two case studies: a synthetic energy model and a real building model, each under the same conditions. The conclusions that we have reached in this paper and future research considerations are presented in Section 5.

2. List of a Number of Error Metrics

Having a “reasonable” idea of the quality of adjustment between real and simulated models is not hard [32], but evaluating the accuracy of a BEM or quantifying the quality of the adjustment is, actually, quite difficult, particularly when this quantification is used to identify the best adjusted model to the real building.

Different indices are used in diverse research branches to define an evaluation criterion for the accuracy of the energy model. For example, efficiency measures are used in hydrology [33,34]. In order to evaluate the performance of the model for energy saving in the Measurement and Verification process (M&V), Goodness-of-fit metrics are generally used [35]. Another measure, known as the uncertainty index, is used in an energy modeling context for the same purpose [36].

Each metric provides a different insight into the model’s performance, and therefore, there is no an ideal metric to identify the best-adjusted model. In fact, researchers suggest “to use the numerical comparison as well as graphical comparison when one decides the base model adequacy” [37]. In practice, several metrics are jointly evaluated and complemented by a graphical analysis (e.g., [38–41]).

2.1. Bias Error Indices

The range for all of these is the whole real line, and the optimal value must be zero (Table 1). The M&V methodology for energy calibrated models considers $p = 1$ for *NMBE*.

Table 1. Bias error measures.

Index	Equation	
Bias Error	$BE = \sum_{i=1}^n (y_i - \hat{y}_i)$	(1)
Mean Bias Error	$MBE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$	(2)
Relative Error	$RE = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)}{y_i}$	(3)
Normalized Mean Bias Error	$NMBE = \frac{1}{\bar{y}(n-p)} \sum_{i=1}^n (y_i - \hat{y}_i)$	(4)
<i>PBIAS</i>	$PBIAS = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{\sum_{i=1}^n y_i} \times 100$	(5)

2.2. Uncertainty Indices Based on Absolute Deviations

These indices consider only the distance between values, omitting the direction of the differences and overcoming cancellation errors. They can take any positive value, and their optimal value is minimum (Table 2).

Table 2. Absolute error measures.

Index	Equation	
Absolute Error	$AE = \sum_{i=1}^n y_i - \hat{y}_i $	(6)
Mean Absolute Error	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	(7)
Relative Absolute Error	$RAE = \sum_{i=1}^n \frac{ y_i - \hat{y}_i }{y_i}$	(8)
Mean Absolute Percent Error	$MAPE = \frac{1}{\bar{y}} \sum_{i=1}^n \frac{ y_i - \hat{y}_i }{y_i} \times 100\%$	(9)
<i>E_{max}</i>	$E_{max} = \max_{1 \leq i \leq n} y_i - \hat{y}_i $	(10)

2.3. Uncertainty Indices Based on Square Deviations

In square deviation measures (Table 3), the M&V methodology for energy calibrated models considers $p = 1$ for *CV(RMSE)*.

Table 3. Square deviations measures.

Index	Equation	Range	Optimal value
Mean Squared Error	$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$	(11) $[0, \infty)$	0
Root Mean Squared Error	$RMSE = \left[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right]^{\frac{1}{2}}$	(12) $[0, \infty)$	0
RMSE-observation Standard Deviationratio	$RSR = \left[\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \right]^{\frac{1}{2}}$	(13) $[0, \infty)$	0
Coefficient of Variation of RMSE	$CV(RMSE) = \frac{1}{\bar{y}} \left[\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p} \right]^{\frac{1}{2}}$	(14) $[0, \infty)$	0
$RMSE/MAE$	$\frac{RMSE}{MAE} = \frac{\left[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right]^{\frac{1}{2}}}{\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i }$	(15) $[1, \sqrt{n}]$	1

2.4. Goodness-of-Fit Metrics

The uncertainty in energy calibrated models is directly related to their goodness-of-fit [42] and is the reason why they are the most popular measures to establish the fitness of a simulated model (Table 4). They measure the quality of the linear relationship between the simulated and observed data. This relationship may be quite strong, but with a substantial bias. Thus, these measures may be completed with the bias measures. It may be said that uncertainty can be assessed with a couple of measures, one of the goodness-of-fit and one of bias.

Table 4. Goodness-of-fit measures.

Index	Equation	Range	Optimal Value
Pearson Correlation Coefficient	$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\left(\sum_{i=1}^n (y_i - \bar{y})^2 \right) \left(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \right)}}$	(16) $[-1, 1]$	$ r = 1$
Spearman Correlation Coefficient	$\rho = \frac{\sum_{i=1}^n (rg(y_i) - \overline{rg(y)})(rg(\hat{y}_i) - \overline{rg(\hat{y})})}{\sqrt{\left(\sum_{i=1}^n (rg(y_i) - \overline{rg(y)})^2 \right) \left(\sum_{i=1}^n (rg(\hat{y}_i) - \overline{rg(\hat{y})})^2 \right)}}$	(17) $[-1, 1]$	$ \rho = 1$
Coefficient of Determination	$R^2 = r^2$	(18) $[0, 1]$	1
bR^2	$bR^2 = \begin{cases} b R^2 & \text{if } b \leq 1 \\ R^2/ b & \text{if } b > 1 \end{cases}$	(19) $[0, \infty)$	1
GoF*	$GOF = \left[\frac{1}{2} (CV(RMSE)^2 + NMBE^2) \right]^{1/2}$	(20) $[0, 1]$	0
Index ZEC	$ NMBE + CV(RMSE) + (1 - R^2)$	(21) $[0, \infty)$	0
Ratio of Standard Deviations	$rSD = \left(\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2} \right)^{\frac{1}{2}}$	(22) $[0, \infty)$	1

2.5. Efficiency Criteria

This indices are measures of how well a model simulation fits the real observations [43], and they are widely used, for instance, to evaluate the performance of hydrological models [33,34]. Most of the efficiency criteria include notions of distance and variance between real and simulated values in order to analyze the adjustment, both in terms of location and variability. Table 5 shows the indices considered here.

Table 5. Efficiency criteria.

Index	Equation	Range	Optimal Value
Nash–Sutcliffe efficiency	$NSE = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$	(23) $(-\infty, 1]$	1
Modified NSE	$mNSE = 1 - \frac{\sum_{i=1}^n y_i - \hat{y}_i ^j}{\sum_{i=1}^n y_i - \bar{y} ^j}$	(24) $(-\infty, 1]$	1
Relative NSE	$rNSE = 1 - \frac{\sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}{\sum_{i=1}^n \left(\frac{y_i - \bar{y}}{\bar{y}}\right)^2}$	(25) $(-\infty, 1]$	1
Logarithmic NSE	$\log NSE = 1 - \frac{\sum_{i=1}^n (\log y_i - \log \hat{y}_i)^2}{\sum_{i=1}^n (\log y_i - \log \bar{y})^2}$	(26) $(-\infty, 1]$	1
Index of Agreement	$d = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (\hat{y}_i - \bar{y} + y_i - \bar{y})^2}$	(27) $[0, 1]$	1
Modified Index of agreement	$md = 1 - \frac{\sum_{i=1}^n y_i - \hat{y}_i ^j}{\sum_{i=1}^n (\hat{y}_i - \bar{y} + y_i - \bar{y})^j}$	(28) $[0, 1]$	1
Relative Index of Agreement	$rd = 1 - \frac{\sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}{\sum_{i=1}^n \left(\frac{ \hat{y}_i - \bar{y} + y_i - \bar{y} }{\bar{y}}\right)^2}$	(29) $(-\infty, 1]$	1
Coefficient of Persistence	$cp = 1 - \frac{\sum_{i=2}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n-1} (y_{i+1} - y_i)^2}$	(30) $(-\infty, 1]$	1
Volumetric Efficiency	$VE = 1 - \frac{\sum_{i=1}^n y_i - \hat{y}_i }{\sum_{i=1}^n y_i}$	(31) $[0, 1]$	0

2.6. Indices for Model Discrimination

The Akaike Information Criterion (*AIC*) and Bayesian Information Criterion (*BIC*) are used for choosing among different models, as Table 6 indicates.

Table 6. Comparison model measures. In the equations, *d* represents the number of parameters of the model.

Index	Equation	Range	Optimal Value
Akaike Information Criterion	$AIC = n \log(MSE) + 2d$	(32) $\mathbb{R}$	lower value
Bayesian Information Criterion	$BIC = n \log(MSE) + d \log(n)$	(33) $\mathbb{R}$	lower value

2.7. Proximity Measures

p – factor (34) is the proportion of simulated values that are within a given band of the observed values and, therefore, takes values in $[0, 1]$.

$$p - factor = \frac{A}{n} \quad (34)$$

$$A = \{\hat{y}_i \in [y_i - \lambda, y_i + \lambda], i = 1, \dots, n\}$$

If the simulated model represents exactly the behavior of the real model, $p - factor = 1$ for small λ -wide uncertainty bounds, which is equivalent to obtaining 100% of observations within the uncertainty band [44]. Therefore, the model error could be referenced by $(1, p - factor)$. A perfect adjustment of models has a $p - factor$ equal to one.

This measure tries to catch the graphical behavior of the adjustment. In the equation, A is the uncertainty λ -wide band, with width equal to 2λ . The choice of λ is not an easy task in general and depends on the data. In this work, λ corresponds to degrees Celsius, which will be chosen in such a way that it correlates well with both uncertainty and energy. We initially propose $\lambda = 0.5$ as a quality criterion, which can collect random and measured errors. In practice, $p - factor$ is widely used by researchers to validate model adjustment [45].

3. Cases Studies, Building Description, and Models' Preparation

In order to carry out this methodology, the calibration process was checked under two assumptions. In Section 4, two cases were developed. In the first case, the real data were produced synthetically from a BEM, as recommended by the *ASHRAE Fundamentals Handbook* [35], with the idea of avoiding the inaccuracy of the temperature meters. In this case, the quality of the parameters resembled quite faithfully the parameters of the model that originated the data. In the second case, the model was calibrated with real data from meters inside the building. On this occasion, the gap between real and simulated data was clearer, as will be seen in the results.

The building selected for generating both case studies explored in this paper was the Architecture School administrative building of the University of Navarra (Figure 1).



Figure 1. The Architecture School administrative building of the University of Navarra.

The Architecture School was designed by the architects Rafael Echaide, Carlos Sobrini, and Eugenio Aguinaga and was built between 1974 and 1978. It won the “National Award for Architecture in Brick” in 1980. The building is organized along an interior garden with four zones at different levels that accommodate the needs of the school.

Through a transparent gallery, connected to the main building, people can access the office area, which is the building object of this paper. It is mainly used as an administration building and by postgraduate students of the different master’s programs of the School of Architecture, and it mainly keeps business hours.

It is a freestanding single-story building of almost 760 square meters. It is a porticoed structure of concrete, and the interior and exterior walls were made of red clinker brick fabric, while the building frames were made in situ of aluminum with an air chamber and a light gold color.

The space allocation consists of a succession of offices for personnel that face southeast and northwest, an administration zone facing northwest, an open working space and master classrooms facing southeast, and a corridor in the middle connecting the spaces.

The building energy model has been divided into 25 thermal zones, one for each room (Figure 2). The HVAC system has been introduced through the option of ideal loads offered by EnergyPlus.

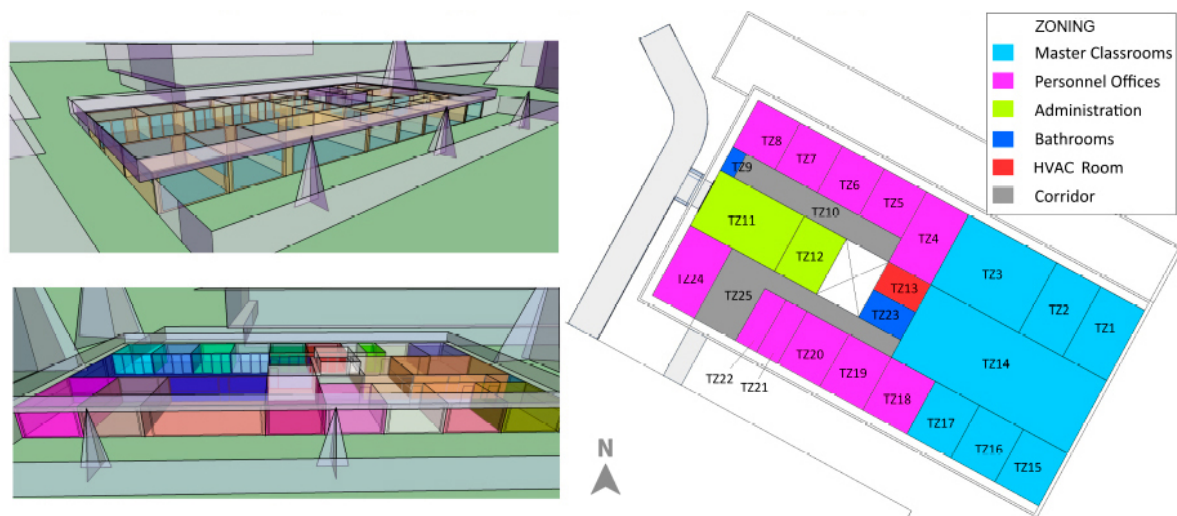


Figure 2. Energy model and zoning map of the Office building, School of Architecture (University of Navarra).

The calibration methodology was carried out by ZEC, described in the previous paragraphs, the process of which is defined in Figure 3.

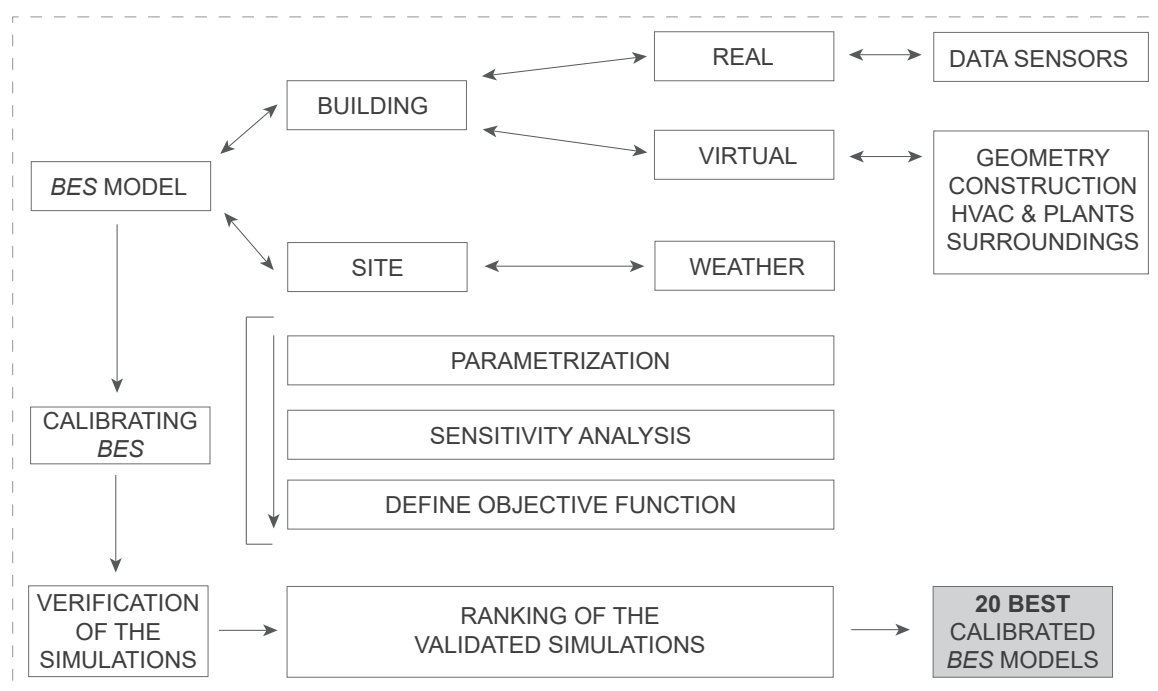


Figure 3. ZEC calibration methodology and model evaluation.

The last step in the ZEC methodology after calibration is to obtain the 20 best models of the calibration process for each period.

4. Methodology to Evaluate Energy Models: Analysis of Case Studies

For evaluation of the models, a global checking period is defined to validate the best models of each calibration period. In the ZEC methodology, the evaluation involves performing an uncertainty

analysis that compares the simulated temperature during the free oscillation times of the checking period with the measured temperature from the real building. This allows the analysis of all the models on equal terms to generate a ranking of simulations in order to choose the best solution (Figure 4).

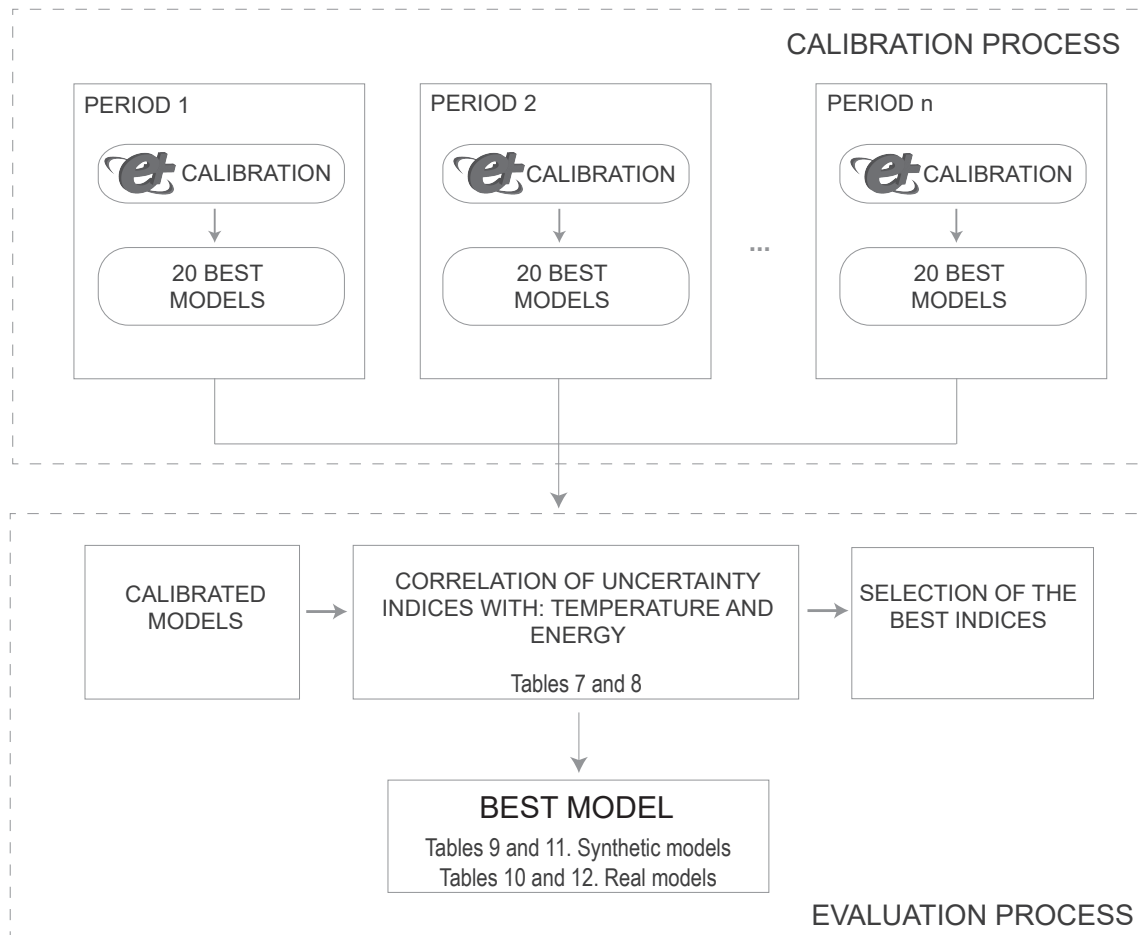


Figure 4. Calibration and evaluation process by ZEC for the different calibration periods.

To carry out this research, the model has been calibrated, using the ZEC methodology, in 16 different calibrated periods choosing the 20 best models, with the lower energy of each period, generating a total of 320 models. The models have been identified by Pk_Mj , where Pk is the calibration period (from 1–16) and Mj is the model with respect to its position in the energy ranking (from 1–20). These models are evaluated in a common checking period, obtaining the results of their indices of uncertainty and the energy consumed. This study was conducted for both a model with synthetic data and a model with real data.

In the following section, a methodology will be developed to choose the best model among these 320. The methodology proposed a correlation analysis and was performed between the uncertainty indices described in Section 2 and the energy consumption and measured temperature. The energy consumption was calculated from 320 simulations checked over the same period, corresponding to the BEM described in Section 3. The uncertainty indices were calculated using the measured temperatures from a synthetic and a real model in 25 thermal zones.

The correlation was calculated over the mean temperature of these 25 zones, and the real and simulated temperatures were weighted for the relative volume of every zone. Thus, the real (Y^p) and simulated ($\hat{Y}^p$) mean weighted temperature vectors were defined as:

$$Y^p := \sum_{j=1}^{25} y_{i,j} \frac{V_j}{\sum_j V_j}, \quad \hat{Y}^p := \sum_{j=1}^{25} \hat{y}_{i,j} \frac{V_j}{\sum_j V_j} \quad (35)$$

where $y_{i,j}$ and $\hat{y}_{i,j}$ are the real and simulated temperatures of the thermal zone j at time i and V_j is the j -thermal zone volume in cubic meters (35).

For a given model, the best uncertainty indices must have the higher p – factor for small values of λ ; therefore, the correlation between error indices and p – factor(λ) could help to identify them. The highest values of these correlations are reached for narrow λ -wide bands, shown in Table 7 for a synthetic model and in Table 8 for a real model.

Another relevant point to determine which indices are appropriate for the best-performing model is the correlation between them and energy consumption. The right column of Tables 7 and 8 shows the calculated values.

Table 7. (Synthetic case) Group index segmentation by the λ value where the uncertainty indices calculated on weighted mean temperatures and p – factor(λ) reach the maximum correlation, in absolute value. The calculated p -values for these correlations are above 0.97 in most cases. The considered λ values are taken within the interval [0.05, 2].

Uncertainty Index (UI)	Maximum Correlation between UI and p -Factor(λ)	Reached at λ	Correlation between UI and Consumed Energy
GROUP 1			
VE	0.9907	0.25	0.9055
MAE	0.9907	0.25	0.9055
mNSE	0.9906	0.25	0.9055
md	0.9895	0.25	0.9089
MAPE	0.9886	0.25	0.9115
GROUP 2			
CV(RMSE)	0.9923	0.30	0.9001
NSE	0.9923	0.30	0.9001
cp	0.9923	0.30	0.9001
RSR	0.9923	0.30	0.9001
d	0.9918	0.30	0.9015
rNSE	0.9897	0.30	0.9056
logNSE	0.9896	0.30	0.9061
ZEC_index	0.9898	0.30	0.8933
bR^2	0.9779	0.30	0.8633
GoF	0.9906	0.30	0.8818
r.Spearman	0.9206	0.30	0.9274
rSD	0.7720	0.30	0.6035
GROUP 3			
rd	0.9896	0.40	0.9081
E _{max}	0.9660	0.40	0.9276
R^2	0.9177	0.40	0.9611
AIC	0.9923	0.30	0.9001
RMSE/MAE	0.6999	0.05	0.5411
PBIAS%	0.5802	1.35	0.1628
NMBE	0.5802	1.35	0.1628

Table 8. (Real case) Group index segmentation by the λ value where the uncertainty indices calculated on weighted mean temperatures and $p - factor(\lambda)$ reach the maximum correlation, in absolute value. The calculated p -values for these correlations are above 0.95 in most case. The considered λ values were taken within interval [0.05, 2].

Uncertainty Index (UI)	Maximum Correlation between UI and p -Factor(λ)	Reached at λ	Correlation between UI and Consumed Energy
GROUP 1			
VE	0.9811	1.50	0.8006
MAE	0.9811	1.50	0.8007
mNSE	0.9811	1.50	0.8007
md	0.9831	1.50	0.8056
MAPE	0.9834	1.45	0.8220
GROUP 2			
CV(RMSE)	0.9905	1.45	0.8275
NSE	0.9905	1.45	0.8275
cp	0.9905	1.45	0.8275
RSR	0.9905	1.45	0.8275
d	0.9907	1.45	0.8259
rNSE	0.9947	1.30	0.8712
logNSE	0.9968	1.20	0.9137
ZEC_index	0.9716	1.50	0.7999
bR ²	0.9456	1.60	0.7575
GoF	0.9897	1.40	0.8291
r.Spearman	0.9653	0.45	0.5900
rSD	0.8189	1.00	0.8556
GROUP 3			
rd	0.9941	1.35	0.8575
E _{max}	0.9877	1.75	0.7260
R ²	0.9484	0.45	0.6663
AIC	0.9904	1.45	0.8275
RMSE/MAE	0.9155	0.30	0.4085
PBIAS%	0.9635	0.60	0.5335
NMBE	0.9635	0.60	0.5335

In both cases, there are groups of indices differentiated by the λ -value where they reach the maximum correlation:

- In the first group, the indices whose maximum correlation is reached at $\lambda = 0.25$ for a synthetic model and $\lambda = 1.45$ for a real model are measures calculated by the absolute value of the distances.
- The second group reaches the maximum at $\lambda = 0.3$ and $\lambda = 0.45$ for a synthetic model and real model, respectively, and they are calculated with squared distances.
- In the third group, the value of λ varies from $\lambda = 0.05$ to $\lambda = 1.35$ for a synthetic model and from $\lambda = 0.3$ to $\lambda = 1.75$ for a real model. They are not related to a specific distance measure.

The indices *BE*, *MBE*, and *RE* are omitted in the results tables since their method of calculation was subjected to cancellation errors and their performance was poor. Another group of indices, *AE*, *RAE*, *MSE*, *RMSE*, *BIC*, and the Pearson Correlation Coefficient (*r*), were omitted because they had redundant information; that is, they were a direct part of the calculation of other measures with equal or better performance, and their correlations were equal to one in the temperature datasets analyzed here. *MSE* and *RMSE* had a similar performance to *CV(RMSE)*, and the same happens between the indices pairs: *AE* and *MAE*, *RAE* and *MAPE*, *BIC* with respect to *AIC*, and the Pearson correlation coefficient with respect to the Spearman correlation coefficient.

With the results obtained in the previous Tables 7 and 8, we can make the selection of the indices for the evaluation of the models. These indices must meet two objectives: good correlation of the indices with temperature and with energy. *logNSE* is one of the indices that fulfills these two premises for a synthetic model and a real model.

Once the indices with which to evaluate the models have been chosen, we proceed to compare them with the old methodology (*ZEC_index*) for the proposed cases: synthetic and real.

In Table 9 (synthetic case) and Table 10 (real case), we have the twenty best models ranked by *ZEC_index* (old methodology). In the first one (synthetic case), thirteen out of twenty of these models

are among the best of the energy ranking. The best model of Table 9 (P13_M10) was the twenty fifth in the energy ranking.

Table 9. (Synthetic case) On the left of the table, ranking ascending by *ZEC_Index* and with reference to its energy position. On the right, ranking ascending by energy. The shaded text corresponds to the 20 best energy models.

Model	<i>ZEC_Index</i>	Energy Ranking	Model	Energy Ranking
P13_M10	1	25	P5_M7	1
P13_M20	2	35	P5_M1	2
P5_M4	3	10	P5_M2	3
P5_M3	4	5	P5_M9	4
P13_M15	5	27	P5_M3	5
P5_M8	6	8	P13_M1	6
P5_M6	12	12	P5_M5	7
P13_M4	8	14	P5_M8	8
P5_M9	9	4	P5_M10	9
P5_M5	10	7	P5_M4	10
P5_M17	11	16	P5_M13	11
P13_M1	12	6	P5_M6	12
P13_M6	13	22	P5_M14	13
P5_M10	14	9	P13_M4	14
P5_M2	15	3	P5_M15	15
P5_M14	16	13	P5_M17	16
P5_M16	17	20	P5_M11	17
P13_M9	18	32	P5_M12	18
P13_M13	19	29	P5_M19	19
P13_M8	20	30	P5_M16	20

In the second one (real case), three out of twenty of these models were among the best of the energy ranking. The best model of Table 10 (P10_M2) was the twenty ninth in the energy ranking.

Table 10. (Real case) On the left of the table, ranking ascending by *ZEC_Index* and with reference to its energy position. On the right, ranking ascending by energy. The shaded text corresponds to the 20 best energy models.

Model	<i>ZEC_Index</i>	Energy Ranking	Model	Energy Ranking
P10_M2	1	29	P5_M6	1
P13_M10	2	19	P9_M8	2
P13_M5	3	30	P5_M1	3
P13_M12	4	39	P16_M3	4
P13_M3	5	45	P16_M5	5
P13_M4	6	46	P5_M7	6
P13_M18	7	31	P5_M12	7
P10_M3	8	84	P5_M3	8
P16_M4	9	10	P5_M19	9
P13_M9	10	47	P16_M4	10
P9_M8	11	2	P5_M10	11
P13_M1	12	62	P9_M14	12
P10_M6	13	89	P9_M6	13
P13_M2	14	56	P9_M5	14
P14_M4	15	52	P6_M17	15
P14_M1	16	63	P5_M8	16
P13_M11	17	81	P9_M1	17
P13_M14	18	75	P9_M3	18
P9_M1	19	23	P13_M10	19
P9_M2	20	70	P6_M10	20

With the new methodology, the results obtained for the synthetic and real case can be evaluated in Tables 11 and 12. In these tables, the twenty best models are ordered by index *logNSE*. For a synthetic case, Table 11, we can see how seventeen out of twenty of these models were among the best for the energy ranking. The best model of Table 11 (P5_M4) was the tenth in the energy ranking and the

number one in the rest of the indices. For the real case, Table 12, ten out of twenty of these models were among the best of the energy ranking. The best model of Table 10 (P9_M8) was the second in the energy ranking and the number one in the rest of the indices.

Table 11. (Synthetic case) Ranking ascending by $\log NSE$. The text in bold corresponds to the models with the lowest energy consumption. The row shaded in gray corresponds to the best model selected by the old methodology.

Model	Energy	MAPE	rNSE	logNSE	rd	p-Factor (0.20)
P5_M4	10	1	1	1	1	100.0%
P5_M3	5	2	2	2	2	100.0%
P5_M8	8	3	3	3	3	100.0%
P5_M6	12	4	4	4	4	100.0%
P5_M5	7	6	5	5	5	100.0%
P5_M9	4	5	6	6	6	100.0%
P5_M17	16	7	7	7	7	100.0%
P5_M2	3	9	8	8	8	100.0%
P5_M14	13	10	9	9	9	99.8%
P5_M16	20	12	10	10	10	99.9%
P5_M10	9	8	11	11	11	100.0%
P13_M15	27	13	12	12	12	99.9%
P5_M7	1	11	13	13	13	100.0%
P5_M11	17	17	14	14	14	99.9%
P13_M10	25	18	15	15	15	99.7%
P5_M18	21	19	16	16	18	100.0%
P13_M4	14	15	17	17	16	99.5%
P13_M1	6	14	18	18	17	99.3%
P5_M19	19	22	19	19	19	99.8%
P5_M12	18	23	20	20	20	99.9%

Table 12. (Real case) Ranking ascending by $\log NSE$. The text in bold corresponds to the models with lowest energy consumption. The row shaded in gray corresponds to the best model selected by ZEC_index.

Model	Energy	MAPE	rNSE	logNSE	rd	p-Factor (1.00)
P9_M8	2	1	1	1	1	90.1%
P9_M14	12	5	2	2	3	87.5%
P16_M5	5	24	6	3	7	88.9%
P16_M4	10	3	3	4	2	87.0%
P9_M6	13	12	5	5	5	86.9%
P9_M5	14	11	4	6	6	86.9%
P9_M3	4	66	12	7	8	88.5%
P16_M7	37	29	11	8	13	87.8%
P10_M2	29	2	7	9	4	84.7%
P9_M4	17	16	9	10	10	84.5%
P9_M3	18	15	10	11	11	84.5%
P13_M18	31	4	8	12	9	83.3%
P9_M12	22	21	13	13	14	83.1%
P9_M9	24	19	15	14	15	82.5%
P13_M10	19	17	14	15	12	82.1%
P16_M20	38	25	18	16	16	82.7%
P16_M18	48	51	20	17	25	84.1%
P9_M11	27	23	17	18	17	82.8%
P9_M13	32	20	16	19	23	82.5%
P16_M17	68	40	28	20	30	82.8%

Depending on the methodology used to choose the best model, the results obtained are different.

For the case of synthetic models, if we rely on the old methodology, the best selected model is P13_M10 and would be ranked twenty fifth in the energy ranking. With the new methodology, the best model was P5_M4 and it had the tenth position in the energy ranking. Analyzing both models, we can see that the model ranked by $\log NSE$ (new methodology) had a better performance with respect to the temperature curves, as shown by its p -factor, 100% with a $\lambda = 0.2$; while the model chosen with ZEC_index (old methodology) had a p -factor for a $\lambda = 0.2$ of 99.7 %.

The same situation would occur if we analyze the real case. The best model selected with ZEC_index (old methodology) was P10_M2 with a position of 29 in the energy ranking, and if we ranked by $\log NSE$ (new methodology), the best model would be P9_M8, being the Number 2 model in the energy ranking. By carefully examining both models, we can conclude that the model classified with the new methodology was better than the one chosen by ZEC_index , as shown by its p -factor. The model P9_M8 had a p -factor for a $\lambda = 1$ of 90.1 %, while for the model P10_M2, it was 84.7 %.

Choosing the best model from a list of calibrated models is crucial in many applications like model predictive control (MPC), where the optimization is based on an hour by hour control of the energy demand of the model in order to reach the goals of the objective function that are related to an increase/decrease of energy consumption during specific time periods. Therefore, having a reliable methodology that gives us this result is paramount. The variation of energy that the two selected models have for the real case, using the old and the new methodology, is significant, as can be seen in Figures 5 and 6, where the accumulated energy at hourly time steps has been represented for heating and cooling demand.

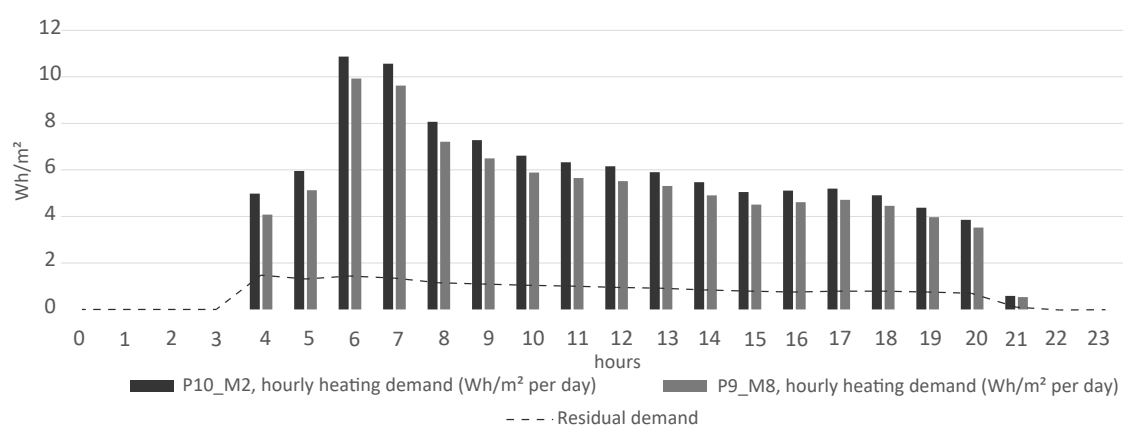


Figure 5. Comparison of heating demand in Wh/m² per day in the months of April and May of 2017 of the models evaluated with the old methodology: ZEC_index (P10_M2) and with $\log NSE$ (P9_M8).

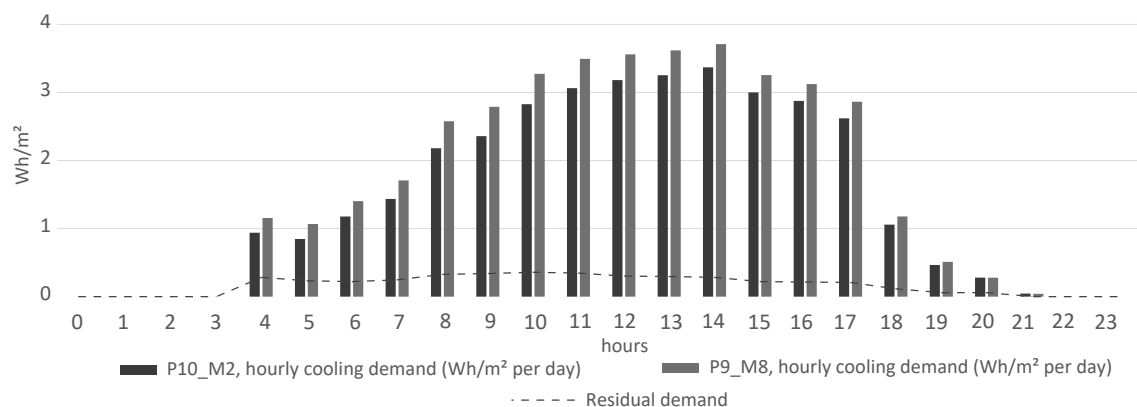


Figure 6. Comparison of cooling demand in Wh/m² per day in the months of June and July of 2017 of the models evaluated with the old methodology: ZEC_index (P10_M2) and with $\log NSE$ (P9_M8).

5. Conclusions and Future Research

After obtaining the results of the cases described in the previous sections, it is clear that a single index is not enough in order to select the best model. $\log NSE$, $rNSE$, rd , $MAPE$, $CV(RMSE)$, cp , and $p - factor$ seem to be the best group of indices to find the best model in both case studies: the synthetic model and the real model. An agreement between all the indices would be desirable in order to choose the best model, as has been demonstrated in this study. In the case of the $p - factor$, this index not only helps to rank the models, but also can be used as a measure to quantify the quality of the model. This value ($p - factor$) demonstrates the actual gap between the calibration process carried out with synthetic data or with real data.

The results that were presented within Section 4 show that the chosen indices based on MSE ($\log NSE$, $rNSE$, rd , $CV(RMSE)$, cp) worked well for time series of temperature data. We estimate that they would still work well in more general scenarios, but some other indices given in this paper could appear to be more appropriate for particular situations. The index $\log NSE$ had the best correlation between energy and temperature with respect to uncertainty indices in the real case and good performance in the synthetic case. This index was computed after a log-transformation of the data; then it was one minus the ratio between the MSE and the difference between the log of the mean and the mean of the log of the observed temperatures. The $rNSE$ index was again based on a ratio of the MSE and, now, the square of the coefficient of variation of the observed temperatures. Index rd was the ratio between the relative MSE and a kind of MSE comparing the observed and simulated temperatures to the mean of the observed temperatures. The $CV(RMSE)$ index was based again on the square of the cover the mean. The cp index was also based on the MSE , now controlled by the consecutive jumps of temperatures. This is especially interesting since it was the only index that took into account the possible correlation between near measures in time. The $MAPE$ index was a relative absolute error index. It can be seen that most of these indices were based on an appropriate ratio of the MSE . Finally, the $p - factor$ was rather intuitive, measuring the observations in a suitable band around the simulated values.

The past results obtained with the ZEC_Index have been improved with this procedure, as well as the concept of ZEC ; calibration by energy was strengthened with this methodology because more models with low energy consumption were among the best models, and there was no uncertainty about the selection of the model in the evaluation process because there was a general agreement between different indices about which one was the best.

A big difference between the synthetic model and real model has been observed, and the new methodology performed better under the real case scenario. This premise has proven to be true, since the methodology presented in this paper is being applied to different buildings and in different calibration periods, showing similar results to those obtained in the previous sections. It is a promising area of research where more calibrated buildings in different environments could be studied. The SABINA project will offer this opportunity.

While in this study, we have used the values of the index to rank the best model offered by the calibration process, in future research, specific values of these indices, in a similar way to that provided by ASHRAE Guideline 14 [16], could be obtained in order to give an idea of the quality of the model. The next generation of BEMs should be classified as complying with the level of quality fed by these indices, depending on the types of applications that are required.

Author Contributions: V.G.G. and L.Á.C. supervised the methodology used in the article, performed the simulations and the analysis, and wrote the manuscript. G.R.R. developed the EnergyPlus model and participated in the data analysis. J.F.L.F. and L.Á.C. developed the methodology that was proposed in the article. All of the authors revised and verified all the manuscript before sending it to the journal.

Funding: The work is funded by the research and innovation program Horizon 2020 of the European Union under Grant No. 731211, project SABINA.

Acknowledgments: We wish to acknowledge the assistance of the University of Navarra (Spain) and, in particular, its maintenance staff for providing us with both the building documentation and data from the sensors placed in it.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Arg	Argument
AE	Absolute Error
AIC	Akaike Information Criterion
BES	Building Energy Simulation
BE	Bias Error
BIC	Bayesian Information Criterion
bR^2	Multiplied by the coefficient of the regression line (b)
CV(RMSE)	Coefficient of Variation of RMSE
d	Index of agreement
GoF	Goodness-of-fit index
M&V	Measurement and Verification process
MAE	Mean Absolute Error
MAPE	Mean Absolute Percent Error
MBE	Mean Bias Error
md	Modified index of agreement
$mNSE$	Modified Nash–Sutcliffe efficiency
MSE	Mean Squared Error
NMBE	Normalized Mean Bias Error
NSE	Nash–Sutcliffe Efficiency
PBIAS	Percent Bias
R^2	Coefficient of determination
rd	Relative index of agreement
RMSE	Root Mean Squared Error
$rNSE$	Relative Nash–Sutcliffe Efficiency
rSD	Ratio of Standard Deviations
UI	Uncertainty Index
ZEC	Zero Energy for Calibration

References

1. Arcipowska, A.; Anagnostopoulos, F.; Mariottini, F.; Kunkel, S. *Energy Performance Certificates across the EU; A Mapping of National Approaches*; The buildings Performance Institute Europe (BPIE): Bruxelles, Belgium, 2014.
2. Lewry, A.J.; Ortiz, J.; Nabil, A.; Schofield, N.; Vaid, R.; Hussain, S.; Davidson, P. *Bridging the Gap Between Operational and Asset Ratings—The UK Experience and the Green Deal Tool*; BRE Group: Watford, UK, 2013.
3. IPMVP Committee. *International Performance Measurement and Verification Protocol: Concepts and Options for Determining Energy and Water Savings Volume I*; Technical Report; Efficiency Valuation Organization: Washington, DC, USA, 2012.
4. Fernández Bandera, C.; Ramos Ruiz, G. Towards a New Generation of Building Envelope Calibration. *Energies* **2017**, *10*, 2102.
5. Florita, A.; Henze, G. Whole Building Fault Diagnostics—A Bayesian Approach; In Proceedings of the Intelligent Building Operations Workshop, Boulder, CO, USA, 20–22 June 2013.
6. Haberl, J.; Bou-Saada, T. Procedures for calibrating hourly simulation models to measured building energy and environmental data. *J. Sol. Energy Eng.* **1998**, *120*, 193–204.

7. Dhar, A.; Reddy, T.; Claridge, D. A Fourier series model to predict hourly heating and cooling energy use in commercial buildings with outdoor temperature as the only weather variable. *J. Sol. Energy Eng.* **1999**, *121*, 47–53.
8. Haberl, J.S.; Sreshthaputra, A.; Claridge, D.E.; Kisoock, J.K. Inverse model toolkit: Application and testing. *ASHRAE Trans.* **2003**, *109*, 435.
9. Kisoock, J.K.; Reddy, T.A.; Claridge, D.E. Ambient-temperature regression analysis for estimating retrofit savings in commercial buildings. *J. Sol. Energy Eng.* **1998**, *120*, 168–176.
10. Crawley, D.B.; Lawrie, L.K.; Winkelmann, F.C.; Buhl, W.F.; Huang, Y.J.; Pedersen, C.O.; Strand, R.K.; Liesen, R.J.; Fisher, D.E.; Witte, M.J.; et al. EnergyPlus: Creating a new-generation building energy simulation program. *Energy Build.* **2001**, *33*, 319–331.
11. Sonderegger, R. *Diagnostic Tests Determining the Thermal Response of a House*; Lawrence Berkeley National Laboratory: Berkeley, CA, USA, 1977.
12. Vuelvas, J.; Ruiz, F.; Gruosso, G. Limiting gaming opportunities on incentive-based demand response programs. *Appl. Energy* **2018**, *225*, 668–681.
13. Communication from the Commission. *A Policy Framework for Climate and Energy in the Period From 2020 to 2030*; European Commission: Brussels, Belgium; 2014.
14. Sabina-Project. SABINA. SmArt BI-Directional Multi eEnergy gAteway. 2016. Available online: <https://sabina-project.eu/> (accessed on 31 May 2019).
15. Webster, L.; Bradford, J.; Sartor, D.; Shonder, J.; Atkin, E.; Dunnivant, S.; Frank, D.; Franconi, E.; Jump, D.; Schiller, S.; et al. *M&V Guidelines: Measurement and Verification for Performance-Based Contracts*; Version 4.0; Technical Report; U.S. Department of Energy, Federal Energy Management Program: Washington, DC, USA, 2015.
16. ASHRAE. *ASHRAE Guideline 14–2014, Measurement of Energy, Demand, and Water Savings*; ASHRAE: New York, NY, USA, 2014.
17. Giuliani, M.; Henze, G.P.; Florita, A.R. Modelling and calibration of a high-mass historic building for reducing the prebound effect in energy assessment. *Energy Build.* **2016**, *116*, 434–448.
18. Tahmasebi, F.; Zach, R.; Schuß, M.; Mahdavi, A. Simulation model calibration: An optimization-based approach. *BauSIM* **2012**, *1*, 386–391.
19. Hong, T.; Lee, S.H. Integrating physics-based models with sensor data: An inverse modeling approach. *Build. Environ.* **2019**, *154*, 23–31.
20. Cipriano, J.; Mor, G.; Chemisana, D.; Pérez, D.; Gamboa, G.; Cipriano, X. Evaluation of a multi-stage guided search approach for the calibration of building energy simulation models. *Energy Build.* **2015**, *87*, 370–385.
21. Bandera, C.F. La Inteligencia Artificial como Inspiración Para la Generación y Diseño de Modelos térmicos de Edificios. Ph.D. Thesis, Universidad de Navarra, Pamplona, Spain, 2016.
22. Bandera, C.F.; Mardones, A.F.M.; Du, H.; Trueba, J.E.; Ruiz, G.R. Exergy As a Measure of Sustainable Retrofitting of Buildings. *Energies* **2018**, *11*, 1–19.
23. Manfren, M.; Aste, N.; Moshksar, R. Calibration and uncertainty analysis for computer models—A meta-model based approach for integrated building energy simulation. *Appl. Energy* **2013**, *103*, 627–641.
24. Robertson, J.J.; Polly, B.J.; Collis, J.M. Reduced-order modeling and simulated annealing optimization for efficient residential building utility bill calibration. *Appl. Energy* **2015**, *148*, 169–177.
25. Nguyen, A.T.; Reiter, S.; Rigo, P. A review on simulation-based optimization methods applied to building performance analysis. *Appl. Energy* **2014**, *113*, 1043–1058.
26. Ruiz, G.R.; Bandera, C.F. Analysis of uncertainty indices used for building envelope calibration. *Appl. Energy* **2017**, *185*, 82–94.
27. Farhang, T.; Ardeshtir, M. Monitoring-based optimization-assisted calibration of the thermal performance model of an office building. In Proceedings of the International Conference on Architecture and Urban Design, Tirana, Albania, 19–21 April 2012.
28. DoE, U. *Energyplus Engineering Reference*; The Reference to Energyplus Calculations; EnergyPlus: Washington, DC, USA, 2010.
29. Wetter, M.; Wright, J. A comparison of deterministic and probabilistic optimization algorithms for nonsmooth simulation-based optimization. *Build. Environ.* **2004**, *39*, 989–999.
30. Chaudhary, G.; New, J.; Sanyal, J.; Im, P.; O'Neill, Z.; Garg, V. Evaluation of Autotune calibration against manual calibration of building energy models. *Appl. Energy* **2016**, *182*, 115–134.

31. Ruiz, G.R.; Bandera, C.F.; Temes, T.G.A.; Gutierrez, A.S.O. Genetic algorithm for building envelope calibration. *Appl. Energy* **2016**, *168*, 691–705.
32. Mustafaraj, G.; Marini, D.; Costa, A.; Keane, M. Model calibration for building energy efficiency simulation. *Appl. Energy* **2014**, *130*, 72–85.
33. Zambrano-Bigiarini, M. *hydroGOF: Goodness-of-Fit Functions for Comparison of Simulated And Observed Hydrological Time Series*; R Package Version 0.3-8. 2014.
34. Abbaspour, K.C.; Rouholahnejad, E.; Vaghefi, S.; Srinivasan, R.; Yang, H.; Kløve, B. A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model. *J. Hydrol.* **2015**, *524*, 733–752.
35. ASHRAE. *Fundamentals Handbook*, IP ed.; ASHRAE: New York, NY, USA, 2017.
36. EVO. *Uncertainty Assessment for IPMVP—Release of a New IPMVP Application Guide*; EVO: Washington, DC, USA, 2018.
37. Yoon, J.; Lee, E.J.; Claridge, D. Calibration procedure for energy performance simulation of a commercial building. *J. Sol. Energy Eng.* **2003**, *125*, 251–257.
38. Touzani, S.; Granderson, J.; Fernandes, S. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy Build.* **2018**, *158*, 1533–1543.
39. Glasgo, B.; Hendrickson, C.; Azevedo, I.L. Assessing the value of information in residential building simulation: Comparing simulated and actual building loads at the circuit level. *Appl. Energy* **2017**, *203*, 348–363.
40. Chakraborty, D.; Elzarka, H. Performance testing of energy models: Are we using the right statistical metrics? *J. Build. Perform. Simul.* **2018**, *11*, 433–448.
41. Coakley, D.; Aird, G.; Earle, S.; Klebow, B.; Conaghan, C. Development of Calibrated Operational Models of Existing Buildings for Real-Time Decision Support and Performance Optimisation. In Proceedings of the CIBSE Technical Symposium, Edinburgh, UK, 14–15 April 2016.
42. Reddy, T.A.; Claridge, D.E. Uncertainty of measured energy savings from statistical baseline models. *HVAC R Res.* **2000**, *6*, 3–20.
43. Beven, K.J. *Rainfall-Runoff Modelling: The Primer*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
44. Abbaspour, K. *User Manual for SWAT-CUP, SWAT Calibration and Uncertainty Analysis Programs*; Swiss Federal Institute of Aquatic Science and Technology, Eawag: Duebendorf, Switzerland, 2007.
45. Kamali, B.; Abbaspour, K.C.; Yang, H. Assessing the uncertainty of multiple input datasets in the prediction of water resource components. *Water* **2017**, *9*, 709.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).