Building and Environment 144 (2018) 94–107

Contents lists available at ScienceDirect

Building and Environment

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

Selecting statistical indices for calibrating building energy models

Marcus Vogt*, Peter Remmen, Moritz Lauster, Marcus Fuchs, Dirk Müller

RWTH Aachen University, E.ON Energy Research Center, Institute for Energy Efficient Buildings and Indoor Climate, Mathieustraße 10, 52074, Aachen, Germany

ARTICLE INFO

Keywords:
Dynamic building energy simulation
Statistical indices
Automated model calibration
Measured energy data
Non-residential sector

ABSTRACT

A well-known problem in the dynamic simulation of buildings energy consumption are the discrepancies between the simulated and measured data, which call for calibration techniques to obtain more accurate and reliable building models. The most recognized calibration techniques use statistical indices to assess and improve the quality of simulation models. While there are already well known statistical indices available to evaluate the simulation outputs, the combination of indices offers potential for further improvements in this field. To assess the procedure of calibrating building simulation models, we present a ranking of six tested statistical indices and their combinations (63 statistical metrics), produced by an automated evaluation procedure, in the specific case of calibrating to annual heat demand curves. The developed evaluation procedure is also able to account for eventual deterioration of other statistical metrics, which are not tuned during the calibration. We apply the new method in dynamic, hourly simulations to a use case with 200 buildings, for which extensive measurement data are available. Based on the generated ranking, we recommend using combinations of four statistical indices: the Coefficient of Variation of Root Mean Square Error ($CV(RMSE)$), the Normalized Mean Error (NME), the standardized contingency coefficient (C_{χ^2}) and the coefficient of determination (R^2). In our use case, these combinations lead to better results than the commonly used indices $CV(RMSE)$ and Normalized Mean Bias Error (). In addition, we could show that it is beneficial to use another index for evaluation than for calibration, because it detects eventual deterioration of the simulation output results.

1. Introduction

Since buildings account for one third of the total global energy consumption [1], lowering the energy demand of buildings is an important element in the overall objective of energy savings, which also includes the retrofitting of existing buildings [2]. In this context, building performance simulation (BPS) can be used as an attempt to simulate the thermal building performance as good as possible, allowing to conduct studies leading to reduce the energy demand of buildings. BPS has been increasingly used in post-construction stages [3]. Despite continuous progress in the field, the discrepancies between simulated results and measured data in actual buildings remain an issue in BPS [4].

Therefore, improving the match between simulated and measured performance has become of high importance for the broader practical use of BPS [3]. The process of minimizing these discrepancies is known as calibration. One important aspect in calibrating building energy models is the evaluation of the simulated output results against the measured data, as part of the process to propose energy efficiency retrofit measures or predict future energy savings [3,5]. For this purpose, there are two main techniques applied: graphical and statistical approaches [6]. The graphical technique is used for manual calibration, generating a graphical representation of both data sets, which has to be interpreted by the user. As this technique depends a lot on user experience, it should always be accompanied by the statistical technique [6]. The statistical technique uses statistical indices to quantify the discrepancies between both data sets. Although the statistical technique also has its limitations (e.g. cancellation effect for some indices), it is the most recognized way to check if a model can be considered as sufficiently calibrated, i.e. the discrepancies are within an acceptable tolerance range [7]. The statistical technique can be of use for manual or automated calibration approaches [8].

In the following, we want to highlight the main differences between manual and automated calibration approaches. Manual calibration approaches require hand-operated skill and they are time consuming. Furthermore, they often depend upon modeler skills and cannot be easily scaled up for more complex cases [9]. Automated approaches could be more broadly used, since they depend less on hand-operated skills and user expertise. Still, in the extensive literature review by Coakley et al. [8], 74% of the presented calibration approaches were

* Corresponding author.
E-mail addresses: marcus.vogt@rwth-aachen.de (M. Vogt), PRemmen@eonerc.rwth-aachen.de (P. Remmen), MLauster@eonerc.rwth-aachen.de (M. Lauster), MFuchs@eonerc.rwth-aachen.de (M. Fuchs), DMueller@eonerc.rwth-aachen.de (D. Müller).

https://doi.org/10.1016/j.buildenv.2018.07.052
Received 8 March 2018; Received in revised form 28 July 2018; Accepted 30 July 2018
Available online 04 August 2018
0360-1323/

manual, while only 26% were automated ones. Furthermore, there is currently no single methodology generally adopted for calibration of building energy models [8,10].

In the context of automatic calibration methods, Bayesian calibration has been increasingly applied to building energy models [11]. Bayesian calibration is a probabilistic calibration approach and has the advantage that the model predictions can consider all sources of uncertainties through the use of prior input distributions, even of over fitted parameters [12]. Furthermore, the Bayesian calibration method attempts to correct any inadequacy of the model, which is revealed by a discrepancy between the observed data and the model predictions from even the best-fitting parameter values [8].

In this article, we focus on one particular aspect of the calibration process. We evaluate the error between measured and simulated data using different statistical indices or their combinations to objective functions. In this study, we do not cover the accuracy assessment of the final models based on "input-side error metrics" regarding the building descriptors matching the real building. By considering a pool of statistical indices, this study shall contribute to a better practical understanding of the use of these indices for the automated building model calibration to heat demand curves at an hourly time step.

To evaluate the calibration results, most commonly two statistical indices are used: the Normalized Mean Bias Error ($NMBE$) and the Cumulative Variation of Root Mean Squared Error ($CV(RMSE)$). Three guidelines specify the acceptable calibrations for these indices: American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Guideline 14 [6], International Performance Measurement and Verification Protocol (IPMVP) [13] and Measurement and Verification of Federal Energy Projects (FEMP) [14]. The acceptable tolerance ranges depend on the time step. In this article we consider an hourly time step. The index $NMBE$ is useful to evaluate the overall positive or negative bias of a model, while $CV(RMSE)$ measures the variance of the model. Therefore, the indices $NMBE$ and $CV(RMSE)$ express uncertainty in different ways and do not always correspond [15]. Moreover, in the calculation of $NMBE$ a cancellation effect can occur [8]. This effect can be significant, especially at hourly time steps.

Another evaluation index is the coefficient of determination R^2, which gives the proportion of the variance in measured data that is predictable by the model [16]. For R^2 acceptable tolerance ranges are defined by the ASHRAE Handbook [17] and IPMVP [13]. However, also R^2 has its weaknesses: This statistic is sensitive to extreme values (outliers) and it is not sensitive to additive and proportional differences between simulated and measured data [16].

Furthermore, several researchers use statistical hypothesis tests, e.g. chi-square [18,19], to assess the calibration results. There is also the possibility of combining the statistical indices, as seen for the cost function f_i, a combination of $CV(RMSE)$ and R^2 [20], and the combination between $CV(RMSE)$ and $NMBE$ as proposed by Yoon et al. [21]. In the ASHRAE Research Project 1051-RP [22] $CV(RMSE)$ and $NMBE$ are combined using the Goodness-Of-Fit index (GOF).

Ramos Ruiz and Fernandez Bandera (2017) [23] tested the indices GOF, $CV(RMSE)$ and f_i, as well as their combinations, in a genetic algorithm for calibrating building models and conclude that all of them are suitable for calibration, but the authors recommend to use GOF.

Garrett and New (2016) [24] question the suitability of the metrics $CV(RMSE)$ and $NMBE$ canonized by ASHRAE Guideline 14 to reduce the input-side error of calibrated models. In their work, they investigate the correlations between input and output errors measured by four alternative metrics, especially at an hourly time step and the possibility that the input error decreases, while the output error is tuned with a better suitable statistical index. If the correlations between output-side and input-side errors were high enough using these alternative metrics, then these metrics should be proposed in the ASHRAE Guideline 14. However, the authors find that these correlations were not statistically significant, implying that $CV(RMSE)$ and $NMBE$ are equally suitable as the other four tested alternative metrics. Only for the hourly interior equipment and lighting electricity, correlations were high and future work should investigate whether tuning based on these metrics produces lower input error.

Besides the similar questioning on the use of the industry standard metrics ($CV(RMSE)$ and $NMBE$) and the search for other suitable metrics, the work of Garrett and New (2016) [24] differs from this present work in several aspects. The most evident differences lay on the used method and the focus on the source of the error. Garrett and New (2016) [24] analyze the correlations among input and output error measures to test the suitability of an statistical index in a more general manner.

While there are well known statistical indices available for the simulation output evaluation, the effect of combining indices offers potential for further studies benefiting the field. Our goal is to narrow down suitable statistical indices, or combinations of these, to evaluate the simulation output results through a comprehensive, automated methodology. Moreover, most calibration procedures tune only one output error metric and cannot account for eventual deterioration of other error metrics. In this context, the following research question arises: Which combinations of indices should rather be used for calibrating whole-building heat demand curves at an hourly time step, if eventual deterioration of other output error metrics is taken into account?

The aim of this paper is to address this knowledge gap and to test six statistical indices and their combinations (in total 63 statistical metrics) to evaluate the suitability for calibrating building models based on a ranking, which is provided by our developed automated evaluation procedure. The main contribution of this paper is to answer the above stated research question. To answer this question, we developed an automated calibration and evaluation procedure.

The automated evaluation procedure enables a ranking of all considered six statistical indices and their combinations for the building energy model calibration. The peculiarity of the automated evaluation method is that every calibrated model is evaluated by all considered indices (including combinations of indices). This allows a more objective evaluation of the calibration results, as during the calibration process only one output error metric is tuned and eventual deterioration of the other indices is taken into account. Based on the results of the automated evaluation procedure, we recommend suitable statistical indices and combinations for the building energy calibration. Consequently, the use of the proposed combinations of statistical indices can lead to better calibration results for the calibration of building model heat demand curves at an hourly time step.

As a use case, we consider more than 200 buildings of a research campus located in Germany of which extensive measurements of whole-building heat demand are available. Due to the local climate, most buildings need to be supplied by heat throughout the entire year. As a consequence, the heat demand of the buildings is extensively measured, which mostly provides reliable measured data. In our use case, we focus solely on the heat demand, due to its importance regarding the local energy demand and also to be able to consider as many buildings as possible for the calibration to annual, whole-building heat demand curves at an hourly time step. For an efficient work flow, we need an automated calibration method to process our analysis. We analyze and calibrate the individual building models to better fit their respective measured annual whole-building heat demand curves.

We want to stress, that the further use of the word "calibration" in this article expresses the tuning of model input parameters using statistical indices. Our developed calibration method is a simplified calibration method and works only with binary input parameters (e.g. construction type of the building: heavy/light or the presence of an air-handling unit (AHU): True/False). These binary input parameters are then processed by a data enrichment procedure to adjust the numerical parameters of the simulation model. Because this is a special case of a calibration method, it cannot replace established automated calibration methods. Therefore, the quantified results produced by this simplified