BUILDING SIMULATIONS FOR CONTROL TUNING: ACCOUNTING FOR MODEL USEFULNESS IN CALIBRATION METRICS CHOICE

A. Bres[1], F. Amblard[2], and S. Hauer[1]
[1]AIT Austrian Institute of Technology GmbH, Center for Energy, Vienna, Austria
[2]HES-SO Valais-Wallis, Sion, Switzerland

ABSTRACT

Calibration, defined as the adjustment of uncertain parameters to achieve a better agreement between simulation and measurements, is a key task in modeling existing systems. Different calibration metrics may be used to quantify this agreement. In the field of building performance simulation, a few metrics are established in practice but there is still little evidence on how to assess and select calibration metrics. This paper investigates the assumption that the values of an adequate calibration metric for given simulations models should correlate with the usefulness of these models. Investigations are carried out with synthetic data for an example heat-pump system serving a residential building, looking at model usefulness for the specific task of tuning control parameters.

INTRODUCTION

Building simulation calibration

Simulation models of buildings and heating, ventilation and air-conditioning (HVAC) systems are usually subject to a range of uncertainties. Calibration refers to the adjustment of uncertain parameters to achieve a better agreement between simulation results and measurements (Reddy et al., 2007a). Building simulation calibration as discussed in this paper is a special case of system identification (Ljung, 1999) with grey-box models of nonlinear systems. Wherever measurements are available, model calibration and validation are essential modelling steps contributing to ensuring model quality.

Calibration metrics for building simulation

In order to quantify how well the simulations and measurements agree, various calibration metrics are used. These metrics are usually expressed as a function m taking as input two vectors of the same length (n) corresponding to simulated ($\hat{y}$) and measured (y) values at some time steps (i) and returning a non-negative real number. Note that these calibration metrics are often not metrics in the strict mathematical sense, as they may fail to be symmetric and to satisfy the triangle inequality (Encyclopedia of Mathematics, 2016). A typical metric is the mean squared error defined in Equation (1).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (1)$$

The coefficient of variation of the root mean squared error defined in Equation (2) provides a normalized value which can be expressed in percent.

$$CVRMSE = \frac{\sqrt{MSE}}{\frac{1}{n}\sum_{i=1}^{n} y_i} \quad (2)$$

The normalized mean bias error defined in Equation (3) is also expressed in percent, and is known to be only weakly informative because positive and negative biases at different time steps may cancel each other out (Royapoor & Roskilly, 2015).

$$NMBE = \frac{\sum_{i=1}^{n}\hat{y}_i - \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i} \quad (3)$$

ASHRAE Guideline 14 (ASHRAE, 2002) provides criteria for the successful calibration of simulation models, for instance requiring that $CVRMSE$ should be lower than 30% and $NMBE$ within $\pm$10% when using hourly values of energy consumption. Despite the lack of substantiated justification, these criteria enjoy widespread use. However, limitations of these criteria and the underlying metrics have also been exposed in the literature:

- Quantitative criteria as defined in ASHRAE Guideline 14 lack time resolutions other than monthly or hourly (e.g. sub-hourly, as in the present paper), and variables other than energy use (e.g. temperatures).
- The problem of finding parameter values leading to NMBE and CVRMSE values close to 0 is mostly underdetermined (Reddy et al., 2007a). Low discrepancy values of the output do not exclude high input errors (Garrett & New, 2016).
- Mean squared error and derived metrics may not be adapted to calibrating simulations with discrete behaviour (e.g. on/off cycling) observed at short time steps. Since binary variables can only be changed at predefined time points, which do not necessarily

 DOI: 10.3217/978-3-85125-786-1-79

correspond to the true timestamps, there may be systematic shifts that lead to high errors despite a qualitatively good representation of reality.

- Using the same data to calibrate models and evaluate their goodness-of-fit – as is frequently done - can lead to overfitting and biased evaluation (Chong et al., 2017). Although this is a general issue and not related to any metric in particular, it deserves to be kept in mind when using quantitative criteria.

In the system identification literature, where calibration metrics are referred to as identification criteria, general criteria for their selection include consistency, robustness and ease of computation (Ljung, 1999). In the field of building simulation, the literature does not provide any definite answer to the question of how calibration metrics should be assessed and selected. In a rare attempt, Garrett & New (2016) determined the relevance of calibration metrics by correlating calibration metrics in building simulation outputs with errors in input parameters, resulting in rather weak correlations for all the investigated metrics. Similar questions on the suitability of metrics have been raised in other disciplines such as time series forecasting (Hyndman & Koehler, 2006; Kim & Kim, 2016), leading to the claim that widely used metrics such as mean absolute percentage error (MAPE) have disadvantages and alternatives should often be preferred.

In conclusion, results from the literature suggest that the question of choosing adequate metrics for building simulation calibration should not be overlooked.

Calibration and model usefulness

The approach taken in this paper relates the quality of a calibration metric to the usefulness of models for which this metric is minimized. If "all models are wrong but some are useful" (Box, 1979), as it has often been posited, model usefulness should indeed be the primary criterion in assessing simulation models in general, as well as in assessing calibration metrics. The drawback of this approach is that model usefulness in turn is often not sharply defined, and typically depends on the scientific purpose. Hence, we restrict our investigation to specific simulation use cases. We consider the case where simulation is used to compare different values of control parameters for a heating system and subsequently select the most appropriate values. In this case, the usefulness of a given model can be equated to the quality of decisions made based on this model. Also, since the investment costs for selecting different control parameters are the same (as opposed to, for instance, the costs for different refurbishment measures), one may use the simple decision-making assumption, that the control parameters with the lowest simulated costs are chosen.

Using synthetic ground truth data, it is thus possible to investigate the relations between model usefulness and model discrepancy according to various calibration metrics. Synthetic data are instrumental in evaluating calibration techniques, with several reported applications in the literature (Chaudhary et al., 2016; Reddy et al., 2007b).

Research question

In the context of dynamic building and HVAC simulation, is it possible to compare the relevance of different calibration metrics by correlating their respective values with the usefulness of simulation models? We investigate this question by using dynamic simulation to select control parameters for a hydronic heating system.

SIMULATION EXPERIMENT

Simulation model

A heating system composed of a water-water heat pump, a buffer tank and underfloor heating, as illustrated in Figure 1, is used as a test case for the investigations presented in this paper. This heating system serves a residential building.

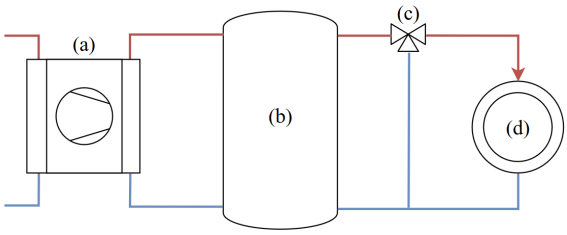

Figure 1: Conceptual system diagram. a) heat pump. b) buffer tank. c) three-way valve. d) consumer (floor heating)

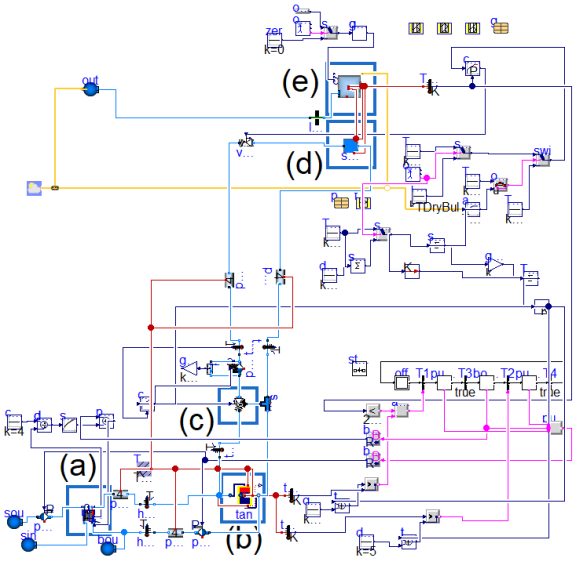

Figure 2: Simulation model diagram, including the same components as in Figure as well as a single-zone building model (e) and control blocks

 DOI: 10.3217/978-3-85125-786-1-79

The system is modeled using Modelica and the Modelica Buildings library (Wetter et al., 2014) Version 6.0.0. The main components and the corresponding models from the Modelica Buildings library are illustrated in Figure 2. For the water-water heat pump, a modified version of Buildings.Fluid.HeatPumps.Carnot_y accounting for temperature dependency of heating capacity is used.

The heating curve of Equation (4) is used to determine the set point T_{sup}^{set} for the supply temperature T_{sup} to the building after a three-way valve, based on a room temperature set point T_r^{set} and a time-averaged outdoor temperature $T_{out,avg}$. The slope S and the level L of the heating curve are examples of parameters to be tuned from the perspective of an installer.

$$T_{sup}^{set} = T_r^{set} + S\,(T_r^{set} - T_{out,avg}) + L \qquad (4)$$

The heat pump cycles ON and OFF with a deadband control on tank temperature, with lower and upper limits T_{tank}^{min} and T_{tank}^{max} aligned with T_{sup}^{set} plus or minus constant offsets to be tuned.

Design of experiment

In terms of parameter variations, a distinction is made between:

- Calibration parameters: uncertain parameters to be calibrated, related to the building, buffer tank and floor heating and summarized in Table 1
- Tuning parameters: control parameters to be tuned, which are related to heat pump controls and summarized in Table 2.

Table 1: Calibration (uncertain) parameters

PARAMETER	RANGE	UNIT
Tank effective volume	1.5 to 2.0	m³
Tank effective height	1.5 to 2.5	m
Tank insulation thickness	0.05 to 0.15	m
Façade solar absorptivity	30 to 70	%
Radiant floor area as fraction of gross area	65 to 85	%

A number $n_{cs} = 20$ of sets of calibration parameter values and a number $n_{ts} = 20$ of sets of tuning parameter values are generated with Latin hypercube sampling (McKay et al., 1979), so as to obtain samples covering the respective parameter spaces effectively. Simulations are then carried out for all pairs $(i_{cs}, i_{ts})_{1\le i_{cs}\le n_{cs}, 1\le i_{ts}\le n_{ts}}$ of calibration parameter sets and tuning parameter sets, resulting in a matrix of n_{cs} by n_{ts} simulation runs.

Table 2: Tuning (control) parameters

PARAMETER	RANGE	UNIT
Heating curve slope S	0.6 to 1.0	-
Heating curve level L	1.0 to 3.0	K
Offset for switching off $T_{tank}^{max} - T_{sup}^{set}$	2.0 to 6.0	K
Offset for switching on $T_{tank}^{min} - T_{sup}^{set}$	-4.0 to 0.0	K
Night setback temperature difference	-4.0 to 0.0	K

Cost functions

A dimensionless cost function is defined in order to quantify the performance of the modeled system in terms of thermal comfort, energy use and equipment cycling. A degree hours criterion is chosen as the cost function for thermal comfort, as suggested in standard CEN/TR 16798-2 (CEN, 2019) and defined in Equation (5), with $\theta_{r,i}$ the simulated room air temperature at time step i, temperature limits $\theta_{r,min} = 21$ °C, $\theta_{r,max} = 25$ °C and a weight w_{tc} in (K.h)$^{-1}$.

$$C_{tc} = w_{tc} \sum_{i=1}^{n} \left(\theta_{r,\min} - \theta_{r,i}\right)^{+} + \left(\theta_{r,i} - \theta_{r,max}\right)^{+} \qquad (5)$$

The costs for energy use are assumed to be proportional to the electrical energy use of the heat pump with a constant unit cost w_{eu} in (W.h)$^{-1}$. The costs for cycling are assumed to be proportional to the number of cycles with a cost w_{cy} per cycle.

The total costs are defined as the sum of these three weighted costs, as in Equation (6).

$$C_{tot} = C_{\mathrm{tc}} + C_{\mathrm{eu}} + C_{\mathrm{cy}} \qquad (6)$$

Note that, while the indices have been omitted in the above equations, these costs are defined for each simulation run: $C_{tot}(i_{cs}, i_{ts})$. Given the relative subjectiveness of the selection of weights, three sets of weights corresponding to different priorities have been defined, as summarized in Table 3:

- A: focus on electricity use;
- B: focus on thermal comfort;
- C: focus on heat pump cycling.

 DOI: 10.3217/978-3-85125-786-1-79

Table 3: Cost function weights

WEIGHT	VALUE WEIGHT SETS			UNIT
	A	B	C	
w_{eu}	0.25	0.25	0.25	$(W.h)^{-1}$
w_{tc}	2.0	10.0	4.0	$(K.h)^{-1}$
w_{cy}	0.1	0.1	1.0	per cycle

Calibration metrics

Calibration metrics are calculated for the heat pump electrical power, which is assumed to be the only measured variable available for calibration.

The calibration metrics summarized in Table 4 are considered. All calibration metrics are expressed in such a manner that they take non-negative values and are equal to zero if the two compared vectors are equal. For instance, one looks at the absolute value of the mean bias error. The mean absolute percentage error often used in other contexts cannot be used because it becomes infinite in the presence of zero values in the reference vector, which is the case here. The frequently used metrics CVRMSE, NMBE and mean absolute error (MAE) are considered. Since costs are defined, the mean absolute error in cost of Equation (7) is also considered.

$$NMAEC = \frac{\left|C_{tot}(i_{cs}, i_{ts}) - C_{tot}(i_{cs,true}, i_{ts})\right|}{C_{tot}(i_{cs,true}, i_{ts})} \quad (7)$$

Table 4: Calibration metrics

METRIC	DESCRIPTION
$CVRMSE$	Coefficient of variation of the root mean squared error, see Equation ()
$\|NMBE\|$	Absolute value of the normalized mean bias error, see Equation ()
MAE	Mean absolute error, see for instance (Willmott & Matsuura, 2005)
$NMAEC$	Normalized mean absolute error in cost, see Equation (7)

In addition, these metrics are calculated at several time intervals where applicable (averages over five minutes (5m), one hour (hr), one day(d)).

Interpretation

The synthetic data generated with the design of experiment described above are then investigated with the following approach: a set $1 \leq i_{cs,\text{true}} \leq n_{cs}$ of calibration parameters is assumed to contain the true values of the system. For each set of calibration parameters i_{cs} and for each pair of tuning parameter sets $(i_{ts,1}, i_{ts,2})$, the cost differences $\Delta C(i_{cs})$ and $\Delta C(i_{cs,true})$ can be determined as in Equation (8).

$$\Delta C(i_{cs}, i_{ts,1}, i_{ts,2}) = C_{tot}(i_{cs}, i_{ts,2}) - C_{tot}(i_{cs}, i_{ts,1}) \quad (8)$$

Additional costs with i_{cs} with reference to $i_{cs,\text{true}}$ are then equal to:

- zero if $\Delta C(i_{cs})$ and $\Delta C(i_{cs,true})$ have the same sign, which means the two alternatives of the pair are ranked correctly and the alternative with lower true cost is chosen;
- the absolute value of $\Delta C(i_{cs,true})$ otherwise, i.e. when the alternative with higher true cost is chosen.

Let $AAC(i_{cs,true}, i_{cs})$ be defined as the average of these additional costs over all pairs $(i_{ts,1}, i_{ts,2})$, divided by the sum of cost differences. This yields a value between 0 (costs ranked correctly for all pairs) and 100 % (costs ranked incorrectly for all pairs).

The correlation of interest to this paper is the one between $AAC(i_{cs,true}, i_{cs})$ and $\overline{D_m}(i_{cs,true}, i_{cs})$. As stated in Equation (9), $\overline{D_m}$ is the mean over all tuning parameter sets of the discrepancy measured by metric m, where $D_m(y, \hat{y})$ is the value of the calibration metric for a ground truth vector y and a simulated vector $\hat{y}$, and $r(i_{cs}, i_{ts})$ is the vector of simulation results for calibration set i_{cs} and tuning set i_{ts}.

$$\overline{D_m}(i_{cs,true}, i_{cs}) = \sum_{i_{ts}=1}^{n_{ts}} D_m(r(i_{cs,true}, i_{ts}), r(i_{cs}, i_{ts})) \quad (9)$$

Moreover, a distinction is made between:

- the period for which the calibration metrics and $\overline{D_m}$ are calculated, referred to as "calibration period", even though calibration is not actually carried out here, but rather only represented by the choice of a calibration set i_{cs} already sampled with Latin hypercube sampling. This calibration period is taken to be either one of the three simulation weeks (W1, W2, W3) or the whole simulation period (total);
- the evaluation period, for which the costs (and AAC) are calculated. The evaluation period is taken to be the whole simulation period.

RESULTS

Results

Example results for two simulation runs are shown in Figrue 3, and illustrate the difficulty of comparing data at sub-hourly time intervals. One may notice the broad early morning peak corresponding to the end of night setback, for which the two simulations are synchronized. During the day, some peaks can be seen to be similar but shifted in the two simulations.

 DOI: 10.3217/978-3-85125-786-1-79

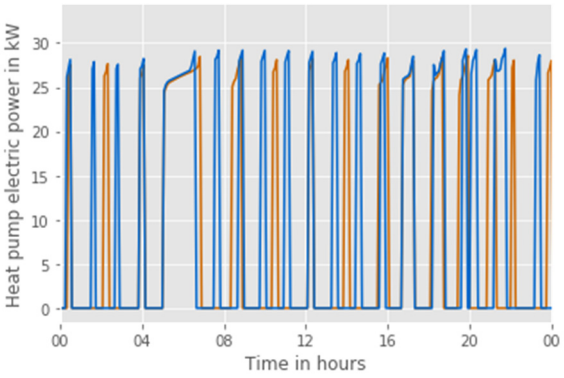

Figure 3: Heat pump electric power for one day at five-minute intervals, for two simulations

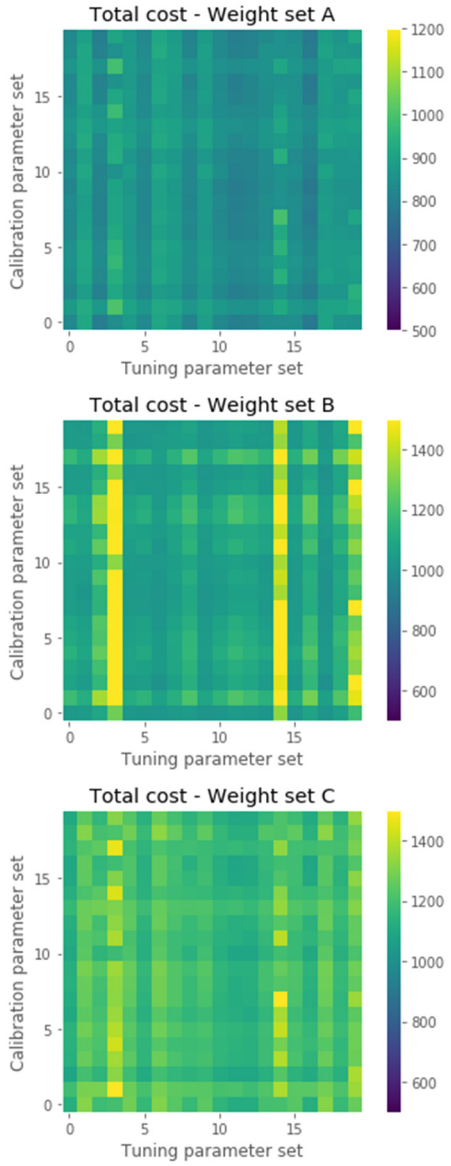

Figure 4: Total costs of the simulation runs

Figure 4 shows the total costs C_{tot} for the matrix of $n_{cs} \times n_{ts}$ simulation runs, for the three different weight sets. The costs of various sets of tuning parameters for a given set of calibration parameters correspond to a row on the diagram, while the costs for a given tuning set correspond to a column. Linear patterns can be distinguished both in terms of tuning parameters (vertical lines) and calibration parameters (horizontal lines). These patterns can be seen to vary with the different weight sets. For weight set B, which gives priority to thermal comfort, two sets of tuning parameters generally leading to worse comfort results are recognizable as yellow vertical lines.

Figure 5 and Figure 6 show scatter plots of the mean calibration error $\overline{D_m}$ against additional costs AAC for several metrics, providing a visual representation of the correlations of interest in this study. In absolute terms, the additional costs always remain below 20 %, which means that – within the bounds of the calibration uncertainty considered here – even uncalibrated models mostly lead to the right decision when comparing two sets of tuning parameters. The (0,0) point, corresponding to the ground truth model parameters, is noticeable for all metrics. Points near the line $y = 0$ correspond to models leading to near-optimal choices of tuning parameters, even though their goodness-of-fit may be low. More problematic are points near the line $x = 0$, corresponding to models with high additional costs despite a high goodness-of-fit. This can be observed especially for the normalized mean bias error in Figure 6. This confirms that a low value of the mean bias error does not imply a "good model", as it may result from errors cancelling each other out. A similar observation can be made with CVRMSE at daily intervals. Mean squared errors and mean absolute errors at time intervals of one hour or five minutes are not affected by this issue, but most imperfect models have high mean errors (above half of the maximal value), which also limits the ability to discriminate between models of different qualities.

Correlations

As revealed by visual inspection of Figure 5 and Figure 6, there are patterns of association between AAC and $\overline{D_m}$, but these patterns are neither sharp nor clearly linear. Correspondingly, the Pearson correlation coefficient, which indicates the degree of linear dependence between the two quantities, takes positive but rather low values, up to 0.5, as reported in Figure 7. The correlation coefficients are lowest for the absolute value of the mean bias error and for CVRMSE with daily values. For the former, one might conclude to a total lack of correlation. The Pearson correlation coefficients are highest for the four metrics corresponding to mean absolute error and CVRMSE with five-minute (5m) and hourly (hr) intervals. Of these four metrics, MAE 5m seems to have the highest correlation with additional costs on average, but the difference does not seem significant in comparison to the deviations across calibration periods and across weight sets.

 DOI: 10.3217/978-3-85125-786-1-79

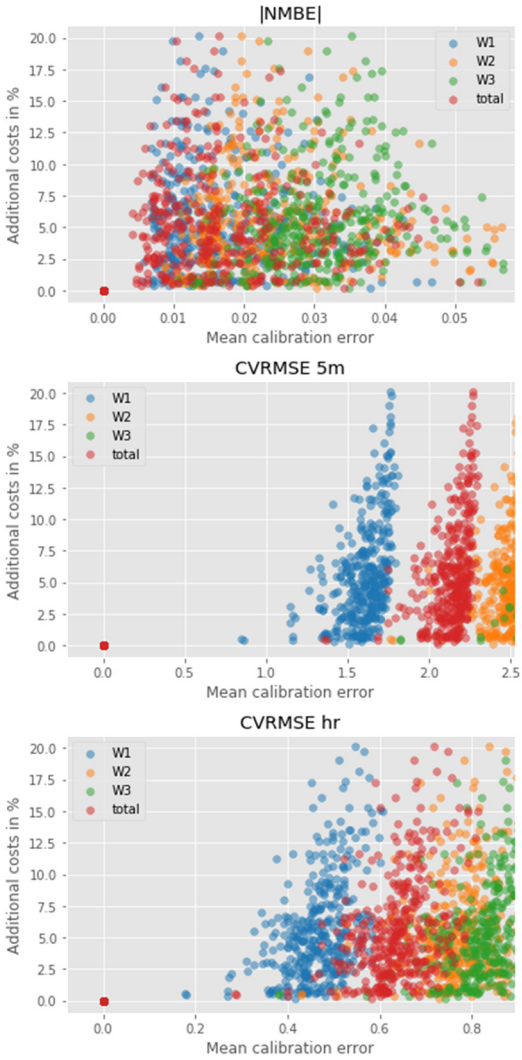

Figure 5: Scatter plot of mean calibration errors |NMBE| and CVRMSE against average additional costs (for cost weighting set A)

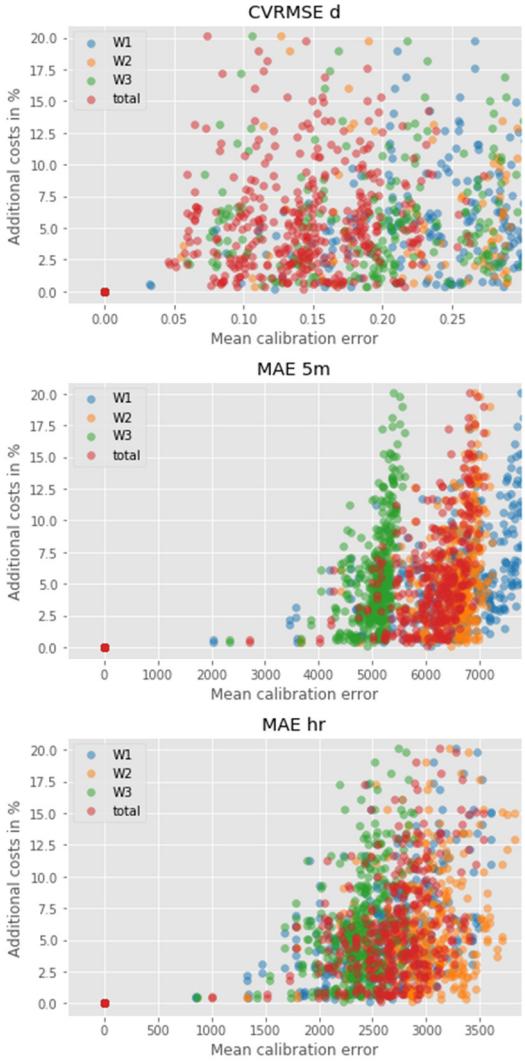

Figure 6: Scatter plot of mean calibration errors CVRMSE and MAE against average additional costs (for cost weighting set A)

Beyond linear correlation, nonlinear dependence is also of interest. A monotonic relationship between values of a calibration metric and the additional costs related to model error, as revealed by high values of the Spearman's rank correlation coefficient (Hauke & Kossowski, 2011), would be a positive property of this calibration metric. Results in terms of Spearman's rank correlation coefficient are shown in Figure 8. They show that, in this respect, metrics calculated on the basis of five-minute values perform best, with the highest rank correlation coefficients for MAE 5m, followed by CVRMSE 5m.

Discussion

ASHRAE Guideline 14 (ASHRAE, 2002) provides criteria for the successful calibration of simulation models which are widely used. However, the relevance of calibration metrics for building simulation has not been subject to much scrutiny until now.

In this paper, calibration metrics are compared by assessing their correlation with model usefulness in the context of using dynamic simulations to select control parameters for a heating system. Accordingly, the results about the usefulness of various metrics are valid in the presented case and not in general. Our results confirm the failure of the mean bias error alone to assess model quality in a useful way. They also show that calibration based on daily values is of limited use when tuning control parameters of the heating system. This is expected, as the most relevant time scale in the studied system is related to the loading and unloading of the buffer and clearly below one day. The use of sub-hourly values may bring additional challenges but more information. Metrics based on mean squared error or absolute error are shown to behave similarly, with a significant dependence on the time interval at which they are calculated. Mean absolute error shows better correlations with additional costs as mean squared error.

 DOI: 10.3217/978-3-85125-786-1-79

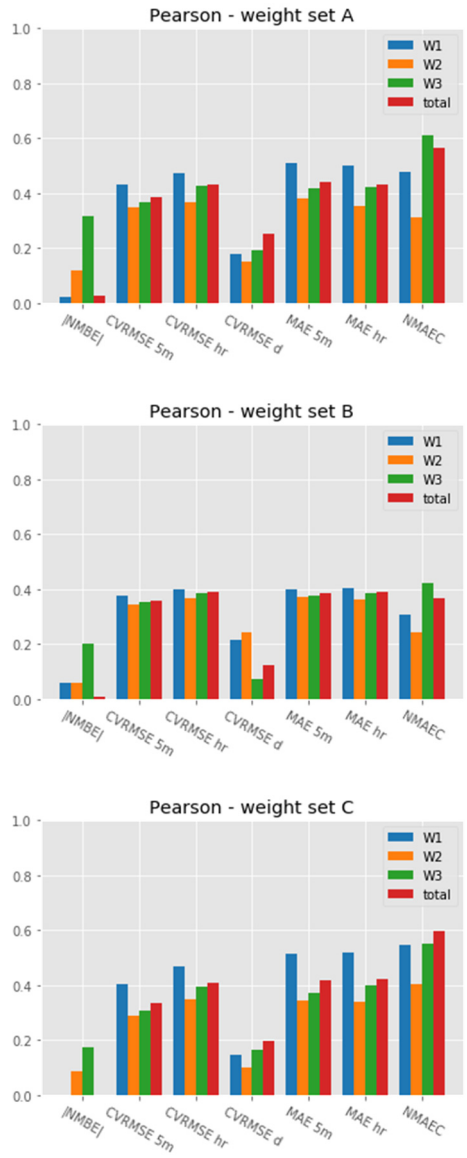

Figure 7: Correlations between AAC and $\overline{D_m}$ in different periods, expressed as Pearson's correlation coefficients r

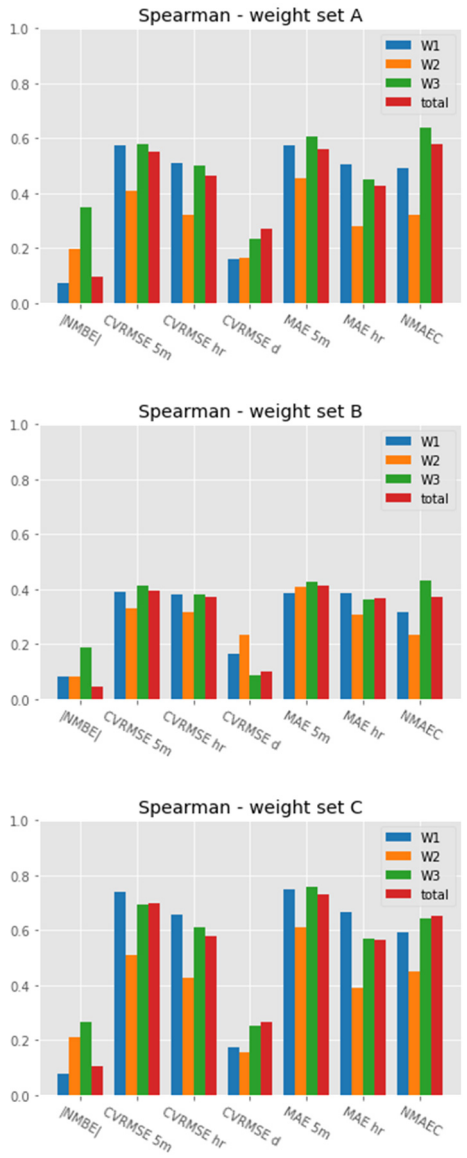

Figure 8: Correlations between AAC and $\overline{D_m}$ in different periods, expressed as Spearman's rank correlation coefficient

The results also illustrate the significant variations affecting calibration metrics calculated in different periods, and the importance of distinguishing calibration and evaluation periods. On another level, the rather low values of AAC reported in this study show that simulation generally yields useful insight even in the presence of parameter uncertainties. A similar approach may be applied to other simulation use cases, such as model predictive control and simulation-based selection of refurbishment options. An application to model predictive control would probably imply models of different structures, whereby complex simulation models may be used as test environments for simpler models with computational requirements adapted to optimization. Applying the approach to planning or refurbishment support might be more challenging, as other costs and functions not represented in simulation would have to be considered.

The use of synthetic data, which makes these investigations possible, has some limitations, such as assuming a perfect model and ignoring measurement errors. We only considered parametric uncertainties, leaving aside the issue of model adequacy. Moreover, we accounted for a limited number of uncertain parameters. These limitations may be addressed in more extensive and elaborate experiments with synthetic data, for instance by considering more sources of uncertainties and simulating measurement noise. Still, experiments with real data may yield additional insight. This would imply not only measurements, but also interactions with a real system (in the present case changes in control parameters) which would have to be sufficiently monitored. Thus, experiments with real data would be significantly more expensive.

 DOI: 10.3217/978-3-85125-786-1-79

CONCLUSION

This paper proposes a new perspective on calibration metrics, by relating them to the usefulness of simulation models in specific cases, here the tuning of control parameters of a heating system. The approach is investigated on the basis of synthetic data obtained from combined variations of control parameters and uncertain calibration parameters. The results show moderate correlations between the values of calibration metrics and the average additional costs of decisions made with an imperfectly parametrized model. The correlations are shown to be particularly weak for mean bias error and errors calculated on daily values. Better correlations are obtained with mean absolute error calculated on hourly and sub-hourly time steps. Further research may investigate calibration metrics in different building performance simulation use case. Other metrics, which can for instance be obtained by filtering time series in various ways, may also be considered.

ACKNOWLEDGEMENT

This research was performed within the SIM4BLOCKS project, funded from the European Union's Horizon 2020 research innovation program under grant agreement No. 695965.

REFERENCES

ASHRAE. (2002). *ASHRAE Guideline 14 – Measurement of Energy and Demand Savings* (ASHRAE Guideline 14). ASHRAE.

Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* (S. 201–236). Elsevier.

Bres, A., Amblard, F., Page, J., Hauer, S., & Shadrina, A. (2019). Now it looks more real—A Study of Metrics and Resolution for the Calibration of Dynamic Simulation. *Proceedings of the 16th IBPSA Conference*, 4609–4616.

CEN. (2019). *CEN/TR 16798-2. Energy performance of buildings—Ventilation for buildings—Part 2: Interpretation of the requirements in EN 16798-1—Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics.*

Chaudhary, G., New, J., Sanyal, J., Im, P., O'Neill, Z., & Garg, V. (2016). Evaluation of "Autotune" calibration against manual calibration of building energy models. *Applied Energy*, *182*, 115–134.

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

Encyclopedia of Mathematics. (2016). Metric. In *Encyclopedia of Mathematics*. Springer. http://www.encyclopediaofmath.org/index.php?title=Metric&oldid=38939

Garrett, A., & New, J. (2016). Suitability of ASHRAE Guideline 14 Metrics for Calibration. *ASHRAE Transactions*, *122*(1).

Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, *30*(2), 87–93.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679–688.

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679.

Ljung, L. (1999). *System Identification—Theory for the User*. PTR Prentice Hall.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239–245.

Reddy, T. A., Maor, I., & Panjapornpon, C. (2007a). Calibrating detailed building energy simulation programs with measured data—Part I: General methodology (RP-1051). *Hvac&R Research*, *13*(2), 221–241.

Reddy, T. A., Maor, I., & Panjapornpon, C. (2007b). Calibrating detailed building energy simulation programs with measured data—Part II: application to three case study office buildings (RP-1051). *Hvac&r Research*, *13*(2), 243–265.

Royapoor, M., & Roskilly, T. (2015). Building model calibration using energy and environmental data. *Energy and Buildings*, *94*, 109–120.

Wetter, M., Zuo, W., Nouidui, T. S., & Pang, X. (2014). Modelica buildings library. *Journal of Building Performance Simulation*, *7*(4), 253–270.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79–82.

 DOI: 10.3217/978-3-85125-786-1-79