Evaluating different metrics for inter-model comparison of urban-scale building energy simulation time series

Johra, Hicham; Mans, Michael; Filonenko, Konstantin; De Jaeger, Ina; Saelens, Dirk; Tvedebrink, Torben

Published in:
Proceedings of Building Simulation 2021: 17th Conference of International Building Performance Simulation Association. Bruges, Belgium, 1-3 September 2021

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Evaluating different metrics for inter-model comparison of urban-scale building energy simulation time series

Hicham Johra[1], Michael Mans[2], Konstantin Filonenko[3], Ina De Jaeger[4], Dirk Saelens[4], Torben Tvedebrink[1]
[1]Aalborg University, Aalborg, Denmark
[2]RWTH Aachen University, Aachen, Germany
[3]University of Southern Denmark, Odense, Denmark
[4]KU Leuven, Leuven, Belgium

Abstract

This article discusses different statistical methods to compare the predictions (in the form of time series) of urban-scale building energy models. It focuses on inter-model comparison without empirical measurements as a reference case. It is thus suggested to build a reference time series as a point-to-point mean average of all the models to be compared (ensemble methods). Different comparison metrics found in the literature are then reviewed and tested on simulation data from a modelling common exercise (DESTEST project). Most of those metrics behave similarly. Finally, a Python-based time series comparison tool is presented. It uses three simple metrics to compare building models: NMBE, CVRMSE of hourly data and CVRMSE of daily amplitudes.

Key Innovations

- Suggestion of a simple solution to build a reference case for inter-model comparison of building energy models without measurement reference case.
- Critical review of different comparison methods and metrics for building energy simulations.
- Presentation of an open-source Python-based tool to compare building energy simulations.

Practical Implications

How to assess the accuracy of building energy models, especially when there is no empirical reference case from a monitored building? This article suggests some simple solutions and introduces an open-source tool using them.

Introduction

To tackle the current environmental and sustainability challenges we are facing, our society needs to drastically decarbonize our cities while improving the quality and reliability of urban energy systems and transportation services. Concerning energy distribution networks, most long-term planning strategies lean towards the establishment of smart grids coupling different energy networks (electricity grid, district heating, district cooling, gas distribution, etc) to enable demand-side management and energy flexibility measures. These smart grids should be able to handle the ever-increasing share of intermittent renewable energy sources.

To that matter, smart grids must account for the capacity, dynamics and interactions of all energy suppliers and end-users. The design, optimization and operation of such a complex system require accurate urban-scale models that can perform dynamic numerical simulations of multiple buildings, infrastructures and energy distribution networks. The multi-physics modelling language Modelica is particularly well-suited for that purpose.

In that context, the *IBPSA Project 1* develops an open-source BIM/GIS and Modelica framework for simulating clusters of buildings connected to energy networks at a city scale (IBPSA, 2017). Within the *IBPSA Project 1*, the District Energy Simulation Test (DESTEST) aims at testing urban-scale energy system simulation tools and, in particular, validate the district energy systems models of dedicated Modelica libraries: *AixLib* (RWTH Aachen University, Germany), *Buildings* (LBNL, USA), *BuildingSystems* (UdK Berlin, Germany), *IDEAS* (KU Leuven, Belgium). All these libraries are based on the core *Modelica IBPSA library* (https://github.com/ibpsa/modelica-ibpsa).

Inspired by the principles of the BESTEST (ANSI/ASHRAE Standard 140-2017), the DESTEST consists of a series of common exercises used for comparison, benchmarking and thorough verifications of urban-scale energy system simulation tools. In each common exercise, different participants are modelling and simulating a given case of buildings and/or energy grid with well-defined properties, characteristics, grid topology, weather conditions, and boundary conditions. The participants can use any suitable commercial and non-commercial simulation tools or the dedicated Modelica libraries. To stay within the scope of the *IBPSA Project 1*, the DESTEST common exercises are restricted to buildings and cluster of buildings connected to district heating or district cooling networks. However, the DESTEST procedure can be extended to other types of systems (Saelens et al., 2019). The DESTEST is also the occasion for participants to discuss common mistakes and pitfalls that are encountered when modelling such systems. The experience and feedback from these common exercises will be gathered into guidelines for good modelling practices. All materials developed for the DESTEST can also be used for training researchers, engineers and students who are willing to gain expertise in dynamic simulations of urban-scale energy systems.

The raw outcome of these common exercises is a set of numerous time series (TS) for some key simulated parameters. The manual qualitative comparison of all of

these TS is very tedious and impractical for a large number of participants. It is thus important to analyse this data with a limited number of appropriate comparison metrics (CMs) to assess the performance of a large number of participants (with their various simulation tools) and establish clear rankings.

To analyse the simulated data generated by the participants of the DESTEST, a Python-based TS comparison tool has been developed. It enables modellers to easily compare their simulation results with the pool of data already generated and uploaded to an online public repository. This DESTEST comparison tool employs some common CMs to assess the differences between the simulated result TS of the user and a reference TS generated from the pool of all vetted results uploaded to the repository. If the differences between the tested TS and the reference TS are large, the user is encouraged to perform a thorough analysis of the model to possibly identify and resolve modelling mistakes.

This article aims at discussing the comparison methods of building energy model (BEM) simulation results. It opens with a suggestion for building a reference for inter-model comparison when there is no empirical reference case. CMs found in the literature are then reviewed and tested. Finally, the DESTEST Python-based TS comparison tool is presented. Although the *IBPSA Project 1* and the DESTEST are also considering district heating and district cooling systems, the scope of this article is restricted to the comparison of BEMs.

Study case

To test the different CMs and illustrate the Python-based TS comparison tool hereafter, results generated by the participants of the DESTEST are chosen among the common exercises concerning single-family dwellings. The study case is a single-family house with simple geometry. It is located in the heating-dominated climate of Belgium. It is assumed to have been constructed in the 1980s and has thus a high space heating demand. The building model has two thermal zones: ground floor (with kitchen and living rooms) and first floor (with only bedrooms). Standard occupancy, internal gains and indoor temperature setpoint schedules are assumed. The infiltration rate is constant. There is no ventilation system. The heating system is an ideal radiator.

The primary question regarding model comparison and validation is what simulated state variables (virtual sensors) should be recorded as output results. However, this crucial discussion is out of the scope of the current article. In the present case, the selected virtual sensors are the ones chosen in the DESTEST common exercises for single-family dwellings: the indoor temperature in each of the two thermal zones and the building heating power use. The simulation period is an entire year. The output data is sampled at a 10-minute interval. More details about this common exercise, the study cases, simplification assumptions and choice of the virtual sensors can be found in Saelens et al. (2019).

At the moment, seven modellers have simulated this case and have submitted their results. These modellers have used commercial simulation tools (IDA ICE, TRNSYS and DIMOSIM) or Modelica with different libraries (*AixLib*, *Buildings*, *BuildingSystems* and *IDEAS*). These seven sets of result data are used for this study.

Reference choice for inter-model comparison of building energy simulations

The quality and validity of a BEM are usually defined by its capacity to accurately predict indoor environment state variables and energy demand. The output of dynamic BEMs usually takes the form of TS for the different variables of interest. Assessing the quality of a BEM thus consists in determining how different the TS of the tested BEM are from a reference TS.

The TS from the tested BEMs are usually denominated “simulated data”, “modelled data”, “predicted data”, “test data”, or “fitted data” (for a model that is a fitted function by regression analysis). The TS from the reference are usually denominated “reference data”, “observed data”, “measured data”, “empirical data” (especially for data measured in an existing building).

Ideally, a BEM validation is performed by comparing the simulation results against empirical measurement data where the measurements, the boundary conditions, the building and its systems are described in detail (ANSI/ASHRAE Standard 140-2017, Annex B23). However, such a benchmarking procedure requires a tremendous amount of time and effort. It would be very expensive and cumbersome to develop one for the validation of urban-scale energy models comprising multiple buildings and/or building occupants.

Contrary to a validation procedure with a measured reference case, the DESTEST is based solely on simulations from various BEMs. Hence, the question of what reference to choose for inter-model comparison is crucial. Some could say that there is no good answer to that question as all models are wrong and should not be taken as a reference. Nevertheless, a reference is necessary so that all models can be compared to each other. For each system’s variable of interest, it is thus suggested to build a reference TS made of the point-to-point mean average or the point-to-point median of the TS of all the BEMs that are to be compared (see Figure 1).

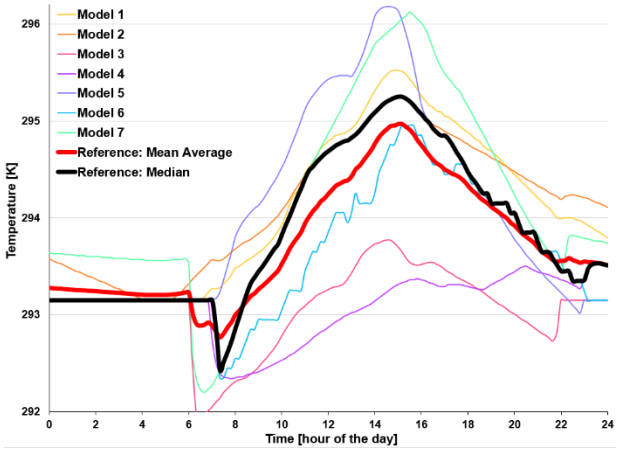

Figure 1: Building reference point-to-point mean average or median time series.

Although not a perfect solution, one can argue that it is the least worst of all and a very simple one. For a large collection of simulation results of the same study case, one could reasonably assume that the different biases of the numerical models and mistakes of the modellers are independent and will compensate each other. Actually, the elimination of individual errors by aggregating the results of many participants has been observed in numerous situations. This effect is commonly designated as the "wisdom of the crowd" (Galton, 1907) or "ensemble methods". The estimation of a single continuous quantity by aggregating the answers (mean average or median) of a large number of independent participants outperforms the majority of individual solutions and falls very close from the true value. The phenomenon has also been observed for higher-dimensional problems, and recent studies indicate that it could be used for a larger spectrum of problem-solving and decision-making situations (Yi et al., 2012). Currently, ensemble methods are frequently used for numerical weather prediction and machine learning. This strengthens the validity of a point-to-point mean average or median reference TS for inter-model comparison purpose. With the increasing number of participants using different numerical models, the reliability of such a reference should only improve.

A new question then stems from the choice of an aggregated reference: should it be a point-to-point mean average or median? The mean average estimates well the central tendency, but it is sensitive to outliers and skewed distributions. On the other hand, the median is less affected by outliers and skewed data, but it is not necessarily affected by new data, it does not estimate well central tendency for small sample size, and it can create an unsmooth TS profile (see Figure 1). As an alternative to those two options, one could use a point-to-point winsorized mean (with moderate winsorizing, e.g., 7.5–22.5% on each side) or trimmed mean (with moderate trimming, e.g., 5–15% on each side) to build the reference TS (Jose and Winkler, 2008).

However, the importance of this question can be moderated by the fact that, for distributions with finite variance (which is the case here), the distance between the median and the mean average is bounded by the standard deviation (Mallows, 1991). Consequently, for a given time step, the distance between the mean average reference TS and the median reference TS is smaller than the standard deviation of the data points used to build these reference TS.

Table 1: Comparison metrics calculated with a mean average reference or median reference (colours are only intended to emphasize extrema).

Model	MBE		RMSE		RMSE 24h-Amp	
	Mean average	Median	Mean average	Median	Mean average	Median
1	0.14	0.12	0.49	0.35	0.36	0.42
2	0.67	0.52	1.08	0.96	0.41	0.43
3	-0.46	-0.49	0.87	0.91	0.68	0.73
4	0.18	0.16	0.68	0.56	0.74	0.77
5	0.14	0.25	0.71	0.75	0.69	0.61
6	-0.78	-0.58	1.16	0.83	0.98	0.91
7	0.10	0.09	0.43	0.39	0.38	0.27

Table 1 presents the calculation results of the Mean Bias Error (MBE), the Root Mean Square Error (RMSE) and the RMSE of the daily amplitude (RMSE 24h-Amp) for the ground floor indoor temperature of the study case. Each CM is calculated against both the mean average and the median reference TS. One can observe in this example that there is not a large difference between the CMs using the mean average reference and the ones using the median reference. Furthermore, the performance ranking order of the different models is fairly well preserved. Similar results are obtained for the two other variables of interest (first floor indoor temperature and heating use).

The current number of modellers participating in the DESTEST common exercises is relatively modest. However, modellers have the occasion to compare their simulation results with that of other participants, and thus notice large deviations due to mistakes that can be eliminated. The chances of large outliers occurring in the pool of results should thus be reduced.

For all those reasons, the point-to-point mean average of all vetted results uploaded to the repository is chosen to build the reference TS. The CMs of the DESTEST comparison tool are thus calculated with this temporary reference. A new modeller joining the common exercise is encouraged to use this tool to compare its simulation results to the current reference. If large deviations are observed, the new modeller is encouraged to analyse and revise its BEM. If the simulation results are deemed to be correct, the new modeller can upload them to the dedicated repository. These new results are integrated into the pool of vetted data and a new temporary reference is generated for the comparison tool.

Review and testing of the comparison metrics for building energy simulations

CMs that are commonly used in the scientific literature to assess TS differences for BEM are tested and compared hereafter. In the equations of these CMs, the TS data points of the tested model are noted m_i, and the TS data points of the reference are noted r_i, with $i \in [1,n]$ and n the number of data points in the TS.

Common qualitative graphical comparisons

Five graphical representations are commonly used to qualitatively assess the differences between BEM simulation outputs: stacked lines plot of the simulated variables as a function of time (see Figure 1), stacked lines plot of the model residuals (r_i - m_i) as a function of time, boxplot (see Figure 2), time distribution plot (also known as "load duration curves") (see Figure 3) and prediction vs reference plot (see Figure 4).

The stacked lines plot is convenient for detailed analysis of TS, but it becomes unreadable for yearly simulations with many daily periodic patterns. The boxplot is a very efficient way to grasp the key statistical characteristics of the data result distribution and to compare many cases against each other. However, it does not preserve any temporal aspects of the TS (e.g., time offsets). Alternatively, the load duration curve provides more temporal information, but it does not show the periodic

patterns and it becomes unreadable with numerous data sets. Finally, the predicted vs reference plot is very convenient to observe global biases, correlations and distortions, but it becomes impractical to display more than one data set on the same figure.

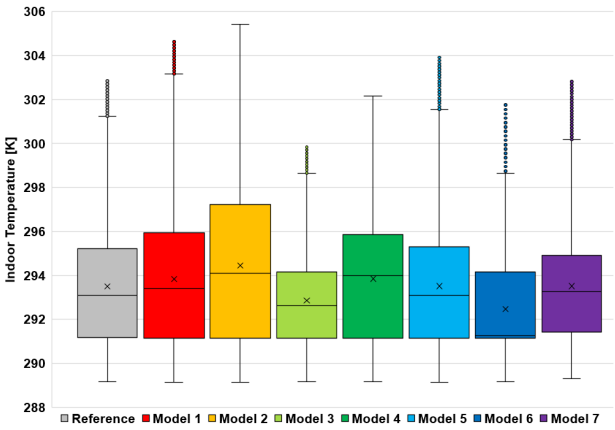

Figure 2: Boxplot of the ground floor indoor temperature during a year (median and quartiles form the box; mean average and outliers are visible; whiskers are set according to Tukey's definition).

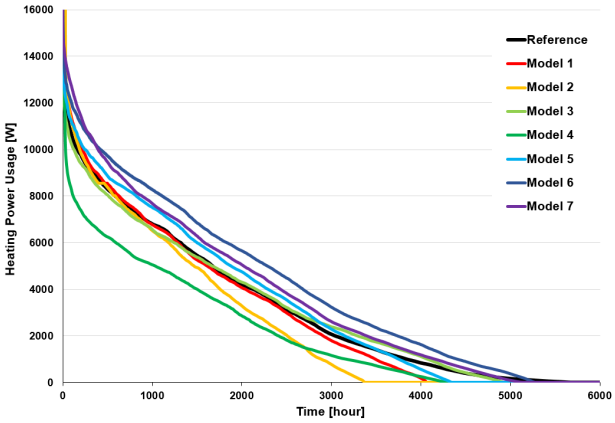

Figure 3: Annual load duration curve for heating power usage (data points sorted in descending order).

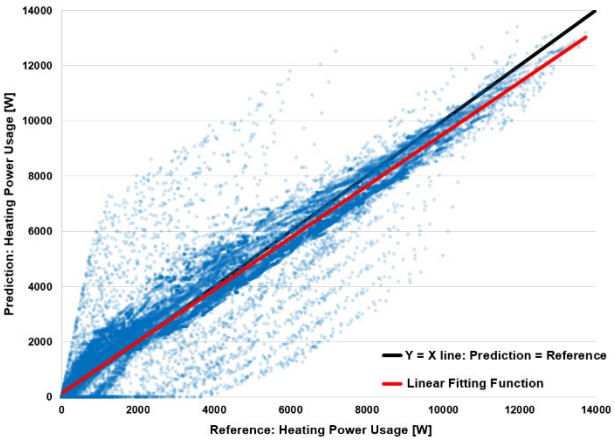

Figure 4: Prediction vs reference plot for heating power usage of model 3.

Common simple comparison metrics

Although intuitive and very informative, a qualitative graphical comparison is not convenient to summarize the differences between numerous TS and rank the quality of multiple models. To that matter, synthetic CMs have been created to condense all the complex differences between two TS into a single number. Of course, a lot of information is lost during this simplification process (in comparison to the qualitative graphical analysis of the human eyes), but it provides a clear and objective assessment method that can easily be automated for a large number of data sets.

The Mean Bias Error (MBE) is probably the simplest CM of all. It calculates the mean average point-to-point difference between the model and the reference:

$$MBE = \frac{\sum_{i=1}^{n}(m_i - r_i)}{n} \quad (1)$$

The MBE is meant to be an indicator of the general bias of the tested model with regards to the reference. Here, a positive MBE indicates that the tested model globally over-predicts the results. Conversely, a negative MBE indicates that the tested model globally under-predicts the results. There are two main issues with this CM. Firstly, the MBE is not normalized, which makes it difficult to give general indications about acceptable MBE ranges for the different variables of a BEM. The MBE cannot be used to compare the performance of a model for simulated variables with different units, scales or natures. Secondly, the MBE can be subjected to cancellation or compensation effects, i.e., local biases in opposite directions compensate each other. For example, local under-estimations would compensation local over-estimations, leading to a globally low MBE despite large local discrepancies.

The most common CMs found in the reviewed studies performing BEM validations are the Normalized Mean Bias Error (NMBE) and the Coefficient of Variation of Root Mean Square Error (CVRMSE). This confirms the observation already made by Coakley et al. (2014) and Ruiz and Bandera (2017).

The NMBE [%] is a normalization of the MBE by the mean average of all the reference data points ($\bar{r}$):

$$NMBE = \frac{\sum_{i=1}^{n}(m_i - r_i)}{n} \times \frac{100}{\bar{r}} \quad [\%] \quad (2)$$

The NMBE of different simulated variables can thus be compared. Similarly to the MBE, the NMBE informs about the global bias of the model: negative values for general under-predicts, and vice versa. However, the NMBE is also prone to compensation effects.

The CVRMSE [%] indicates the variability or randomness between the tested model and the reference:

$$CVRMSE = \sqrt{\frac{\sum_{i=1}^{n}(m_i - r_i)^2}{n}} \times \frac{100}{\bar{r}} \quad [\%] \quad (3)$$

The CVRMSE is a normalization of the Root Mean Square Error (RMSE) by the mean average of all the reference data points ($\bar{r}$). Contrary to the previous metrics, the CVRMSE is not subjected to compensation effects. It is thus well-suited to assess BEM fit and accuracy. However, the CVRMSE does not inform about the direction of global systematic bias. Hence, the calculation of both the CVRMSE and NMBE is recommended.

The popularity of the NMBE and the CVRMSE can be explained by the fact that they are easy to implement into calculation spreadsheet tools like Microsoft Excel, and that they are recommended by well-established international guidelines such as the ASHRAE Guideline 14-2014 (Ruiz and Bandera, 2017).

The coefficient of determination R^2 (R squared) is widely used to measure the goodness of fit for statistical modelling and regression models. It is also very often misunderstood and misused because there are multiple definitions and formulations of R^2, but they are not all necessarily equivalent, and certain formulations present some significant pitfalls (Kvålseth, 1985). R^2 is generally defined as the proportion of variance in the output of a linear model that can be explained by the input variables to that model (the remaining unexplained variability being attributed to unknown variables and/or inherent variability). The common formulation of the coefficient of determination is based on the ratio between the Sum of Squares of the Residuals SS_{res} and the Total Sum of Squares SS_{tot}:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(m_i - r_i)^2}{\sum_{i=1}^{n}(r_i - \bar{r})^2} \quad (4)$$

For BEMs, R^2 is commonly used to assess the accuracy of sub-system time-independent linear regression models over a range of input parameters (e.g., validation of a heat pump model against measurement data during steady-state operation). However, this metric is more seldomly used to assess the accuracy of an entire dynamic BEM by analysing its TS outputs.

The two following metrics are not subjected to compensation effects, but they are not normalized. Similarly to MBE, it is thus not possible to compare their results for simulated variables of different natures. Although very common in statistical modelling, the Mean Squared Error (MSE) and the RMSE are less common for BEM comparison. The RMSE is the standard deviation of the model's residuals or errors (differences between the model's prediction and the reference):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(m_i - r_i)^2}{n}} \quad (5)$$

As an alternative to the CVRMSE, the RMSE can also be normalized by the range (amplitude) of the reference data, or by the interquartile range (IQR) of the reference data:

$$NRMSE = \frac{RMSE}{r_{max} - r_{min}} \quad (6)$$

$$RMSEIQR = \frac{RMSE}{IQR} \quad (7)$$

However, the NRMSE and the RMSEIQR are very rare, in comparison to the CVRMSE, for BEM validations.

The Root Mean Squared Logarithmic Error (RMSLE) is a metric that is commonly found in the Machine Learning community but that is rare for BEM accuracy assessment:

$$RMSLE = \sqrt{\frac{\sum_{i=1}^{n}(\log(m_i + 1) - \log(r_i + 1))^2}{n}} \quad (8)$$

Because of the logarithmic functions in its formulation, the RMSLE is less sensitive to large outliers, in comparison to the RMSE: it penalizes much less very large differences between the tested model and the reference when both the prediction and the reference are large numbers. Besides, the RMSLE penalizes model under-estimations more severely than model over-estimations.

Advanced time series distance metrics

One of the main limitations of the aforementioned simple CMs is that they are calculated point-to-point for synchronized TS with a constant sampling rate. This means that the model prediction at a given time is compared to the reference for the same corresponding time step. This can be very problematic in the case of time offsets between the model output and the reference, especially if there are multiple rapid peaks and drops of the variable of interest. Point-to-point simple metrics would then significantly over-penalize such models, even if the time offset is only a single sampling time step.

To tackle that issue, one can use advanced distance calculation metrics for TS analysis. These distance and shape comparison methods have been developed for classification purposes such as clustering analysis, but they can also be applied to BEM validation. The *TSdist* (Time Series Distance) package is a library of the programming environment "R" that is dedicated to advanced TS analysis, elastic distances calculation and shape comparison. It can be used to measure the dissimilarity between model TS to perform model validation. The package includes four groups of times series distance measurement:

- General shape comparison with lock-step (point-to-point) or elastic distance measurement (measure the distance to the closest point on the other TS, disregarding its time position).
- Feature-based distance measurement: Fourier, wavelet coefficients, autocorrelation values, etc.
- Structure-based distance measurement: a model is fitted to the TS and then compared to that of other TS (measure the quantity of shared model information).
- Prediction-based distance measurement: comparison of the predictions made with the different TS.

Among those, the elastic general shape comparison methods could be interesting for BEM comparison: Dynamic Time Warping (DTW); Dissimilarities based on Pearson's correlation (COR); Dissimilarity index combining temporal correlation and raw value behaviours (CORT); Frechet distance (Mori et al., 2016).

Only a few BEM studies were found using advanced TS analysis accounting for time shift or shape distortion. This is expected since those methods are harder to implement, require some programming skills and have a longer computation time.

DESTEST comparison metrics for building models

In the DESTEST Python-based TS comparison tool, three simple CMs are recommended for BEM validation: 1)

- NMBE.
- CVRMSE on hourly-averaged data (Hourly CVRMSE).
- CVRMSE of daily amplitude (CVRMSE 24h-Amp).

As explained before, the NMBE indicates the overall bias of the model. The Hourly CVRMSE is meant to represent the goodness of fit of the model but without penalizing small-time offsets too much. It is assumed that the hourly averaging of data sampled at a 10-minute rate is a reasonable smoothing of small-time shifts in the BEM predictions. Finally, the CVRMSE 24h-Amp is calculated as the CVRMSE of the daily amplitude (from midnight to midnight) of the simulated variables of interest. This metric was intended to assess how well a model can predict the dynamics of a building: amplitude of indoor temperature variations over 24 hours, daily maximum heating power or cooling power peak, etc.

Testing comparison metrics

In this section, the aforementioned CMs are computed with the TS of the seven BEMs simulating the study case presented before. The reference is constructed as the point-to-point average of the TS of all the models. Table 2 presents a qualitative comparison of all CMs for the simulated heating power usage and the first floor indoor temperature. For clarity, the CMs have been grouped into four categories:

- CMs based on the Mean Bias Error.
- CMs based on the Sum of Squared Errors (SSE).
- Elastic distance metrics between TS.
- CMs based on the daily amplitude.

Table 2: Qualitative comparison of the comparison metrics for building energy model accuracy (colours are only intended to emphasize extrema).

Comparison Metrics	Heating Power Usage						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
ABS(MBE)	105.77	382.00	4.75	135.06	636.61	386.77	675.42
ABS(NMBE) [%]	4.84	17.47	0.22	6.18	29.12	17.69	30.90
MSE	249 025	827 675	558 352	1 474 803	1 207 072	717 745	1 464 974
RMSE	499	910	747	1 214	1 099	847	1 210
RMSLE	2.31	3.05	1.46	2.11	1.22	1.38	2.26
CVRMSE [%]	22.83	41.61	34.18	55.55	50.26	38.75	55.36
NRMSE [%]	3.63	6.63	5.44	8.85	8.00	6.17	8.82
RMSEIQR [%]	12.87	23.47	19.28	31.33	28.34	21.86	31.22
R^2	0.97	0.91	0.94	0.83	0.86	0.92	0.83
Hourly CVRMSE [%]	19.08	35.42	30.32	54.86	44.97	34.13	51.53
DTW	291 424	873 785	293 529	686 479	879 240	522 246	2 203 699
COR	0.17	0.25	0.32	0.57	0.24	0.21	0.29
CORT	13 937	29 589	54 724	102 307	63 101	46 776	53 249
Frechet	4 481	7 388	2 256	4 105	3 318	5 199	6 501
CVRMSE 24h-Amp [%]	74.32	118.73	11.94	23.27	31.15	61.96	87.92

Comparison Metrics	Indoor Temperature of the First Floor						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
ABS(MBE)	0.21	0.70	0.53	0.22	0.20	0.03	0.44
ABS(NMBE) [%]	0.07	0.24	0.18	0.08	0.07	0.01	0.15
MSE	0.19	1.23	0.74	0.52	0.35	0.37	0.96
RMSE	0.44	1.11	0.86	0.72	0.59	0.61	0.98
RMSLE	0.0015	0.0037	0.0029	0.0024	0.0020	0.0021	0.0033
CVRMSE [%]	0.15	0.38	0.29	0.25	0.20	0.21	0.33
NRMSE [%]	3.00	7.57	5.85	4.92	4.03	4.14	6.70
RMSEIQR [%]	11.48	28.92	22.35	18.81	15.39	15.82	25.58
R^2	0.98	0.88	0.93	0.95	0.97	0.96	0.90
Hourly CVRMSE [%]	0.15	0.38	0.28	0.24	0.20	0.20	0.33
DTW	111	1 308	655	311	293	314	657
COR	12.27	35.68	29.71	21.17	20.21	17.02	62.98
CORT	3.76	22.94	18.13	8.57	9.88	6.78	21.86
Frechet	0.77	2.77	1.88	1.57	1.32	1.75	2.95
CVRMSE 24h-Amp [%]	7.30	20.91	15.83	29.98	15.20	44.84	31.58

Although the absolute values of the CMs are very different from one another, the relative performance ranking of the models is clearly preserved for most metrics in their respective group.

Logically, the MBE and the NMBE perform exactly the same since the latter is just a normalization of the former. One can also observe that all the SSE-based metrics perform very similarly for ranking the models, except for the RMSLE that shows very different results for the heating power usage comparison. The latter could be due to the asymmetry of the RMSLE penalization of over-estimations and under-estimations. It indicates that RMSLE is probably not an adequate metric for BEM comparison.

Regarding the advanced elastic distance metrics, their ranking is very similar to that of the SSE-based metrics for the indoor temperature comparison. For heating power usage, however, agreement within the elastic distance metrics group or with the SSE-based group is much less clear, with the Frechet distance diverging clearly. Further investigations are needed to identify which elastic distance metrics are the best suited for BEM TS comparison.

Finally, the ranking of the CVRMSE 24h-Amp is significantly different from that of all the other metrics. This is intended since its purpose is to specifically analyse the goodness of fit of the daily amplitude rather than that of the data points themselves.

DESTEST comparison tool

The Python-based TS comparison tool has been developed for the DESTEST participants to easily compare their simulation results to each other when working on the same common exercise. The participants are encouraged to use the tool from early stage to quickly receive feedback about how their simulation results fall in comparison to the other participants who have uploaded their vetted results. The tool is intended to give a synthetic overview of how far is the user's data from the reference TS (key CMs), but also provide informative figures allowing for more detailed analysis. If the user finds its results to be outliers among the pool of data generated by the other participants (large differences with the reference TS), the former is encouraged to analyse further its model to identify possible modelling errors. When the simulation results are deemed to be correct, the user can upload them to a dedicated online GitHub repository. These new results are thus integrated into the pool of vetted data used for the generation of the reference TS.

Three key CMs are suggested by default for this comparative analysis (presented above). However, most of the CMs presented in this article can also be used (selected) in the comparison tool. Because the performance ranking of the different BEMs is most probably different for each CM and each comparison simulated variable, a summary *Accuracy Grade* is calculated as follows: For each simulated variable of interest and for each CM calculated for the latter, the model with the best CM score gets a grade of 100% and the model with the worst CM score gets a grade of 0%. The remaining models receive a grade between 0% and 100% that is linearly proportional to the distance of their score from the "best" and "worst" models. All these

grades are then averaged to form the summary *Accuracy Grade*. The user can assign a weighting factor to the different CMs. The summary *Accuracy Grade* is then calculated as a weighted mean average. By default, all CMs have a weighting factor of 1. If a model scores the best CM for all variables of interest, its *Accuracy Grade* is 100%. Conversely, if a model scores the worst CM for all variables, its *Accuracy Grade* is 0%. However, it is important to note that this *Accuracy Grade* is only intended to rank the models of the data pool in between each other. A low *Accuracy Grade* indicates that the model is an outlier relative to the other models, but it does not necessarily mean that the model is flawed, especially if there are not many models in the data pool.

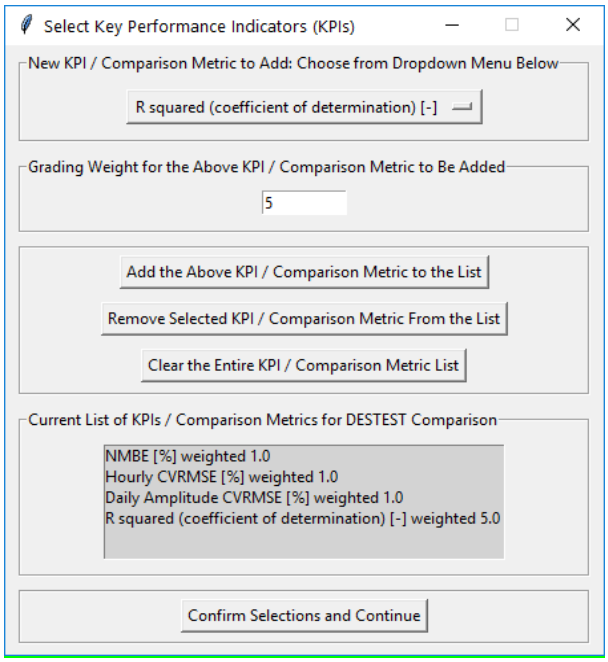

Figure 5: Tool interface for selection of the comparison metrics and weighting factors.

The TS comparison tool has a simple graphical user interface to select a specific common exercise case, select the result data file of the user (optional), and choose the different CMs and associated weighting factors that are used for the comparison process (see Figure 5). The Python-based source code and detailed documentation of the comparison tool, together with all information and DESTEST result data can be found on the dedicated GitHub: https://github.com/ibpsa/project1-destest.

Table 3: Output result table with comparison metrics and summary accuracy grade (colours are only intended to emphasize extrema).

Simulated Variables	Comparison Metrics	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Heating Power	NMBE [%]	-4.84	-17.47	0.22	6.18	29.12	17.69	-30.9
Heating Power	Hourly CVRMSE [%]	19.08	35.42	30.32	54.86	44.97	34.13	51.53
Heating Power	CVRMSE 24h-Amp [%]	74.32	118.73	11.94	23.27	31.15	61.96	87.92
Ground Floor Temperature	NMBE [%]	0.11	0.32	-0.22	0.01	-0.35	0.01	0.12
Ground Floor Temperature	Hourly CVRMSE [%]	0.23	0.47	0.35	0.23	0.49	0.14	0.29
Ground Floor Temperature	CVRMSE 24h-Amp [%]	11.89	11.85	21.17	22.64	31.26	11.77	23.38
First Floor Temperature	NMBE [%]	0.07	0.24	-0.18	0.08	-0.07	0.01	-0.15
First Floor Temperature	Hourly CVRMSE [%]	0.15	0.38	0.28	0.24	0.2	0.2	0.33
First Floor Temperature	CVRMSE 24h-Amp [%]	7.3	20.91	15.83	29.98	15.2	44.84	31.58
Summary Accuracy grade [%]		82.74	30.66	60.6	62.06	38.51	70.27	33.29

The output of the comparison tool is a report with comparison tables gathering the results of all CMs for all variables of interest (see Table 3), and the basic statistical properties of the data sets (see Table 4). The tables are followed by several figures illustrating the detailed comparison of the different models. One can see some of these figures hereafter (see Figure 6-10).

Table 4: Output result table with basic statistical properties of the data sets (colours are only intended to emphasize extrema).

Simulated Variables	Metrics	Reference	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Heating Power	Minimum [W]	-0.48	0	0	0	0	0	-3.38	0
Heating Power	Maximum [W]	13729.28	16573	16573	13404.7	15880.2	16972.2	16471	18223.9
Heating Power	Mean Average [W]	2186.17	2080.4	1804.17	2190.92	2321.23	2822.78	2572.93	1510.74
Heating Power	Standard Deviation [W]	2975.48	3102.47	3054.78	2885.72	3250.96	3506.68	3430.8	2316.73
Ground Floor Temperature	Minimum [K]	289.17	289.13	289.14	289.15	289.13	289.15	289.3	289.15
Ground Floor Temperature	Maximum [K]	302.85	304.7	305.41	299.89	304.02	301.75	302.93	302.15
Ground Floor Temperature	Mean Average [K]	293.5	293.83	294.44	292.86	293.51	292.46	293.51	293.85
Ground Floor Temperature	Standard Deviation [K]	3.08	3.62	3.92	2.58	3.28	2.69	2.92	3.02
First Floor Temperature	Minimum [K]	289.14	289.12	289.13	289.15	289.13	289.15	289.13	289.15
First Floor Temperature	Maximum [K]	303.8	304.58	305.16	302.22	305.61	303.95	305.2	301.3
First Floor Temperature	Mean Average [K]	293.57	293.78	294.27	293.04	293.79	293.37	293.6	293.13
First Floor Temperature	Standard Deviation [K]	3.18	3.48	3.72	2.87	3.44	3.04	3.19	2.91

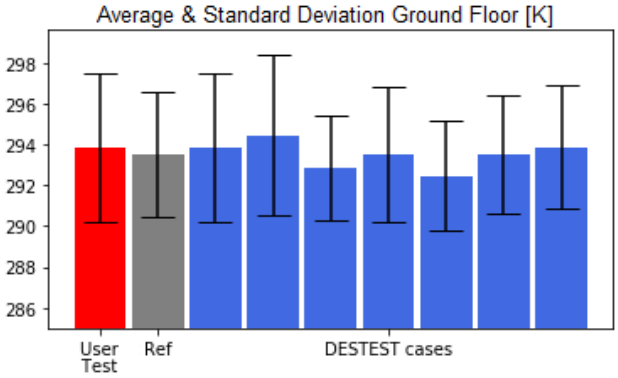

Figure 6: Average and standard deviation of the ground floor indoor temperature.

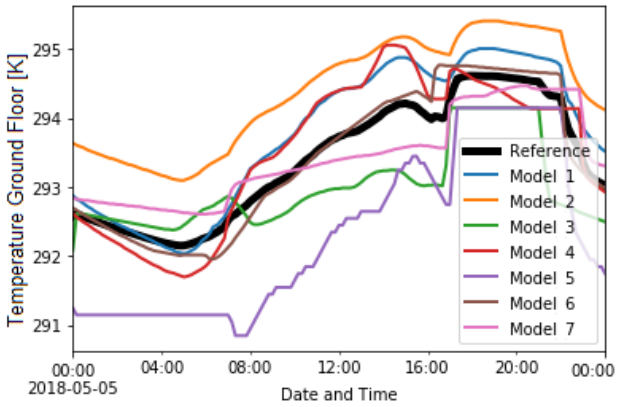

Figure 7: Ground floor indoor temperature for the selected day 5th of May.

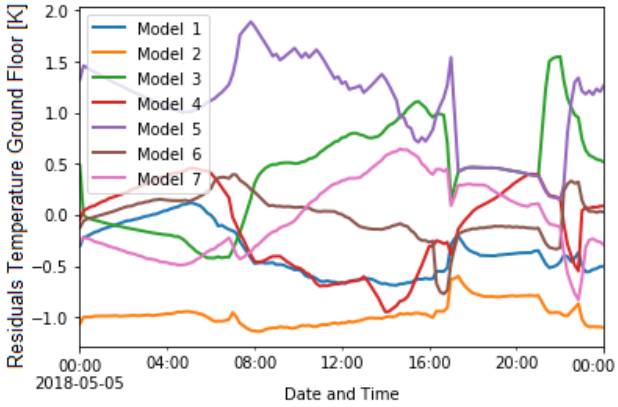

Figure 8: Model residuals for the ground floor indoor temperature for the selected day 5th of May.

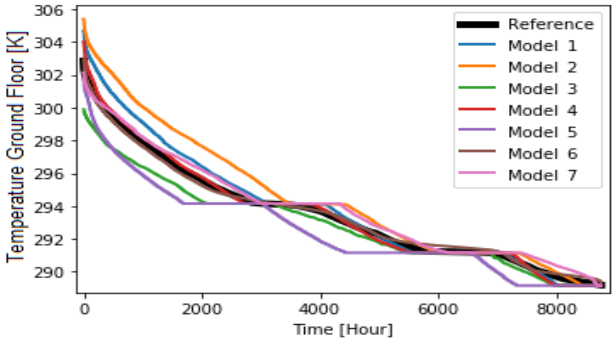

Figure 9: Annual load duration curve for the ground floor indoor temperature (data points sorted in descending order).

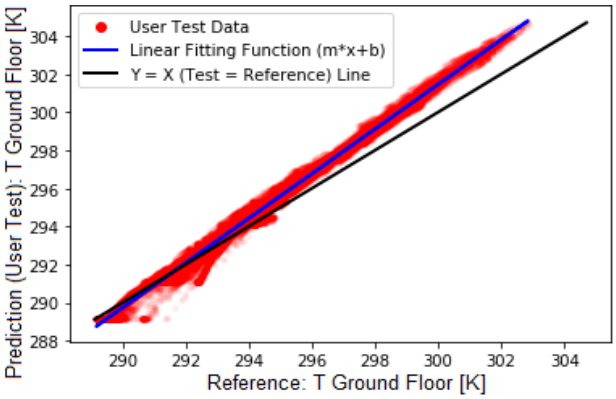

Figure 10: Prediction vs reference plot for the ground floor indoor temperature of the user data.

Conclusions

In this article, different methods for comparing results (in the form of time series) of BEMs are discussed. In the case of inter-model comparison without empirical reference case, it is suggested to build a reference time series as a point-to-point mean average of all the models to be compared (ensemble methods). However, this approach should not be preferred over measurement-based reference when the latter exists. Different simple comparison metrics and advanced time series distance metrics are then reviewed and tested on simulation data from a common exercise on modelling single-family houses. Those metrics are grouped into four categories. Within these categories, most of the metrics behave similarly. For the simple comparison metrics, it is suggested to not use RMSLE but rather use NMBE and CVRMSE. CVRMSE of hourly-averaged data and daily amplitude of data are also suggested as simple metrics that do not over-penalize small time shifts and scrutinize building dynamics, respectively. However, these conclusions are drawn for only two types of virtual sensors which are commonly used for BEM analysis: indoor temperature and energy use. The conclusions might change for simulated variables with different patterns and dynamics. Finally, a Python-based time series comparison tool is presented. It is intended to help the modelling community to compare simulation results and thus improve the accuracy of building models.

In the near future, the work presented in this article will be extended as follows. The ensemble methods for building a reference will be tested further. More comparison metrics will be reviewed, analysed and tested with larger data sets. Advanced time series distance metrics will be thoroughly studied. The analysis will also include various virtual sensors of urban-scale energy systems such as district heating/cooling networks. New metrics and functionalities will be integrated into the DESTEST comparison tool.

Acknowledgement

This work was financed by the InterHUB project (www.interhub.aau.dk). This work emerged from the IBPSA Project 1, an international project conducted under the umbrella of the International Building Performance Simulation Association (IBPSA).

References

American Society of Heating, Refrigerating and Air-Conditioning Engineers (2014). *Measurement of energy, demand, and water savings (ASHRAE Guideline 14-2014)*.

American Society of Heating, Refrigerating and Air-Conditioning Engineers/American National Standards Institute (2017). *Standard Method of Test for the Evaluation of Building Energy Analysis Computer Programs (ANSI/ASHRAE Standard 140-2017)*.

Coakley, D., Raftery, P., Keane, M. (2014). A review of methods to match building energy simulation models to measured data. *Renewable and Sustainable Energy Reviews 37*, 123–141.

Galton, F. (1907). Vox populi. *Nature 75*, 450–451.

IBPSA (2017). IBPSA Project 1 website: https://ibpsa.github.io/project1/.

Jose, V.R.R. and Winkler, R.L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting 24*(1), 163–169.

Kvålseth, T.O. (1985). Cautionary Note about R2. *The American Statistician 39*(4), 279–285.

Mallows, C. (1991). Another comment on O'Cinneide. *The American Statistician 45*(3), 257.

Mori, U., Mendiburu, A., Lozano, J.A. (2016). Distance measures for time series in R: The TSdist package. *The R Journal 8*(2), 451–459.

Ruiz, G.R. and Bandera, C.F. (2017). Validation of calibrated energy models: common errors. *Energies 10*, 1587.

Saelens, D., De Jaeger, I., Bünning, F., Mans, M., Vandermeulen, A., van der Heijde, B., Garreau, E., Maccarini, A., Rønneseth, Ø., Sartori, I., & Helsen, L. (2019). Towards a DESTEST: a District Energy Simulation Test Developed in IBPSA Project 1. *Proceedings from the 16th IBPSA Conference BS2019*, 3569-3577. Rome (Italy), 2-4 September 2019.

Yi, S.K.M., Steyvers, M., Lee, M.D.; Dry, M.J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science 36*(3), 452–470.