What Metrics Does the Building Energy Performance Community Use to Compare Dynamic Models?

Hicham Johra[1], Markus Schaffer[1], Gaurav Chaudhary[2], Hussain Syed Kazmi[3], Jérôme Le Dréau[4], Steffen Petersen[5]
[1]Aalborg University, Aalborg, Denmark
[2]Norwegian University of Science and Technology, Trondheim, Norway
[3]KU Leuven, Leuven, Belgium
[4]La Rochelle University, La Rochelle, France
[5]Aarhus University, Aarhus, Denmark

Abstract

Comparing, validating and assessing the accuracy of dynamic models is crucial for multiple applications in the field of energy, buildings and indoor environmental engineering. To that matter, various comparison metrics and key performance indicators have been developed or borrowed from other fields of science, and a few popular guidelines have recommended some of them for building energy models.

This article aims at giving an overview of what metrics are used by the community of researchers in the field of energy, buildings and indoor environment. This overview is based on a large-scale review work of 259 scientific publications from the last 40 years. This paper also discusses the main trends, reveals certain gaps and suggests several research activities currently being undertaken by a multi-institutional working group of researchers, which should greatly benefit the entire community of building energy and indoor environment simulation.

Highlights

- Review of 259 publications about energy and building model testing, comparison and validation.
- Overview of comparison metrics used by the building energy performance community to test, compare and validate numerical models.
- Analysis of the most popular comparison metrics.

Practical implications

This review identifies clear trends in the practices of the building energy performance community regarding the testing, comparison and validation of numerical simulations. Popular comparison metrics are analyzed, and their shortcomings are pointed out. This analysis can guide researchers in selecting appropriate comparison metrics for numerical model validation and suggest future work for the development of robust validation methods in the field of building physics.

Introduction

Comparing, validating and assessing the accuracy of dynamic models is crucial for multiple applications in the field of energy, buildings and indoor environmental engineering. The output results of these dynamic simulations are most often in the form of time series. The quality assessment and validation of such dynamic models thus consist in determining how different the output result time series from a simulation are when compared to a reference time series (Johra et al., 2021).

Such metrics and key performance indicators have been developed or adopted from other fields of science, and additionally, a few popular guidelines, such as the ASHRAE Guideline 14-2014, the IPMVP (2014) and the FEMP, M&V Guidelines (2015), have recommended some comparison metrics and criteria for building energy modelling testing and validation. Despite the clear importance and influence of such metrics on the model quality, no large-scale comparison review of them has been published to the best of the authors' knowledge, and thus discussions behind the choice of adequate comparison metrics are very seldom and not supported by data. One reason is presumably the human-labour-intensive nature of such a task since there are no easy ways to automate the search and categorisation of equations in scientific publications, especially with large variations in the formulations, naming and acronyms. Such difference in definition, naming, and acronyms not only hinders an automated search but also increases the likelihood of misinterpretation of results by fellow researchers, exacerbates the risk of misunderstandings, and thus potentially hinders research.

In other research communities, such comparisons exist. For example, Lepot et al. (2017) have recently compared metrics for interpolating time-series data, Prema et al. (2021) provided an overview and comparison of metrics for wind and solar power forecasting, while Hewamalage et al. (2022) provide a general overview for forecasting.

However, the adequacy of comparison metrics might be greatly influenced by the dynamic properties of the evaluated time series, e.g., the sampling rate, signal amplitude, frequency spectrum, unit scale (e.g., K, °C, °F), or closeness to the 0 of such unit scale. An analysis of comparison metrics focusing specifically on the building physics and building energy modelling context and supported by data is thus important for the building industry and research community and the IBPSA audience in particular.

This article aims to close this knowledge gap for the dynamic modelling of energy, buildings and indoor environment by providing an overview and discussions on comparison metrics that are based on the review of 259

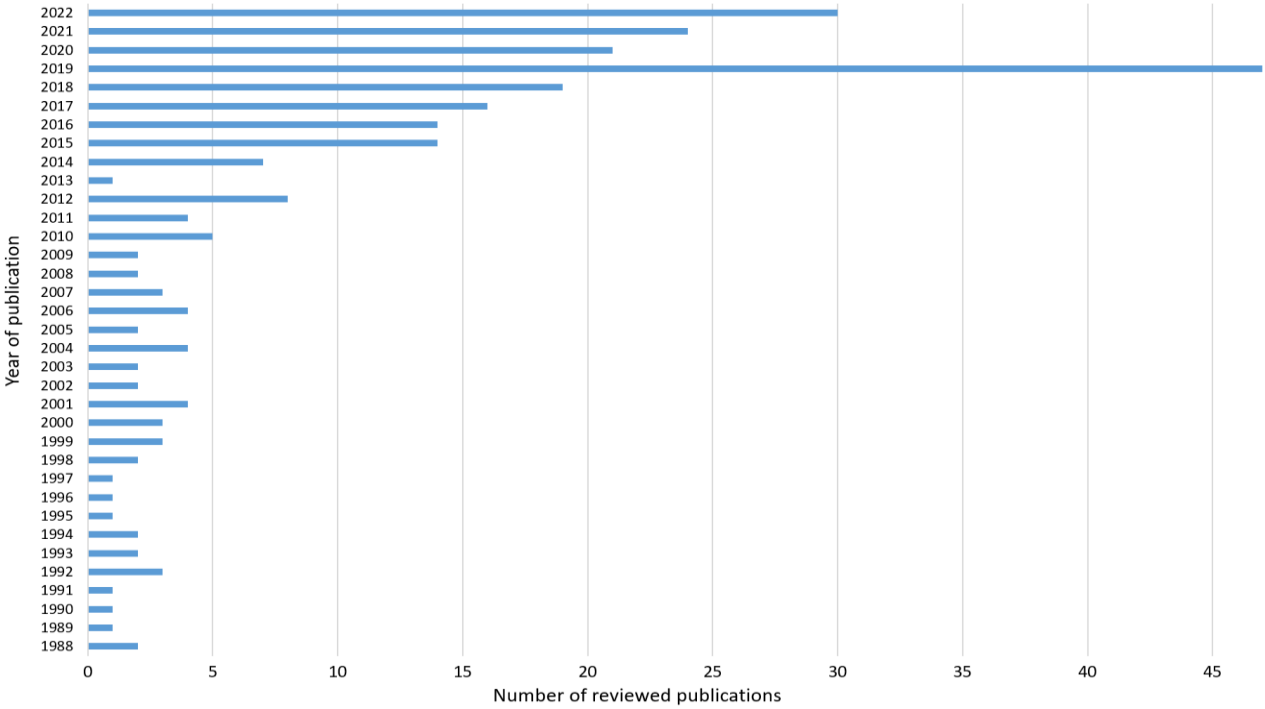

Figure 1: Temporal distribution of the reviewed publications.

scientific publications over the last 40 years. Furthermore, main trends are discussed, gaps are revealed, and future research activities are suggested.

All the references and data collected for this study have been compiled and curated and are available in open access (Johra et al., 2023). Additionally, a unified definition and notation for the 48 metrics found in the review process are provided in supplementary materials (Johra et al., 2023). The supplementary materials to the current study can be directly accessed here: https://doi.org/10.54337/aau533917780

Combined with the provided overview and discussions, such unified notation and definition should greatly benefit the entire building energy simulations community.

Scope and methodology of the current literature review

The conducted structured literature review focused on scientific papers using deterministic building physics (or related) models for dynamic (time-dependent) variables such as temperature, energy demand, heating/cooling/electricity demand, CO2 concentration, relative humidity, fluid mass flow rate, or heat flow in building elements.

The review focuses primarily on peer-reviewed scientific journals and conference proceedings in the fields of energy in buildings, and indoor environment, in which simulation result time series are being compared to reference time series (e.g., empirical reference data) or compared to other numerical dynamic model results. These time series comparisons are intended to assess model accuracy and/or validate the correctness (and thus usefulness) of a tested simulation model.

The literature search covered both journal articles and conference proceedings. The Scopus database was used to search for documents in the field of building physics using the following keywords: *building** AND *energy* AND "*simulation*" OR "*model"* AND *compar** OR *valid** OR *accura** OR *error*. Based on a preliminary screening of the article content, the latter is added to the list of valid documents for thorough review and analysis or disregarded. In addition, searches were carried out over specific time periods to cover the last 40 years of literature in the field of energy and indoor environment modelling. Moreover, the proceedings of the IBPSA, NSB and IEEE conferences were specifically screened using similar keywords.

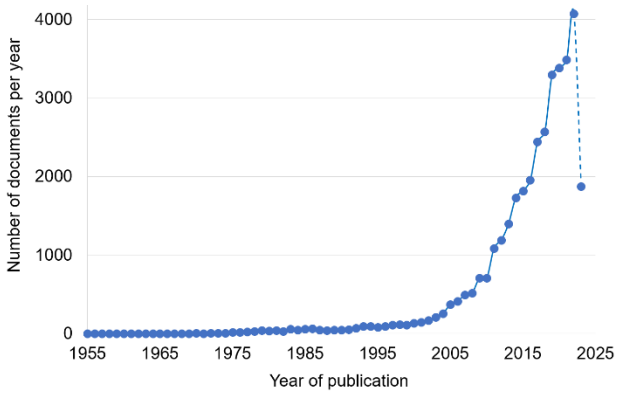

Figure 2: Results of the Scopus search with keywords building AND energy AND "simulation" OR "model" AND compar* OR valid* OR accura* OR error.*

Time series forecasting with probabilistic models is defined to be outside the scope of the current work but should naturally be included in further studies expanding this first research effort. Moreover, the included metrics

are always used for model evaluation but not for case or scenario comparison nor for model training (i.e., hyperparameter estimation in the training process). Model training methods might use other custom loss functions and specific assessment metrics that are out of the scope of this paper.

Overview and statistical insights on the reviewed literature

This section gives an overview of some key characteristics and certain statistical insights on the body of reviewed literature in the field of building model testing and validation. This can inform on modelling practices of the building energy and indoor environmental engineering research community.

The analysed publications cover the last four decades so that a representative picture of the development and usage of metrics can be drawn. However, due to practical reasons and the recent massive expansion of the number of scientific publications in the field of building physics (see Figure 2), work before 2010 is underrepresented (see Figure 1).

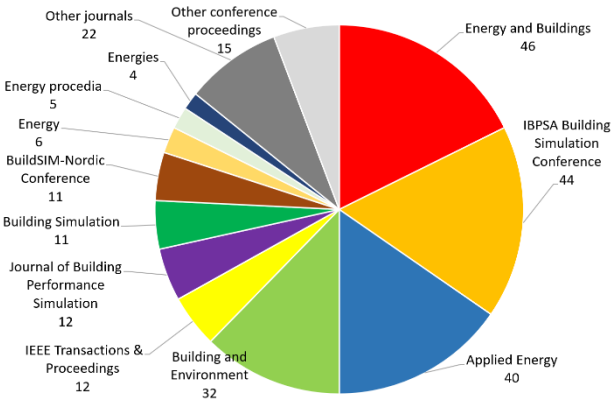

Figure 3: Overview of the different sources for the reviewed publications.

One can see in Figure 3 that most of the reviewed publications are published in the peer-reviewed scientific journals *Energy and Buildings*, *Applied Energy*, *Building and Environment*, and the proceedings of the *IBPSA Building Simulation* conference. The *other journals* source category includes, e.g., ASHRAE publications, *Energy Conversion and Management*, *International Journal of Heat and Mass Transfer*, *Journal of Solar Energy Engineering* or *Renewable Energy*. The *other conference proceedings* source category includes, e.g., *American Control Conference*, *Journal of Physics: Conference Series* or *Nordic Symposium on Building Physics*.

Figure 4 presents the distribution of the main topic of applications for the numerical models in the reviewed publications. More than half of the publications focus on the energy demand simulation of single buildings or clusters of buildings. This dominating trend can be attributed to the large interest, need and funding for estimating, forecasting and explaining the significant share of the total energy demand accounted for by the global building stock.

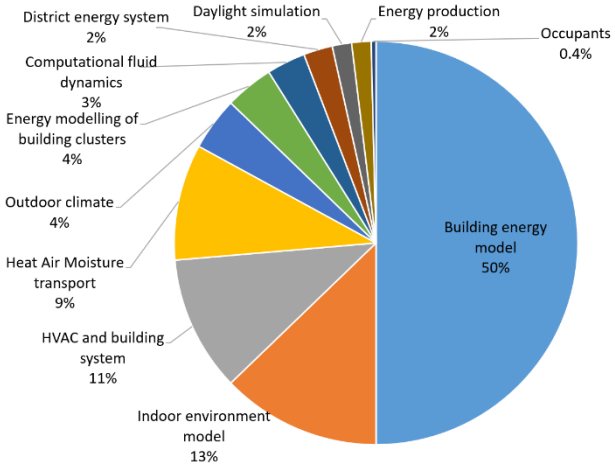

Figure 4: Overview of the main application of the numerical models in the reviewed publications.

As shown in Figure 5, the most popular modelling tools in the reviewed publications are *Energy Plus*, *MATLAB* (custom-made code or existing libraries/packages), *TRNSYS*, Modelica and *IDA ICE*. The *other* modelling tool category includes *DIMOSIM*, *ENVI-met*, *PHPP*, *Radiance*, *COMIS* or *STAR-CCM+*.

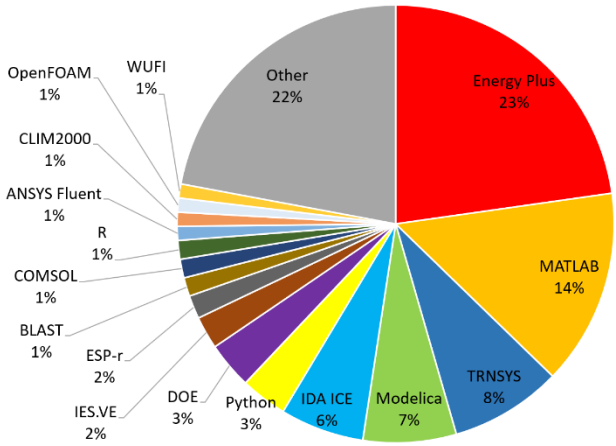

Figure 5: Overview of the numerical simulation tools employed in the reviewed publications.

One can see in Figure 6 that the majority of the simulation tools used in the reviewed publications follow a *White Box* modelling paradigm.

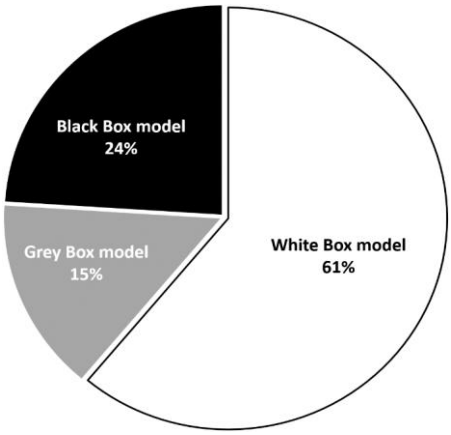

Figure 6: Distribution of the modelling approaches/paradigms in the reviewed publications.

This is clearly correlated to the modelling approach employed in the most popular numerical tools. The *White Box* approach also includes Computational Fluid Dynamics (CFD) and analytical solutions of heat and mass transfer equations. The vast majority of the *Grey Box* models in the reviewed studies are low-order Resistance-Capacitance (RC) networks. A large share of the *Black Box* models are Artificial Neural Networks (ANNs). There are also many linear regressions and auto-regressive models like ARMA, ARMAX or ARIMA.

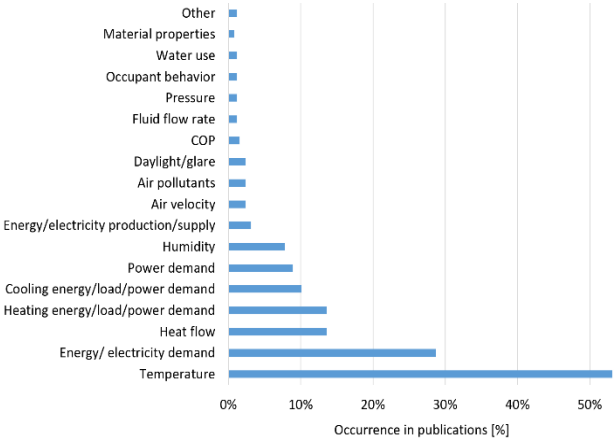

Figure 7: Occurrence distribution of the different analyzed variables of interest in the reviewed publications.

Figure 7 provides insights into what simulated variables are typically analyzed in the reviewed publications. One can clearly observe that, in the case of model comparison, the temperature of the indoor environment, outdoor environment and building systems are systematically analyzed in more than 53% of the publications. Different forms of energy demand (i.e., hourly, sub-hourly or daily heating energy, cooling energy and electricity demand) represent more than 52% of the analyzed variables, which is in line with the main modelling focus emphasized in Figure 4.

If the energy demand is a common simulation result of interest (typically computed as Wh per hour or Wh per day), the heating, cooling or electrical power demand and peak demand (in W) are only compared in less than 10% of the cases.

Results: trends in building modelling comparison metrics

In this review study, the model comparison approach is systematically assessed for the different publications. One can see in Figure 8 that the researchers include a graphical display of the simulation result times series in 86% of the reviewed publications. In 29% of the cases, the model comparison is only based on a qualitative assessment in the form of, most of the time, overlaid time series figures, or sometimes boxplots or predictions-reference plots. These graphical qualitative model comparisons are thus often accompanied by a statement like "the model is in good agreement with the empirical data". If this was common practice in the 80s, 90s and early 2000s, the absence of quantifiable indicators on how well this agreement really is, largely prevents proper reproducibility and comparison between studies.

In 57% of the cases, however, the publication presents a graphical comparison of the model performance together with one or several comparison metrics. More seldomly, only quantitative comparison metrics are presented to justify the model's accuracy without any graphical display. As the lack of quantifiable indicators was a drawback of the early studies, the lack of visual comparison may mask a poor fit (despite quantitative indicators). However, one should note that these results are highly skewed towards recent publications following 2010.

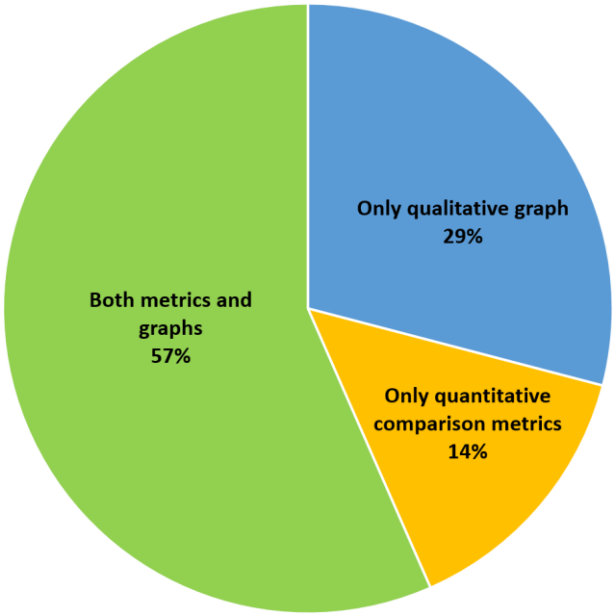

Figure 8: Time series comparison approach for the testing and validation of numerical models in the reviewed publications.

When looking at the historical perspectives of these comparison practices, one can see in Figure 9 that there is a clear evolution in the reviewed publications: older studies (80s, 90s and early 2000s) tend to use only graphical qualitative assessment to report their numerical models' adequacy. This practice has drastically changed after 2014 with the systematic use and reporting of quantitative accuracy performance indicators in the validation of numerical models for energy in building and indoor environmental engineering.

This trend coincides with the publication of the ASHRAE guideline 14-2014, the publication of the IPMVP-Core Concepts (EVO, 2014) and the FEMP, M&V Guidelines 2015, which provided benchmarking methods and the possibility for the building community to use a set of recommended comparison metrics for simulation result time series. 26% of the recent reviewed publications mention or refer to the ASHRAE Guideline 14.

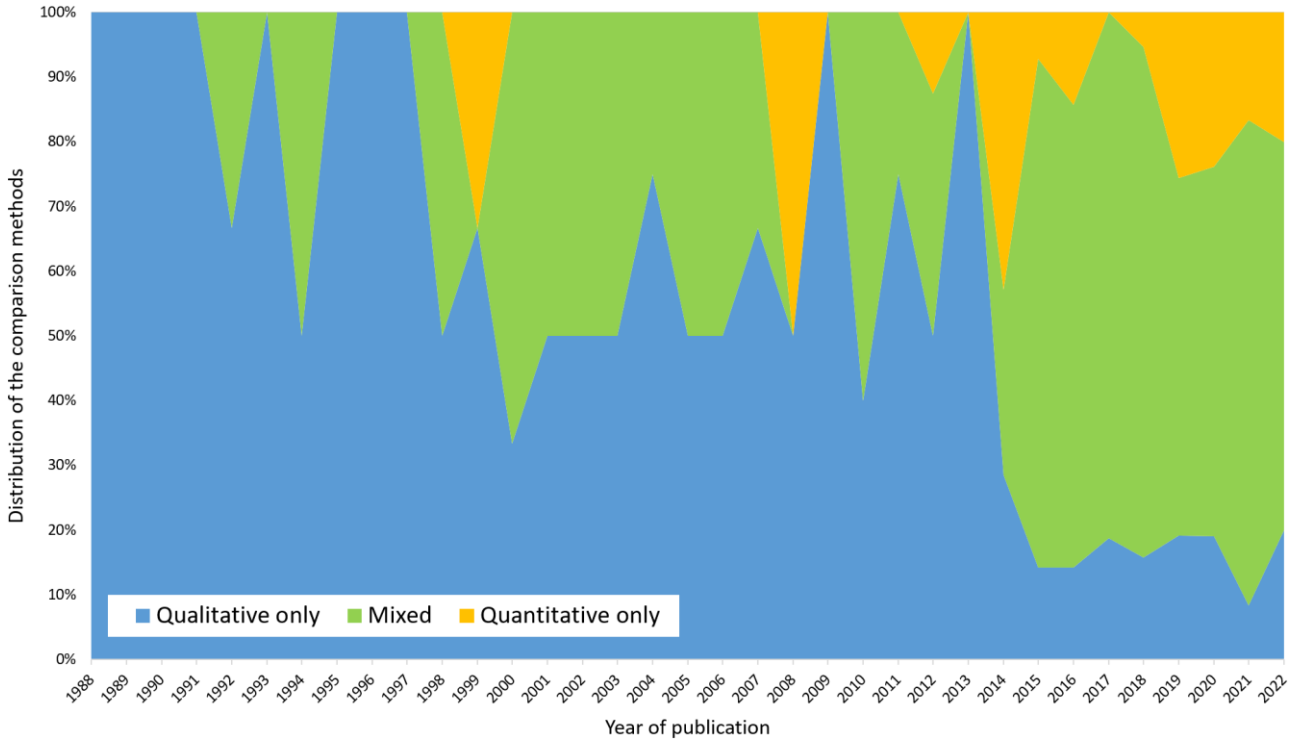

Figure 9: Historical evolution of building model comparison practices in the reviewed publications.

This current review study identified 48 different metrics for comparing simulation time series in the considered publications. The vast majority of these metrics are point-to-point comparison ones, such as the Mean Bias Error (MBE) or the Sum of Squared Errors (SSE). These point-to-point comparison metrics are usually simple to compute, but they assume perfectly synchronized time series data points and regular/constant sampling rates. This is a clear limitation when there is a certain offset between the test time series and the reference one, especially if there are multiple peaks in the signal (Johra et al., 2021). To overcome this limitation and tackle the risks of over-penalization of models in that situation, a few publications report time series elastic distances calculation and shape comparison instead of simpler point-to-point metrics. Examples of these elastic distance metrics and general elastic shape comparison found in the considered literature are the Dynamic Time Warping, dissimilarities based on Pearson's correlation, and the Frechet distance.

In addition, some point-to-point comparison metrics are not applied directly to the entire time series of the building variable of interest but to a transform of that time series. For instance, Panão et al. (2016) compute the Mean Absolute Error (MAE) of the daily max of the time series, and Johra et al. (2021) calculate the Coefficient of Variation of Root Mean Square Error (CVRMSE) of the daily amplitude of the signal (from midnight to midnight each day).

The systematic counting of the comparison metrics in the review publications (see Figure 10) reveals a clear dominating use of 7 popular metrics:

- MBE: Mean Bias Error
- NMBE: Normalized Mean Bias Error
- MAE: Mean Absolute Error
- MAPE: Mean Absolute Percentage Error
- R^2: Coefficient of determination
- RMSE: Root Mean Square Error
- CVRMSE: Coefficient of Variation of Root Mean Square Error

The other less popular metrics with occurrence above 1% are: MaxAE (Maximum Absolute Error), MaxAPE (Maximum Absolute Percentage Error), NMAE (Normalized Mean Absolute Error), SSE (Sum of Squared Errors), MSE (Mean Square Error), RNRMSE (Range Normalized Root Mean Square Error), RMSEP (Root Mean Square Error Percentage), RMSLE (Root Mean Square Logarithmic Error), Pearson correlation coefficient, Spearman's rank correlation and GOF (tailored goodness of fit function consisting in different combinations of other metrics like MBE, NMBE, RMSE and CVRMSE).

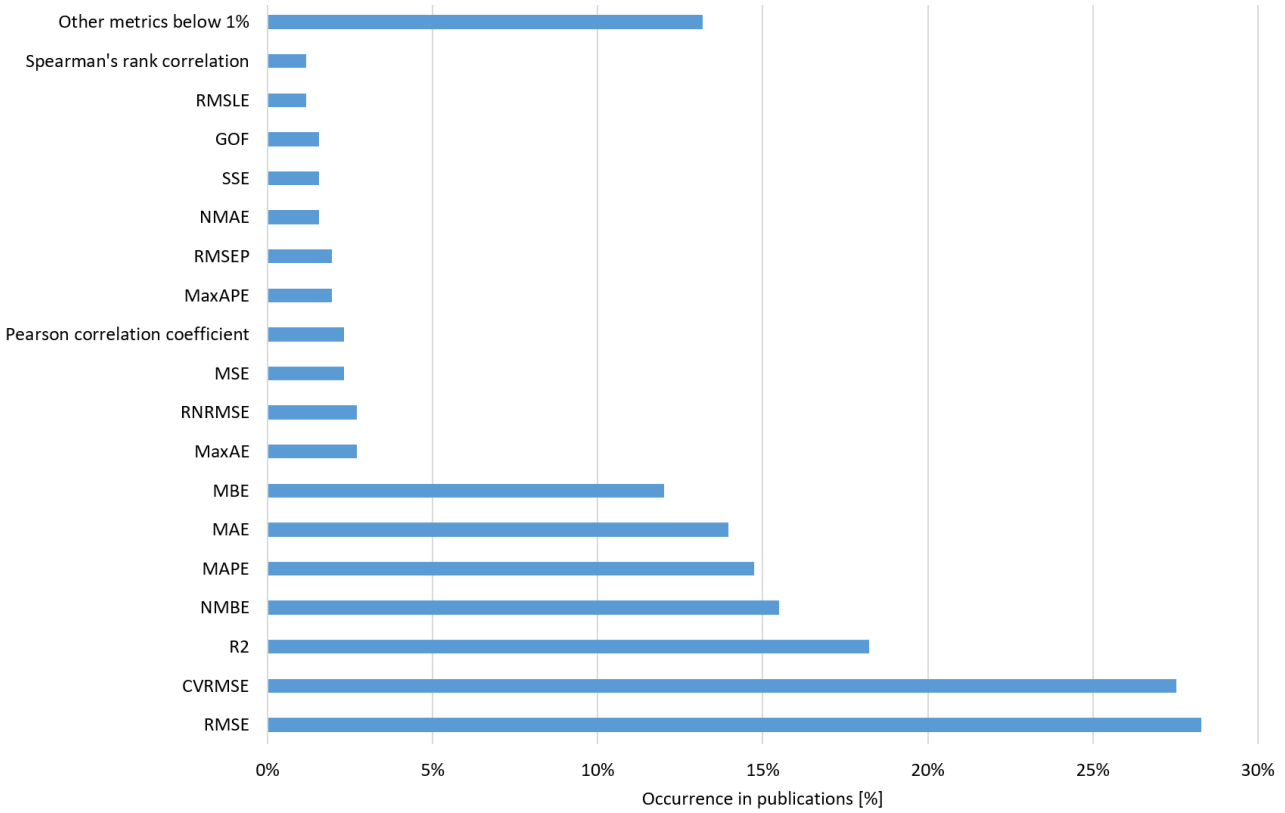

Figure 10: Occurrence of the most popular time series comparison metrics in the reviewed publications.

When focussing on the main variables of interest individually, the review reveals the following popular choices (sorted by decreasing order of occurrence frequency):

- Energy demand or supply/production (Wh per year/month/day/hour): CVRMSE, R^2, NMBE
- Power demand (W): CVRMSE, RMSE, NMBE
- Temperature (°C, °F or K): RMSE, R^2, CVRMSE
- Heat flow (W or W per m^2): RMSE, MAPE
- Moisture content (% relative humidity or kg per kg): R^2, CVRMSE
- Water consumption (L/m^3 per year/month/day/hour): NMBE, R^2, RMSE
- Air pollutant (e.g., CO2 or VOC concentration): MAE
- Daylight/glare discomfort: R^2
- Material properties: RMSE
- COP (coefficient of performance) for HVAC systems: RMSE, CVRMSE, NMBE

Discussions

Before about 2014, there was a clear tendency that mainly qualitative model assessment was used based on time series graphs and a subjective definition that the model agrees well with the (validation) data. From about 2014 on, a clear paradigm shift is visible towards using both qualitative and quantitative assessment, which can be seen as a significant increase in scientific objectivity and transparency. This trend also coincides with the publication of impactful international guidelines (ASHRAE guideline 14-2014, IPMVP-Core Concepts 2014 and the FEMP, M&V Guidelines 2015), which provided benchmarking methods and recommended comparison metrics with validity thresholds for the building community. These guidelines are assumed to be one of the significant reasons for the frequent use of the MBE, NMBE, RMSE and CVRMSE. While these recommended metrics, combined with the suggested thresholds, eased the comparison between different simulation performance reports, they are no panacea.

The MBE and NMBE indicate the global bias of the model (global under-prediction or global over-prediction). However, these metrics are prone to cancellation or compensation effects: local biases in opposite directions compensate each other, i.e., local model under-estimations would compensate for local over-estimations, leading to a globally low MBE and NMBE despite large local discrepancies. Metrics based on squared differences (e.g., RMSE, CVRMSE) or absolute value of differences (e.g., MAPE) are not subjected to compensation effects.

Metrics that are not normalized (e.g., SSE, RMSE) do not allow for the comparison of models on datasets of different sizes or with different unit scales. Normalized metrics should thus be preferred. However, certain normalized metrics have the same or similar names (usually comporting a “normalized” term) but actually do not use the same normalization logic. This is, e.g., the case

for the NRMSE (Normalized RMSE) that is actually normalized by the average of the reference time series in certain publications and is thus the same as the CVRMSE, but, in some other publications, the NRMSE is normalized by the amplitude of the reference time series. These misleading namings can cause severe confusion.

Metrics based on the (squared) difference and normalised by the "total" mean (e.g., NMBE and CVRMSE) do not consider changes in the magnitude of a quantity over time. They are thus biased towards periods with higher magnitudes and, therefore, not necessarily suited in the presence of strong seasonal variation commonly found in, e.g., heating or cooling energy demand time series.

Likewise, widespread metrics such as the RMSE, MAE, MAPE, and MBE suffer the same limitations and are biased towards high magnitude periods and, in the case of the RMSE, are sensitive to outliers/measurement errors.

Certain metrics are very sensitive to values that are close to the 0 of the time series unit scale. This is highly problematic as building energy demand profiles for heating or cooling, when taken separately, are very likely to be at 0 for extended periods of time over the year. The RMSEP and the very popular MAPE are undefined if the data contains zeros.

Certain comparison metrics are also very sensitive to outliers with significant over-penalization effects, which might not be desirable for global model validation. For instance, the CVRMSE emphasises large deviations (due to the squared difference) and can easily lead to poor reported accuracy if, e.g., outliers or measurement errors are present in the (validation) data.

The RMSLE is less sensitive to large outliers, in comparison to, e.g., the RMSE, because it penalizes much less very large differences between the tested model and the reference when both the prediction and the reference are large numbers. However, the RMSLE penalizes more the model under-estimations than the model over-estimations, which is not necessarily a desired feature for building applications.

Regarding the definition and implementation of the popular metrics, it must be noted that the ASHRAE guideline 14-2014 recommends using *n-1* (with *n* the number of samples) for calibration purposes. Yet the current review has found that both *n-1* and *n* are commonly used in the research community. While this is expected to have a negligible impact for most cases, it highlights possible sources for misunderstandings, particularly if no clear definitions are provided, which is frequently the case.

R^2 is another very common comparison metric, yet one that is often misunderstood and misused. Indeed, there exist multiple definitions and formulations of R^2, and they are not all necessarily equivalent, which can lead to significant interpretation mistakes (Kvålseth, 1985). In addition, R^2 does not consider the model complexity and is biased toward high-magnitude periods. The problem of the model complexity can be overcome with the adjusted R^2. Another practical issue is that some R^2 implementations in popular software (e.g., *Scikit-Learn* in Python) can take values below 0 if a model performs worse than the mean of the data would, which is not the case for other R^2 definitions and implementations. This makes the comparison between studies possibly challenging.

Finally, as mentioned in the previous section, point-to-point comparison metrics can over-penalized models when slight time shifts of signal peaks are presented in the test datasets. This behaviour might not be desirable and could be tackled by computing elastic distance and shape comparison metrics.

Conclusion and suggestions for future work

In this work, 259 scientific papers from energy, buildings and indoor environment modelling have been reviewed and analysed regarding their used model evaluation metrics and approach.

Overall, the collected comparison metrics vary in their definition, notation and abbreviations between research papers, leading to possible confusion, misinterpretation of results and misunderstandings. Thus the modelling research community should bundle efforts to establish universal names and definitions. Furthermore, the revised metrics made it obvious that currently, the recommended and most commonly used metrics suffer from various flaws, such as bias towards periods with higher magnitude for quantities with strong seasonal variation, outlier hyper-sensitivity, compensation effect, or impossibility to compute when the considered variables are equal to zero.

Thus future research efforts should focus on establishing more robust comparison metrics adapted to the signal features of specific variables of interest and with clear and unambiguous definitions.

The key conclusions from this study and recommendations for model comparison in the field of building physics can be summarized as follows:

- Qualitative comparison with time series graphical visualization should always be included with different time scales and with good readability (choice of colours and markers)
- Normalized metrics are prefered over absolute metrics for quantitative comparison.
- The equation of the used metrics should always be provided along with the evaluation period, and information on the data treatment for zero-values
- For error evaluation, CVRMSE, RMSE, MAPE and MAE are commonly used metrics
- For bias evaluation, NMBE and MBE are commonly used metrics
- Elastic distance metrics (e.g., Dynamic Time Warping or Frechet distance) should be considered for further analysis

The review and analysis work presented in this paper has led to a unified and coherent definition and notation for the 48 reviewed metrics (Johra et al., 2023). These different metrics will then be systematically tested with well-defined datasets from building physics. This will

thus allow identifying similar behaviour and pitfalls among the reviewed metrics and thus guide practitioners in the selection of adequate comparison tools for specific types of time series generated by building models.

Finally, this review has explicitly focused on deterministic models. Hence probability-based models, which have become increasingly popular, and forecasting, should be included in further extensions of this work.

Acknowledgement

This work has been carried out within the framework of the International Energy Agency (IEA) Energy in Buildings and Communities (EBC) Annex 82: "Energy Flexible Buildings Towards Resilient Low Carbon Energy Systems" (https://annex82.iea-ebc.org/). The authors would like to gratefully acknowledge the IEA EBC Annex 82 for providing an excellent research network and thus enabling fruitful collaborations.

References

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

FEMP, Federal Energy Management Program, M&V Guidelines: Measurement and Verification for Performance-Based Contracts Version 4.0 (2015).

Hewamalage, H., Ackermann, K., Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery 37*, 788–832.

Efficiency Valuation Organization (EVO) (2014). IPMVP, International Performance Measurement & Verification Protocol-Core Concepts 2014.

Johra, H., Mans, M., Filonenko, K., De Jaeger, I., Saelens, D., Tvedebrink, T. (2021). Evaluating different metrics for inter-model comparison of urban-scale building energy simulation time series. *Proceedings of Building Simulation 2021: 17th Conference of IBPSA*, 1556-1563.

Johra, H., Schaffer, M., Chaudhary, G., Kazmi, H.S., Le Dréau, J., Petersen S. (2023). Coherent description of 48 metrics to compare, validate and assess accuracy of building energy models and indoor environment simulations. *Aalborg University, DCE Technical Reports No. 314*.

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

Lepot, M., Aubin, J-B., Clemens, F.H.L.R. (2017). Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment. *Water 9*(10), 796.

Panão, M.J.N.O., Santos, C.A.P., Mateus, N.M., da Graça, G.C. (2016). Validation of a lumped RC model for thermal simulation of a double skin natural and mechanical ventilated test cell. *Energy and Buildings 121*, 92-103.

Prema, V., Bhaskar, M.S., Almakhles, D., Gowtham, N., Rao, K.U. (2022). Critical Review of Data, Models and Performance Metrics for Wind and Solar Power Forecast. *IEEE Access 10*, 667-688.