Comparative study of measurement and verification (M&V) baseline models for quantifying energy savings in building renovations

Adalberto Guerra Cabrera and Dimitris Ntimos

IES R&D, Integrated Environmental Solutions Limited, United Kingdom

adalberto.cabrera@iesve.com

Abstract. Measurement and verification (M&V) is the process of quantifying energy savings originated by one or several energy conservation measures (ECM) in an existing building. The estimation of the savings consist of comparing actual energy consumption to the adjusted baseline model. This paper focuses on comparing three approaches for creating baseline models: linear, symbolic regression (SR) and extreme gradient boost (XGBoost); and discusses the advantages and drawbacks on each of them from a practitioner's perspective. In this paper, these approaches are assessed qualitatively and quantitatively. The qualitative assessment compares the type of model output, interpretability and calibration time. Linear model excels in all three criteria whereas the XGBoost is the worst option for model output and interpretability. SR model is the worst performing in terms of calibration time, but intermediate model for output and interpretability. Quantitative assessment is done through the quantification of prediction errors in 367 buildings after being calibrated with hourly data for a 12-month period. The XGBoost model has the highest prediction accuracy in terms of CVRMSE. The linear model performs significantly well in terms of NMBE. SR preforms well in terms of CVRMSE and has the best median NMBE overall remaining as the most accurate interpretable option. The results show different benefits and drawbacks of each approach and the implementation of SR model for this application is the main innovation of this paper.

1. Introduction

Buildings account for more than one third of energy consumption worldwide, being a significant originator of greenhouse emissions [1] . Out of all the energy used in buildings during their life-cycle, it is estimated that 80-90% is used during the operational phase [2]. To reduce energy consumption during the operational phase energy conservation measures (ECM) are cost effective actions that improve energy performance without compromising occupants' comfort.

Measurement and verification (M&V) is the process of quantifying energy savings originated by one or several ECMs in an existing building. M&V is used whenever the savings need to be verified such as in energy efficiency (EE) projects, energy performance contracts (EPC) as well as to provide proof of the effectiveness of an energy management program. The reliable and interpretable calculation of savings is critical to generate consensus of the estimated savings between building owners and energy services companies (ESCO). M&V is essential and increasingly required for the financing and rebate incentives that make energy efficiency projects viable, as investments in these projects can only be justified by delivering real value to beneficiaries in the form of energy savings.

Regarding M&V guidance, standards, protocols and guidelines have been created in order to promote a unified approach to savings estimation and increase confidence from investors. The International Performance Measurement and Verification Protocol (IPMVP) [3] and ASHRAE Guideline 14 [4] are the most common guidelines for M&V projects in the world today. Recently the ISO: International Organization for Standardization (ISO) released an M&V standard under the name ISO 17741:2016[5].

Regardless the standard/protocol/guideline used, strategies to quantify energy savings will depend on the size of the intervention and the magnitude of the expected savings in relation to the energy consumption of the whole building. When expected savings in relation to the whole building are significant (>10% for IPMVP), the process consist of creating a baseline model usually from utility data, and create an "adjusted baseline" meaning that model predictions have been adjusted to current conditions to account for time-dependent parameters such as weather or thermostat set-points.

During the M&V process, the adjusted baseline is used to represent the energy consumption of the building without the ECM for the reporting period. Hence, estimated savings are the difference between the adjusted baseline and the current energy consumption of the whole building. This strategy is known as "option C" for the IPMVP, "whole building path" for ASHRAE G14, and "adjusted calculation" in ISO 17741. Finally, estimated savings are generally used by the ECM investor e.g. an ESCO which receives the monetary savings equivalent for an agreed reporting period as stablished in a contract. After the reporting period is finalized, savings are translated to the building owner/occupants/manager depending on the case.

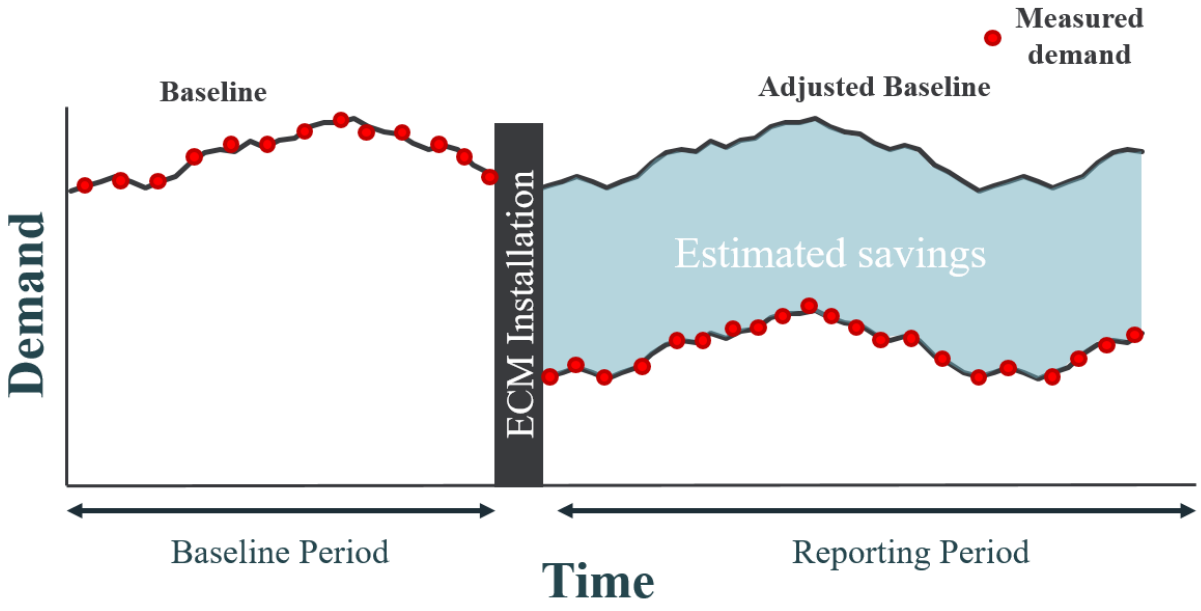

Figure 1 *Estimated savings are calculated as the difference between the adjusted baseline and the actual demand after the installation of the ECM. The period before the ECM installation is known as baseline, the one after is known as reporting period. Measured demand is typically acquired from smart meters.*

In recent times, utility data has become more available in the form of time-series usually with an interval from 15 minutes to 1 hour, as opposed to monthly or bi-monthly utility bills. Creating baseline models using aggregated consumption, e.g. monthly or bi-monthly, is inherently an uncertain process, given that aggregated data does not account for daily demand patterns, weekend/weekday trends, operational changes and other non-routine events within the billing period.

On the other hand, the widespread use of modern building management systems (BMS) and smart meters enables the time-series demand measurements. Consequently, M&V practitioners have a whole new set of options to create baseline models, some of which will be described in the following section.

Current common practices in creating baseline models include day-adjusted models, two/three/four/five parameter change point models, multivariate model, variable-base degree day models, among others [4]. Baseline models face a trade-off between accuracy and costs. Increasing accuracy requires significant sampling, enough metering time-span, modelling efforts and verification of the installation of the ECM. All these actions have cost implications. However, allowable uncertainty in baseline models is limited by the magnitude of the expected savings, otherwise, it becomes impossible to demonstrate that savings are actually due to ECMs over noise. Hence, the minimization of uncertainty by creating higher quality models with a limited budget is one of the main challenges of M&V.

This paper focuses on comparing three approaches for creating baseline models: linear, extreme gradient boost (XGBoost) and symbolic regression (SR) and discussing their advantages and disadvantages from a practitioner's perspective. A practitioner's perspective means that comparison should respond to the typical questions an M&V professional will commonly face on their everyday practice. Overall quality of the models is assessed qualitatively and quantitatively. On the qualitative side, the type of model generated, interpretability and calibration times is discussed. On the other hand, the quantitative assessment is done through an online assessment tool that provides hourly 12-month gap free data for 347 datasets for model calibration. The quantification of the assessment is based on the normalized mean bias error (NMBE) and the Coefficient of Variation of the Root Mean Squared Error (CVRMSE) as well as the 25th and 75th percentile over all the data sets. The code is provided to promote the use of these techniques in everyday M&V practitioner's work.

2. Baseline models

Demand of energy in buildings can be generally attributed to independent variables such as time of the day, outdoor temperature and other weather variables, levels of occupancy and other non-routine changes, such as thermostat set-points.

Baseline models can be used to predict average demand shed, rebound, daily peak demand and/or energy in daily, weekly or monthly basis [6]. Also they can be designed to make short term predictions i.e. using the 10 most recent days [7] or to generate long term predictions (12-months or more) given a similarly long calibration data set [8].

Finally, current M&V guidelines require the model to be reported and agreed among all stakeholders[3].

2.1. Linear model

Linear models are commonly used to create baseline models for M&V. An advantage of this approach is the interpretability of the model as it can be reported in form of a mathematical function across stakeholders for further verification, if required. However, buildings are complicated systems and linear models cannot always capture the underlying relationships between inputs, such as weather, occupancy and baseline adjustments, and an output, such as electricity or gas consumption. An example of a linear model for a baseline model is presented in the following equation. Notice that α is a conversion factor equals to $2 * \pi/(7\ days, 12\ months\ or\ 24\ hours\ depending\ on\ the\ case)$.

$$
\begin{aligned}
&Energy\ demand(t) \\
&\quad = 234.150 - 4.145\big(Days\ in\ month(t)\big) - 3.188\sin(Month\ number(t) * \alpha) \\
&\quad + 4.997\cos(Month\ number(t) * \alpha) + 5.011\sin(Day\ of\ week(t) * \alpha) \\
&\quad - 1.906\cos(Day\ of\ week(t) * \alpha) - 5.795\sin(Hour\ of\ day(t) * \alpha) \\
&\quad - 36.873\cos(Hour\ of\ day(t) * \alpha) - 13.297\big(Bank\ holiday(t)\big) \\
&\quad + 29.460\big(Weekday(t)\big) - 0.0198\big(Diffuse\ horizontal\ radiation(t)\big) \\
&\quad - 0.0029(Direct\ normal\ radiation(t)) - 0.954(Wet\ bulb\ temperature(t)) \\
&\quad - 26.077(Baseline\ adjustment(t)
\end{aligned}
$$

This linear model can be deployed and plotted alongside the electricity metered data, see figure 2. As expected, it has sinusoidal shape during the 24 hours cycle that is modulated (in scale and position) for week/weekend/ holiday day, month and weather variables.

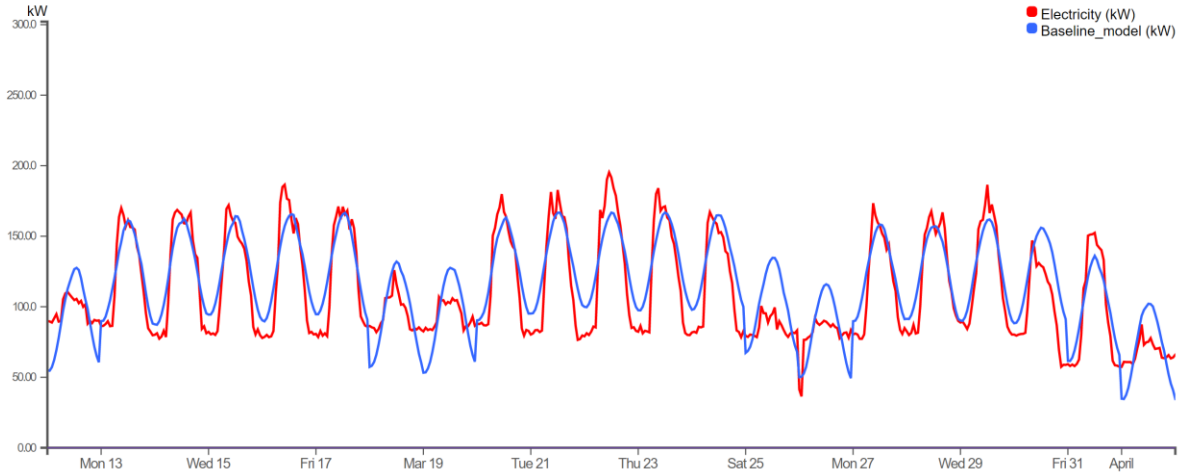

Figure 2 *Baseline linear model example plotted alongside metered data during March 2017. Notice the limitations in terms of flexibility to present hourly variations due to its sinusoidal nature.*

2.2. Symbolic regression model

Symbolic regression (SR) is an application of genetic programming (GP) used to solve a large range of tasks and is considered one of the most successful applications of GP. SR uses genetic algorithm (GA) optimization to create a regression model given a set of inputs and a target value. The pseudo code of SR can be defined as:

```
START
Generate the initial set of randomly created functions
Compute fitness (e.g. mean absolute error)
REPEAT
    Selection of fittest functions
    Crossover (combine bits of fittest functions)
    Mutation (add other random changes)
    Compute fitness
UNTIL population has converged to minimum error or reach generation limit
STOP
```

This approach generates a human-readable model and can potentially generate complex models with similar performance to the ones created using black box machine learning (ML) models, making it a balanced option between the simplicity of linear models and accuracy of ML models, SR is therefore one of the alternatives being explored in the new field called explainable artificial intelligence (XAI)[9]. No literature demonstrating the applicability of SR for baseline models was found and the demonstration of its usefulness is the main innovation presented in this work.

The most relevant parameters are the number of generations (G) and population size (P), which are studied in GA literature[10], and parsimony coefficient. As a rule of thumb, the number of generations varies from ten to fifty [11]. The selection of this parameter has a large impact in the accuracy of the model; higher number of generations tend to increase accuracy of the model, see figure 3. However, a higher number of generations has a time penalty during the creation of the model, which explains why SR models tend to have a larger calibration time than the other types of models presented in this work.

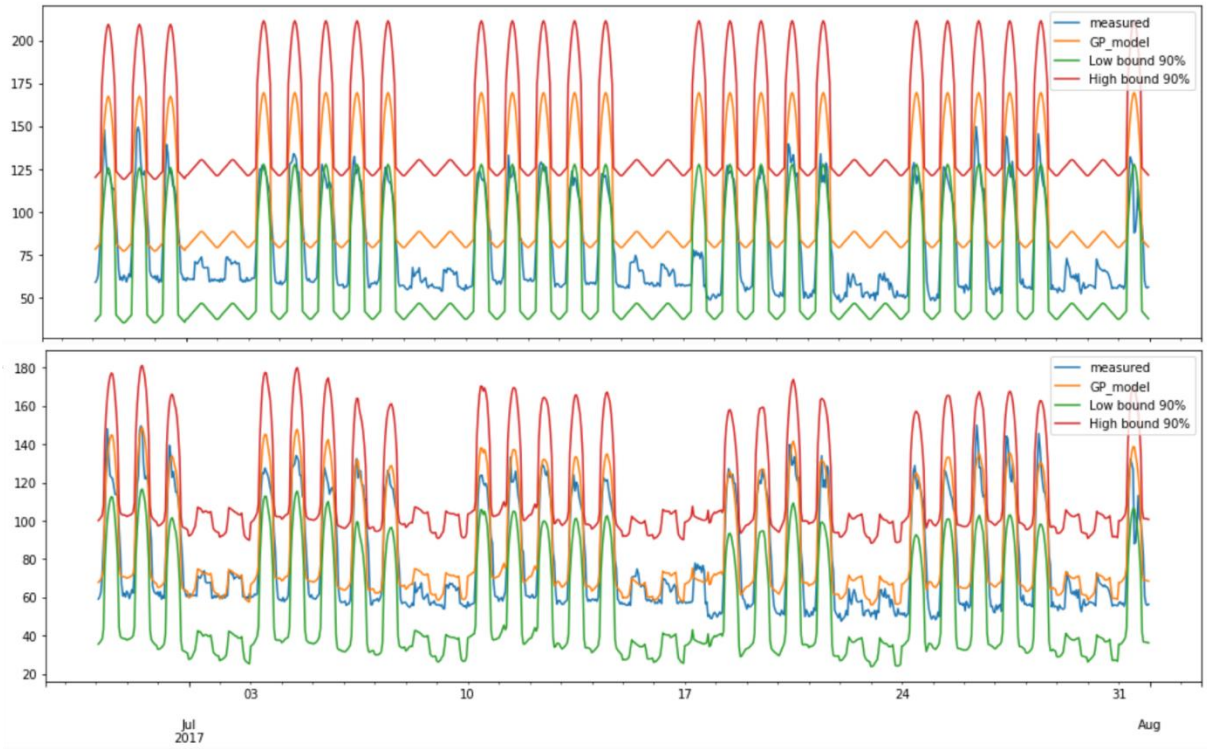

Figure 3 *Example of the impact that generations have on accuracy on the prediction of energy consumption. In the top figure, the model was calibrated with 20 generations. The bottom figure, the model was calibrated with 100 generations, and the output of the model is closer to the measured data as compared to the top figure. However, the second model required 13 times more time to calibrate.*

For the case of population size the situation is similar, it is impossible to suggest an ideal population size as it will vary depending on the application. It was found in literature that this number is generally higher than 500 and that various tests should be carried out in order to determine an adequate trade-off [11].

Increases on the complexity of the functions without significant model improvement is known as bloat. The complexity of the candidate solutions is controlled by a parsimony coefficient and it must be carefully monitored as it is magnitude dependent. This parameter is not as relevant as generations and population size, but it has an impact on the interpretability of the model.

Other challenges related to SR are reproducibility, local adaptation for ongoing calibration, and the requirement of expert knowledge to generate models through SR. Scalability issues are also relevant, as the complexity grows non-linearly as the number of input variables increases.

Once the model is trained, the model produced consists of a function that can be exported to any platform that can carry out functions, i.e. sum, multiplication, subtraction, inverse, square root, etc.; of time series data. Figure 4 shows a calibrated baseline model created using SR.

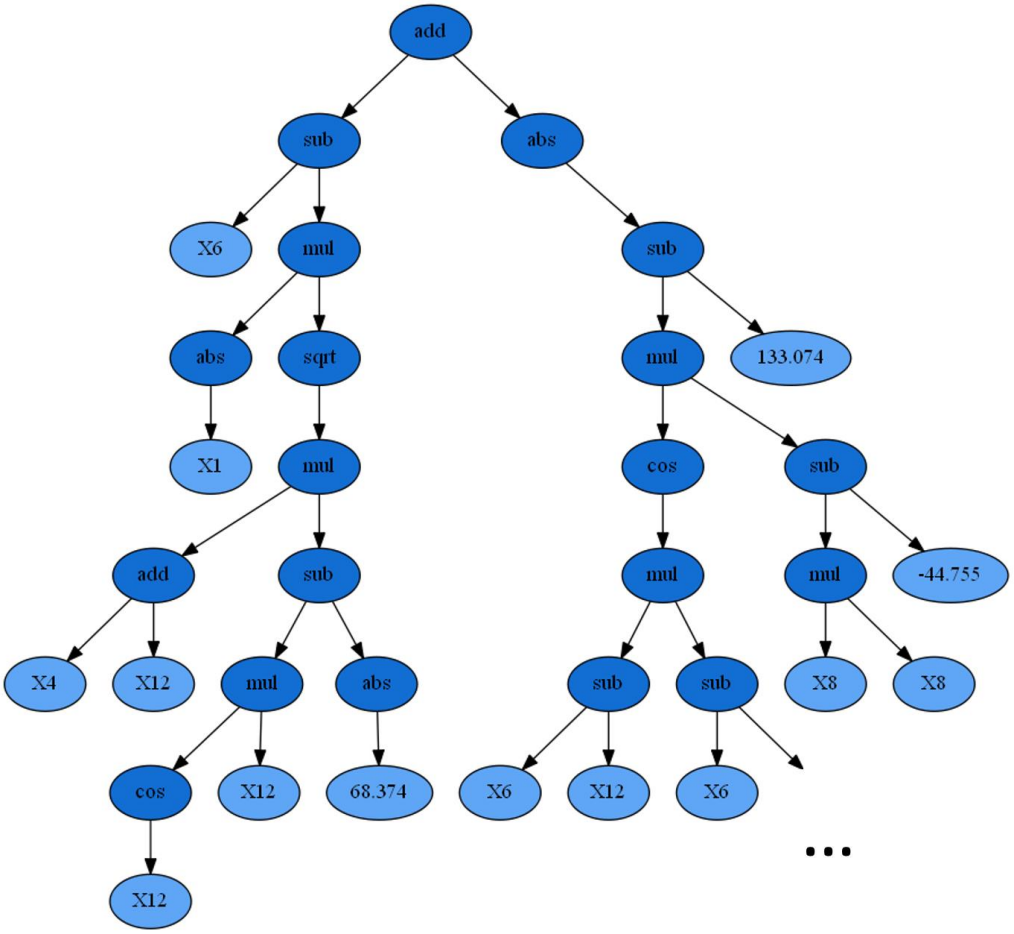

Figure 4 *Graphical representation of a symbolic regression baseline model with an optimal number of generations, population size and parsimony coefficient. X1, X2,X3…XN are model inputs in the forms of time series. The output value is the final numerical result after being processed by each function (sum, sub, mul, abs, cos, sqrt) in a bottom up approach.*

Figure 5 shows the model in figure 4 being deployed in an online platform capable of performing the functions used by the model, i.e. cosine, sine, absolute values, multiplication, sum, etc.

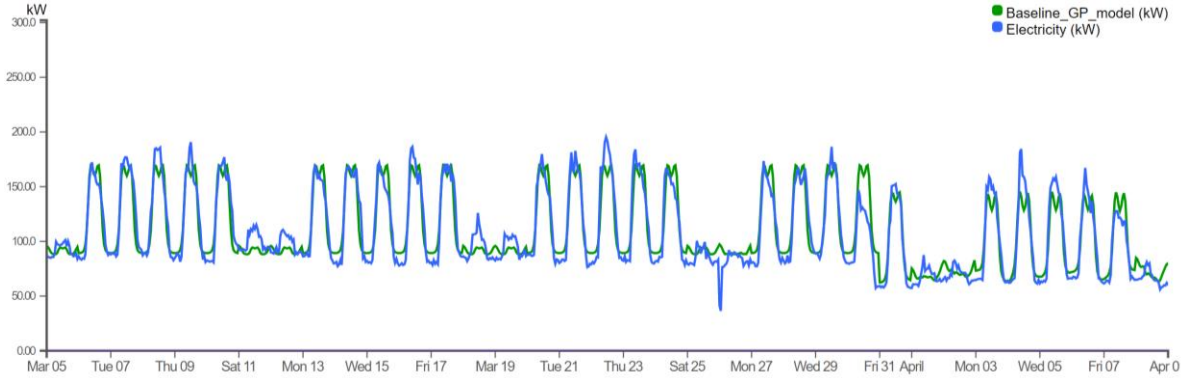

Figure 5 *Example of a baseline model created using a symbolic regression model with an optimal number of generations, population size and parsimony coefficient. In contrast with the linear model, SR model does a better job representing complex patterns existing in the metered electricity values.*

2.3. Extreme gradient boost model (XGBoost)

The rising availability of smart meters and analytical tools, such as machine learning (ML), has proven to be beneficial for baseline M&V models with increased levels of accuracy as compared to traditional

statistical models. However, these new techniques have raised concerns regarding interpretability given that ML models are algorithms hardly understandable to humans, being this a critical drawback.

XGBoost models, a type of ML, provide a feature importance feature shows the most relevant features for the prediction. Due to their partial interpretability, these models have a number of applications in the building sector, from fault detection [12], building demand prediction [13] and energy performance grading and benchmarking [14].

3. Quality assessment

In order to provide a useful description to the M&V practitioner, two types of assessments are presented in this document, qualitative and quantitative. Qualitative covers aspect related to the calibration time, interpretability and the type of output that can be expected for each model. Quantitative assessment focuses on comparing prediction errors for a large set of buildings. Selected parameters for each of the models are presented in the following subsection.

3.1. Selected model parameters

Linear model has the parameters indicated in table 1. Parameters not displayed in the table, suggest that default values were implemented.

Table 1 Linear model parameter values.

Parameter	Value
Library name	sklearn.linear_model.LinearRegression
fit_intercept	True
normalize	False

XGBoost has the parameters indicated in table 2, this model is adapted from the work of [13].

Table 2 XGBoost parameter values.

Parameter	Value
Library name	xgboost
KFold	n_splits=5, shuffle=True
Scoring parameter	mean_squared_error
Number of features for hyper parameter search	20
XGB max depth	20
Learning rate	25 values from 0.001 to 0.1,

Table 2 XGBoost parameter values.

Parameter	Value
Number of estimators	20 values from 100 to1000
Number of iterations	25

The SR model has the parameters indicated in table 3. Notice that a SR model requires more parameters, however, they are not presented in this work as the default values were specified.

Table 3 SR parameter values.

Parameter	Value
Library name	gplearn
function_set	['add', 'sub', 'mul','div', 'sqrt','log','abs','inv','neg','cos','sin','tan']
Range of constants	Max value is the 95% quantile of the dependent variable in the calibration set Min is equals to -1*(Max value)
Population size	12000
Generations	20
Parsimony coefficient	0.001
metric	mean absolute error

3.2. Qualitative assessment

During the M&V process using current guidelines, baseline models should be reported for further verification, if required, hence, the type of model output is important. Traditionally, a baseline model is reported in the form of a function, indicating the input values, units for both input and output and some other uncertainty metrics as defined in [15]. Therefore, the model output is relevant when the M&V plan is adhering to IPMVP.

A good model should also be easily interpreted. Even when a model can be described with mathematical functions, interpretation is not necessarily hassle free. A linear multivariate model of the form:

$$E = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4$$

where:
E = Energy use
β_1 through β_5 = Coefficients
X_1 through X_4 = Independent variables

is certainly easier to interpret than a model created using symbolic regression without an adequate complexity penalty (defined as parsimony coefficient), such as:

$$E = X_7 - X_1X_4 - X_1(X_5X_{16})^{\frac{1}{4}} - \max(X_1, logX_5) + \frac{X_5}{X_7 + X_8 + X_{16} + \sin(X_{14})} + \min(A, B)$$

where:

$$A = \sqrt{\min(0, X_{16})}\left\{X_7 - X_4 + \left(\sqrt{\frac{X_5}{\sqrt{X_4 + X_7} + X_2}}\right) + \frac{X_5}{X_4 + X_7}\right\}$$

$$B = \frac{X_5}{X_{10}}\left(\frac{X_5}{\sqrt{X_9} + X_4} + X_7\right)\sqrt{X_{14} - X_0}$$

which represents a real baseline model with a low interpretability. Yet this model can be considered interpretable if it is compared to another algorithm-based model output, such as the one provided by the XGBoost model. Algorithm-based models are typically binary files interpretable only by their corresponding libraries, e.g. Scikit-learn. The XGBoost model provides the option of displaying feature importance but does not provide a full explanation of how a prediction is made. This is a critical drawback for XGBoost models for situations where interpretability is a critical aspect of the project.

The third and last aspect assessed is the calibration time. There is a direct impact on costs related to the time the M&V practitioner spends fine tuning a model and/or creating a large number of models. A model that is calibrated quickly gives the practitioner time to consider improvements in data pre-processing (e.g. feature engineering), parameter tuning and enables the possibility of on-going re-calibration, which is relevant on M&V 2.0 applications such as automated analysis [16]. In general, linear models have the fastest calibration time, generally under 1 second (fractions of seconds are not documented as it is not relevant information from the practitioner's point of view). XGBoost requires some calibration time during the parameter search using gradient boosting, during this work, XGBoost presented calibration times varying from 5 to 10 minutes. Finally, the SR model, has the largest variation in time. Acceptable results were found using 20 generations and a population size of 12,000, resulting in a training time of approximately 12 minutes. Table 4 shows a summary of the quantitative findings for the three models.

Table 4 Qualitative comparison of the three baseline models.

Model	Output	Interpretability	Calibration time
Linear	A function in form of linear equation.	High	Short (less than 1 second).
Extreme Gradient Boost	A file saved in a compatible format (i.e. *.SAV) with varying size depending on the type of model (i.e. from 500kb to 100 mb)	Generally low, depending on the ML algorithm. Decision trees have some degree of interpretability but it fades away when decision forest are used.	Short-Medium. It took around five minutes (≈300 seconds) for calibration a full year's worth of data.

		Feature importance is useful in many cases.	
Symbolic regression	Function with varied complexity, usually from 100 to 500 terms. But it can be controlled through an adequate parsimony coefficient.	High-medium (depending on number of terms allowed)	Long. It requires a careful parameter selection for generation, population size and parsimony coefficient. In this work calibration time is around 10 to 15 minutes (≈600 to 900 seconds).

In this assessment, the linear model excels in the three criteria whereas the XGBoost model is the worst option for model output and interpretability. The SR model is the worst performing in terms of calibration time, but is an intermediate option between the linear and XGBoost models in terms of the other properties.

3.3. Quantitative assessment

Due to the inherent random component on building loads, the prediction of the demand is always subject to prediction errors. These prediction errors can be quantified using a variety of metrics.

ASHRAE Guideline 14 [4] recommends the use of the Normalized Mean Bias Error (NMBE), which is defined as

$$NMBE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{(n-p) * \bar{y}}$$

NMBE is an indication of the overall bias of the model, in other words, it quantifies the tendency to over or under estimate predictions for a defined period. This metric is independent of time and it can result in overall positive or negative bias cancelling.

Also recommended by ASHRAE is the Coefficient of Variation of the Root Mean Squared Error (CVRMSE), defined as:

$$CVRMSE = \frac{\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(n-p)}}}{\bar{y}}$$

The CVRMSE is a normalized version of the root mean square error (RMSE), which quantifies the accuracy of the prediction, using the standard error, divided by the mean demand.

The coefficient of determination (R^2) is a useful metric to assess how much of the demand is predicted by the model, it is also known as "goodness of fit" or "degree of correlation". R^2 ranges between 0 and 1, with 0 representing no correlation and 1 representing perfect agreement between the model and the metered data. In practice, this metric is used only to perform initial checks in the model and a low R^2 value might indicate the absence of relevant independent variables or that the model has an incorrect functional form.

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{n=1}^{N}(y_i - \mu)^2}$$

Other useful calibration metrics have been suggested, such as the Range-Normalized Root Mean Squared Error (RNRMSE), which eliminates the scale-dependency that the CVRMSE metric presents making it a robust metric to compare model performance when presented alongside the R^2 [17]. The

Unscaled Mean Bounded Relative Absolute Error (UMBRAE) also is useful when assessing time-series forecasting as it is resistance to outliers, symmetric, scale-independent, and interpretable [18].

In this paper, only the NMBE and the CVRMSE metrics will be considered due to their familiarity across the M&V industry. Research has shown that the combination of these two metrics are effective quality indicators[19]. Additionally, the extraordinary large number of buildings used for quantitatively assessing the models is expected to compensate the potential flaws of these two metrics. Initial model check is presented using the R^2 achieved during the calibration phase.

The quantitative assessment is performed by calculating the three modeling approaches with the advanced M&V assessment portal. This portal has been released to the general public in May 2019 with the intention of fulfilling the industry demand for objective testing methods and benchmarking of advanced M&V modelling tools [20].

The datasets that are provided for testing models include buildings of different uses, characteristics and climate zones ensuring that the model can deal with a variety of situations, ensuring (to a certain extent) robustness. The buildings from which the datasets were generated have no known EE projects and data have been cleaned from gross anomalies. This portal provides information from 367 meters.

Model predictive capability is carried out through out-of-bag estimates, meaning that a test dataset is kept away from the calibration data set. A series of 12-month datasets containing time (input), ambient temperature (input) and energy consumption (output) are provided for calibrating the models, then a testing dataset containing only time and ambient temperature is used as the input of the calibrated model to generate predictions which are submitted to the portal for its corresponding assessment. Figure 6 shows conceptual example of this process.

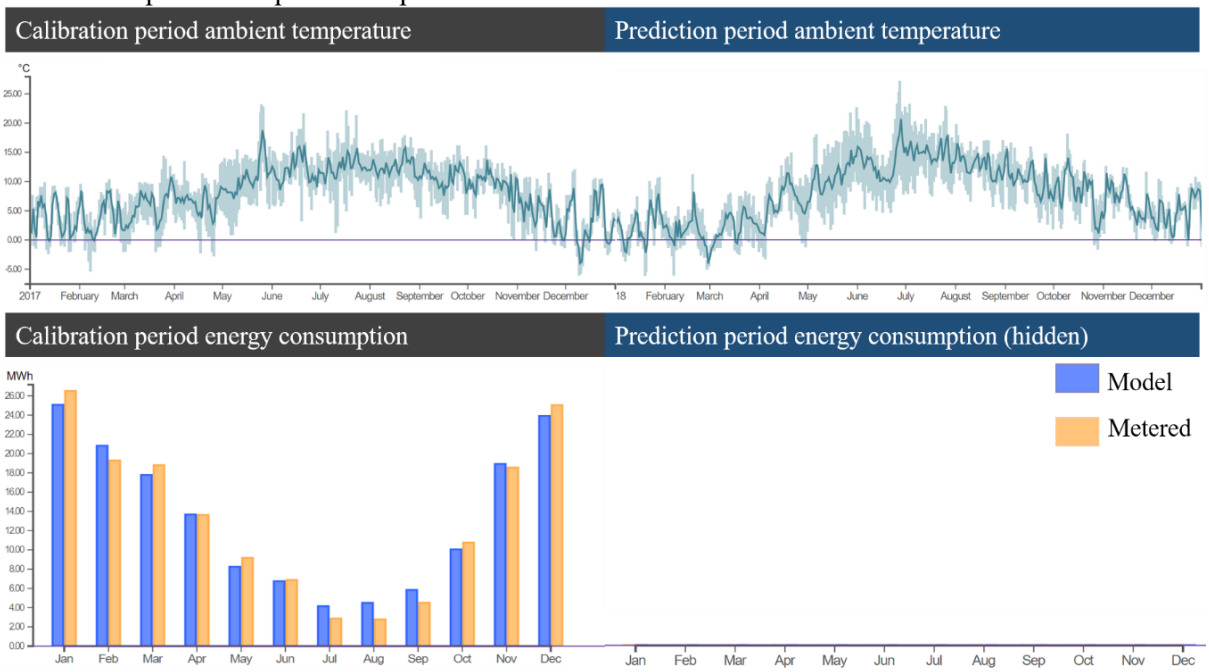

Figure 6 *Out-of-bag estimate example. If left-hand side, metered data and ambient temperature (top) is used to calibrate the models. During the prediction period, right-hand side, ambient temperature is provided and the calibrated model is used to predict the actual metered value. Actual metered values from the prediction period are hidden and only accuracy metrics are reported back.*

Prediction results are provided as the median of scores across the predicted energy consumption in all the 367 meters.

i) Calibration period results

The coefficient of determination, CVRMSE and NMBE are plotted in figures 7, 8 and 9 respectively.

Coefficient of determination is presented only to provide a more complete comparison between models. As discussed before, a low R^2 tends to be indicative of the overall model competence for the prediction.

Notice that these metrics should not be used as an indicative of accuracy on the unseen data but rather as the flexibility of the model to fit data used for calibration. An excess on the flexibility of the model might lead to overfitting, hence it should be controlled for improved predictions.

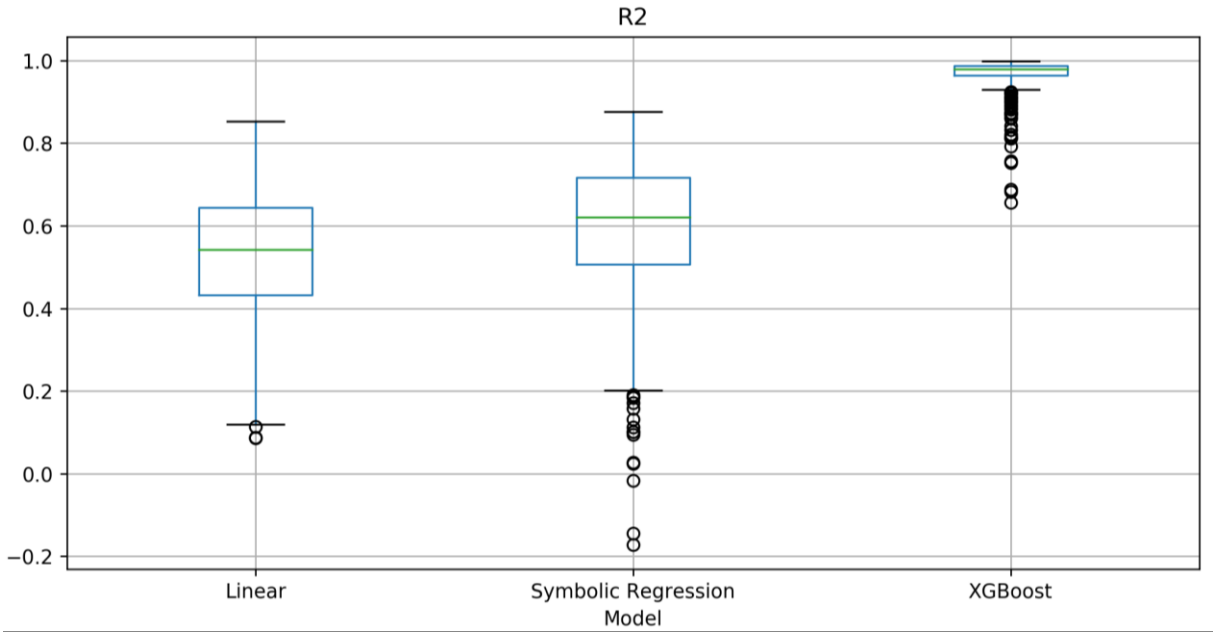

Figure 7 *Coefficient of determination model results for the calibration dataset.*

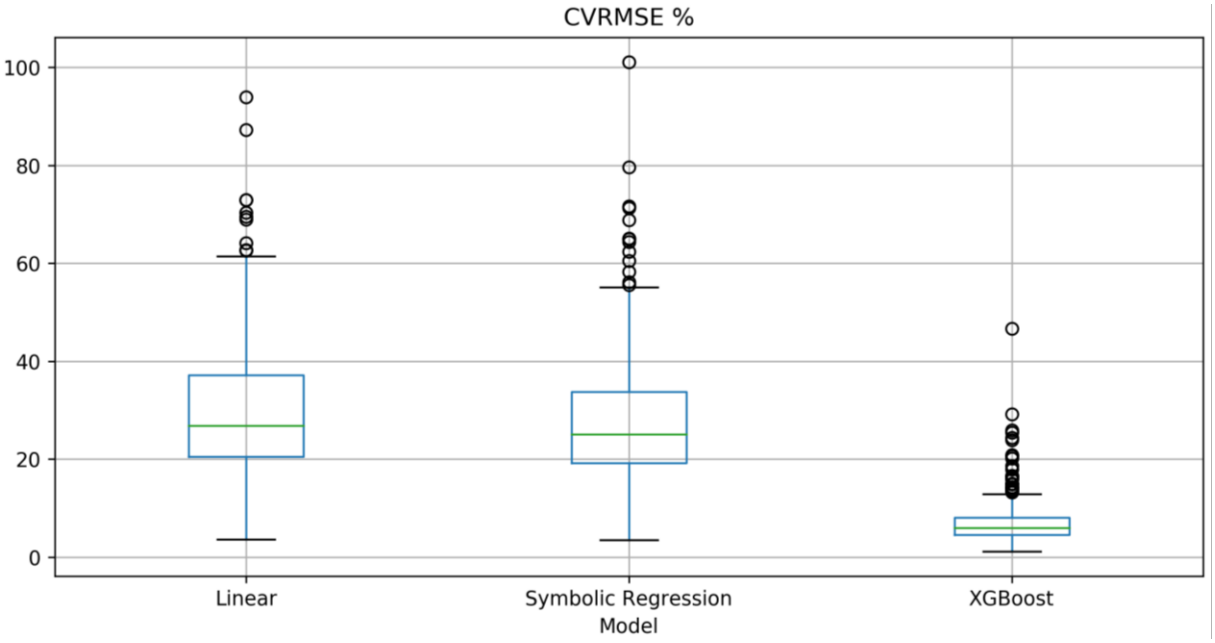

Figure 8 *CVRMSE model results for the calibration dataset.*

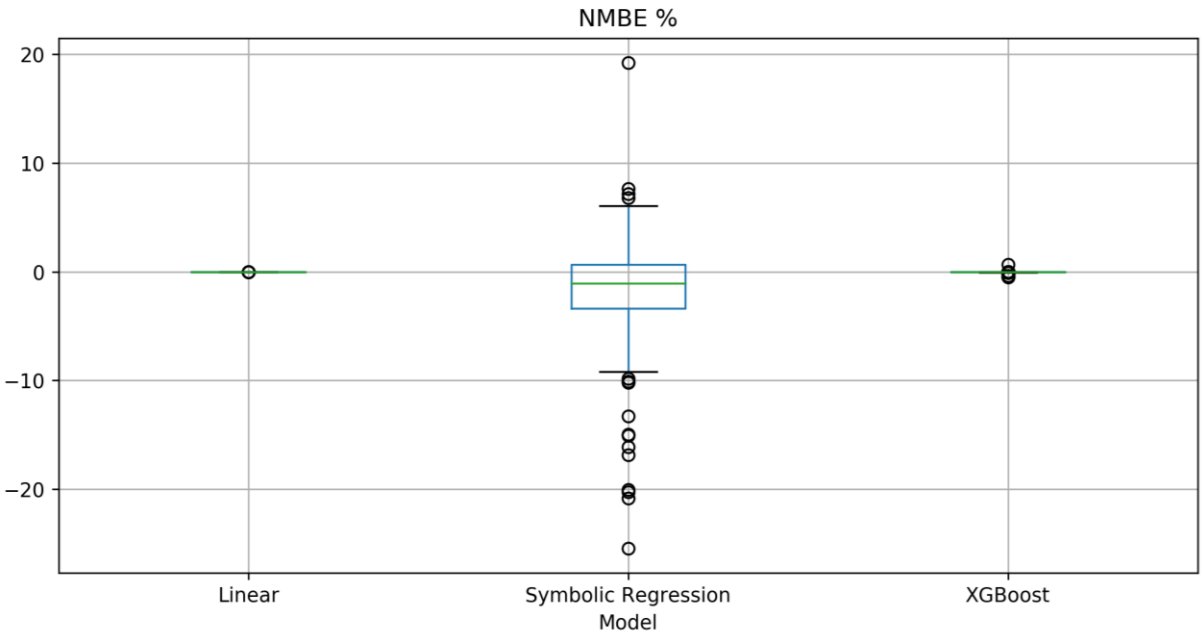

Figure 9 *NMBE model results for the calibration dataset. Notice that due to the nature of the linear model calibration, a NMBE close to 0 is part of the calibration requirements.*

ii) Prediction period results

Prediction results for each of the models are summarized in the following tables. Results consist on the median, 25th and 75th percentile CVRMSE and NMBE for each of the studied model types. Equivalent calibration results are also provided only for guidance. All prediction results have a bias correction using the NMBE obtained in the calibration results. It is worth mentioning that specific prediction results are never provided to avoid M&V tools to be "tweaked" to outperform in this specific test, hence, a more detailed whisker box plot as the one presented in the calibration results section is not possible to generate.

The results are presented in table 5, 6 and 7 for the linear, SR and XGBoost models respectively.

Table 5 Linear model results in the calibration and test dataset. CVRMSE and NMBE percentiles after predicting 367 energy meters.

	CVRMSE (%)		NMBE (%)	
Percentile	Calibration	Test	Calibration	Test
25th	20.44	26.06	0	-10.32
50th (median)	26.87	40.65	0	0.54
75th	37.19	71.78	0	11.03

Table 6 Symbolic regression model results in the calibration and test dataset. CVRMSE and NMBE percentiles after predicting 367 energy meters.

	CVRMSE (%)		NMBE (%)	
Percentile	Calibration	Test	Calibration	Test
25th	19.21	26.03	-3.39	-11.34
50th (median)	25.05	39.67	-1.08	0.46
75th	33.68	73.47	0.68	11.64

Table 7 XGBoost model results in the calibration and test dataset. CVRMSE and NMBE percentiles after predicting 367 energy meters.

	CVRMSE (%)		NMBE (%)	
Percentile	Calibration	Test	Calibration	Test
25th	4.49	23.97	-0.039	-11.35
50th (median)	5.89	37.17	-0.035	-0.52
75th	8.00	72.42	-0.031	9.58

A summary of the test results is presented in figure 10, where the linear (red triangle), SR (green circle) and the XGBoost (blue square) models are displayed in terms of median CVRMSE and NMBE.

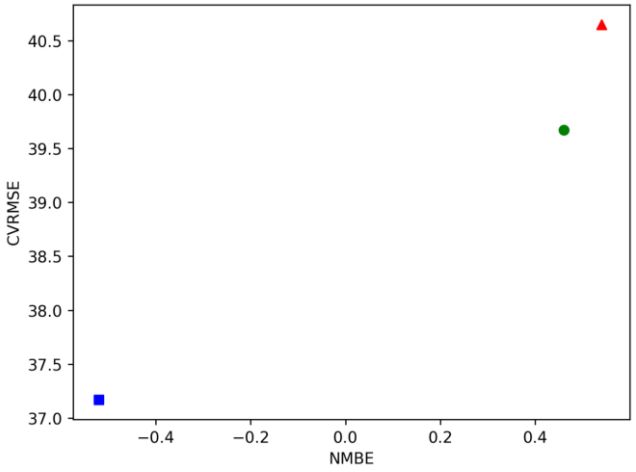

Figure 10 *Test results summary. Linear model (red triangle), SR model (green circle) and the XGBoost (blue square) results are displayed in terms of median CVRMSE and NMBE.*

It can be noticed that linear model performs significantly well in terms of NMBE. This might suggest that for applications where prediction is presented in an aggregated way (i.e. monthly, yearly), linear model is a viable option.

XGBoost model has the highest prediction accuracy in terms of CVRMSE. However, as described in the qualitative section, interpretability has been given up in exchange of improved accuracy.

SR model showed an improved median CVRMSE compared to the linear model option and has the best median NMBE overall, suggesting that when interpretability is required and calibration time is not the main constraint, this model is the best option for time-series predictions.

Prediction results can be likely further improved by implementing change point detection algorithms for all three models [21]. Model specific follow up actions include parameter optimization in the XGBoost model for number of estimators to reduce potential "overfitting", defined as the overrepresentation of the calibration data set reducing prediction accuracy on the unseen testing dataset. Parameter tuning can be also implemented in the SR model for the constant range, population size and parsimony coefficients to reduce its large variability on the calibration results, as shown in figures 7, 8 and 9.

4. Conclusion and future work

Measurement and verification (M&V) is the process of quantifying energy savings originated by one or several ECMs in an existing building. During the M&V process, the adjusted baseline is used to represent the energy consumption of the building without the ECM for the reporting period. Hence, estimated savings are the difference between the adjusted baseline and the current energy consumption of the whole building. In this work, Linear, Symbolic Regression (SR) and XGBoost models were compared quantitatively and qualitatively from a practitioner's approach. The use of the SR models for M&V is the main innovation presented in this work.

Qualitative comparison included model output, interpretability and calibration time. In this assessment, linear model excels in the three criteria whereas the XGBoost is the worst option for model output and interpretability. The SR model is the worst performing in terms of calibration time, but is an intermediate between the linear and XGBoost models in terms of the other properties.

Regarding the quantitative assessment, the XGBoost model has the highest prediction terms of CVRMSE. Linear model performs significantly well in terms of NMBE. This might suggest that in applications where prediction is presented in an aggregated manner (i.e. monthly, yearly) linear model is a viable option. SR model showed an improved median CVRMSE compared to the linear model option and has the best median NMBE overall, suggesting that when interpretability is required and calibration time is not the main constraint, this model is the best option for time series predictions .

Prediction results can be likely further improved by implementing change point detection algorithms for all the three models [21]. Model specific follow up actions include hyper-parameter optimization in the XGBoost model for number of estimators and for the SR model for the constant range, population size and parsimony coefficients.

5. Acknowledgements

This project is funded by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement N. 818329.

Special thanks for Denis Tanguay, Executive Director of the Efficiency Valuation Organization (EVO) for facilitating access to the advanced M&V tools testing portal http://mvportal.evo-world.org/

Additional thanks to Shane Campbell, Senior Project Consultant and Mario Favalli Ragusini, iCL project consultant at IES for their technical advice.

6. References

[1] IEA, "Global Energy and CO2 Status Report 2017," 2018.

[2] T. Ramesh, R. Prakash, and K. K. Shukla, "Life cycle energy analysis of buildings: An overview," *Energy and Buildings*. 2010.

[3] IPMVP, "International Performance Measurement & Verification Protocol: Concepts and Options for Determining energy and Water Savings, Volume 1," 2002.

[4] ANSI/ASHRAE, "ASHRAE Guideline 14-2015 Measurement of Energy and Demand Savings," 2015.

[5] ISO, "ISO 17741 - General technical rules for measurement, calculation and verification of energy savings of projects," vol. 0, no. 0, 2016.

[6] J. L. Mathieu, D. S. Callaway, and S. Kiliccote, "Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices," *Energy Build.*, 2011.

[7] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, "Statistical analysis of baseline load models for non-residential buildings," *Energy Build.*, 2009.

[8] J. Grandersona, S. Touzani, S. Fernandes, and C. Taylor, "Application of automated measurement and verification to utility energy efficiency program data," *Energy Build.*, 2017.

[9] A. Preece, "Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges," *Intell. Syst. Accounting, Financ. Manag.*, 2018.

[10] M. Mitchell, "An Introduction to Genetic Algorithms (Complex Adaptive Systems)," *MIT Press*, 1998.

[11] R. Poli, W. B. Langdon, N. F. McPhee, and J. R. Koza, *Field Guide to Genetic Programing*, no. March. 2008.

[12] D. Chakraborty and H. Elzarka, "Early detection of faults in HVAC systems using an XGBoost model with a dynamic threshold," *Energy Build.*, 2019.

[13] D. Chakraborty and H. Elzarka, "Advanced machine learning techniques for building performance simulation: a comparative analysis," *J. Build. Perform. Simul.*, 2019.

[14] S. Papadopoulos and C. E. Kontokosta, "Grading buildings on energy performance using city benchmarking data," *Appl. Energy*, 2019.

[15] EVO; Efficiency Valuation Organization, "Uncertainty Assessment for IPMVP," *Evo*, vol. 1, no. April, 2018.

[16] E. Franconi *et al.*, "The Status and Promise of Advanced M&V: An Overview of 'M&V 2.0' Methods, Tools, and Applications," *Rocky Mt. Inst.*, no. March, pp. 0–22, 2017.

[17] D. Chakraborty and H. Elzarka, "Performance testing of energy models: are we using the right statistical metrics?," *J. Build. Perform. Simul.*, 2018.

[18] C. Chen, J. Twycross, and J. M. Garibaldi, "A new accuracy measure based on bounded relative error for time series forecasting," *PLoS One*, 2017.

[19] J. Granderson, S. Touzani, C. Custodio, M. D. Sohn, D. Jump, and S. Fernandes, "Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings," *Appl. Energy*, vol. 173, 2016.

[20] EVO; Efficiency Valuation Organization, "EVO's Advanced M&V Testing Portal," 2019. [Online]. Available: http://mvportal.evo-world.org/.

[21] S. Touzani, B. Ravache, E. Crowe, and J. Granderson, "Statistical change detection of building energy consumption: Applications to savings estimation," *Energy Build.*, vol. 185, pp. 123–136, 2019.