Factors contributing to measurement uncertainty of HVAC-enabled demand flexibility in grid-interactive commercial buildings

Vindel E.[1], Akinci B.[1], Bergés M.[1], Kavvada O.[2], Gavan V.[2]
[1] Carnegie Mellon University, USA, [2] ENGIE Lab CRIGEN, France
evindel@cmu.edu

Abstract. The importance of evaluating the performance of electric power demand flexibility in buildings has increased in recent years due to the expected participation of demand resources in reliability-based grid services. All demand flexibility assessments involve a degree of uncertainty when measuring the magnitude of demand response events, due to the uncertainty in the counterfactual load estimation. The objective of this work is to understand how distinct factors contribute to uncertainty in measuring HVAC-enabled demand response in buildings. More specifically, we will vary two factors that are known to contribute to the uncertainty of the observed event: choice of baseline model and choice of assessment boundary. The practical implication of our analysis is to contribute to the design of improved measurement and verification protocols for demand flexibility in buildings.

1. Introduction

1.1 Overview

The value of electric power demand flexibility as a quantifiable and reliable grid resource is increasing with a modernizing electric grid. This is evident by the active participation of demand resources in energy and ancillary services markets within different regulatory bodies (FERC, 2021). Buildings are the largest part of these demand resources and the deployment of grid-interactive technologies in this sector is anticipated (Neukomm, Nubbe and Fares, 2019). A fundamental technology requirement for this transition is the development of improved performance assessments and trustworthy metrics that quantify their demand flexibility (Satchwell *et al.*, 2021). Current measurement and verification (M&V) protocols were designed for an electric grid that benefited from but did not depend on demand flexibility. Performance assessments are critical for reliable system support, transparent financial settlements and, overall increasing the trustworthiness of demand resources (Goldberg and Agnew, 2013). The challenge in designing performance assessment protocols is that it involves the estimation of the unaltered baseline load, and this step inevitably introduces error between the real load modification and the measured load modification. Moreover, the effects of these errors are event-specific due to the varying accuracy of baseline models. The main approach to reduce the uncertainty, when measuring the real load modification, is to improve prediction accuracy through the choice of baseline model. Another known approach involves the choice of assessment boundary (i.e., level at which the demand flexibility is documented) which can improve the accuracy when measuring an event. These choices play a significant role in the design of M&V protocols and are the source of investigation for this paper. The objective of this work is to understand how the choice of baseline model and assessment boundary contribute to the measurement uncertainty of demand flexibility events. We will narrow our focus to heating ventilation and air conditioning (HVAC) flexibility in commercial buildings as it accounts for a substantial portion of flexible electricity consumption in the building sector (Roth and Reyna, 2019). Furthermore, we will extend the analysis to different climate zones (ASHRAE, 2019), given that HVAC demand flexibility heavily depends on this factor. The understanding of how these factors contribute to measurement uncertainty of demand flexibility

https://doi.org/10.7146/aul.455.c214

events can provide insight into the design of improved M&V protocols tailored to different requirements and applications.

1.2 Existing Research

Understanding the sources of measurement uncertainty for demand flexibility (DF) events has been an important topic in the design of M&V protocols (KEMA-XENERGY, 2003). As this problem is inherent to the value of demand flexibility, studies on comparing the accuracy of baseline models are common in settings beyond academic research (KEMA, 2011; Nexant, 2017). Various demand flexibility baseline models have been proposed, but they all generally fall into the following categories: averaging, simple regression, and machine learning (Amasyali and El-Gohary, 2018). Of these, averaging models cover the majority of industry practice (IRC (ISO/RTO Council), 2018). Simple regression models, such as the time-of-the-week and temperature model, have been proposed with an improved accuracy over averaging models (Taylor and Mathieu, 2015). Additionally, machine learning models, in recent years, joined the conversation due to their increased prediction accuracy (Zhang *et al.*, 2021).

Although literature for more accurate baseline models is ample, studies of the effect of baseline prediction on DF measurements is scarcer. Some works have investigated the uncertainty introduced by the source and frequency of the input data used on the model prediction (Coughlin *et al.*, 2009; Granderson and Price, 2014). Similarly, a more recent work explored the bias of common baseline models when evaluating peak electricity load reduction (Granderson *et al.*, 2021). One of the strengths of these studies is the use of real building experimental results as opposed to simulated. However, the main drawback is that the obtained results are not contextualized by the magnitude of potential DF events. A study that does have this context, evaluated the variability of DF measurements when using a specific time-of-the-week and outdoor air temperature baseline model (Mathieu, Callaway and Kiliccote, 2011). One primary finding is that DF measurement error is primarily driven by baseline model error. However, validation of the source and magnitude of this error was complicated due to the comparison being done on only 95 observations throughout multiple buildings.

The other factor explored in this study is the choice of assessment boundary on which to measure the DF event. The assessment boundary refers to the measurement level at which the DF event is documented (Schiller, Schwartz and Murphy, 2020). Although there are a few options for the assessment boundary, in this work we will focus on the HVAC and total building electricity consumption. The main difference between the two is that the building energy consumption is the combination of HVAC load and all other loads. The intuition is that, in some circumstances, divorcing controllable and non-controllable loads (i.e., sub-metering) has the potential to improve the accuracy of the measurement (Cappers *et al.*, 2013). The mechanisms through which this could be true is two-fold. First, baseline models can have a more accurate fit at the HVAC boundary than total building because the explanatory variables commonly used have a more direct relation to power. Second, load modifications have a larger relative impact at more granular assessment boundaries. Several works have studied the accuracy trade-offs of submetering to obtain more accurate DF assessments (Ji *et al.*, 2016; Lei, Mathieu and Jain, 2021). However, these studies start with the assumption that the HVAC boundary will be more accurate and do not explore the trade-offs when compared to using total building power. It is also important to note, that there are studies that have found that the choice of assessment boundary is inconsequential (Motegi *et al.*, 2004).

In this work, we cover the influence of these two factors by generating an extensive simulation dataset on which we empirically analyze their effects on measured DF events. There are four primary distinctions from prior works. First, we expand the comparison to five different

baseline models including averaging, simple regression, and machine learning. Second, we document the accuracy tradeoffs of the total building and HVAC assessment boundary. Third, we apply various performance metrics relative to the magnitude of the simulated events. Fourth, we analyze these factors for five different ASHRAE climate zones (ASHRAE, 2019). The expected outcome of these results is to inform the design of future M&V protocols.

2. Proposed Approach

To evaluate how the selected factors affect measurement uncertainty of demand flexibility we will first generate a dataset of simulated events. We will leverage this dataset to empirically analyze the contribution of each factor to measurement uncertainty. For clarity in comparison, we will group each building model instance as the following triple `(Climate, Assessment Boundary, Baseline Model)` and our dataset will consist of 50 such triples (5 climates × 2 boundaries × 5 baseline models). A visual abstract of the factors explored is shown in Figure 1. Section 2.1 describes the data generation process including a description of the building model used and the procedure to create the demand flexibility events. Section 2.2 describes the selected baseline models, description of the input features used, and the training/testing methodology model fitting. Section 2.3 summarizes the evaluation metrics used to compare the results from the triple instantiations. Finally, Section 2.4 condenses the key steps of the experimental plan to generate the data and calculate the proposed evaluation metrics.

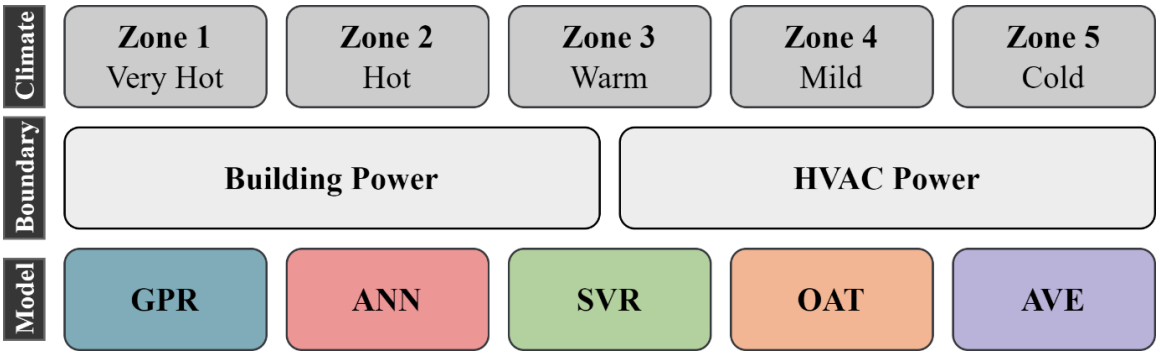

Figure 1: Factors explored - Climate Zone, Baseline Model, and Assessment Boundary (GPR – Gaussian Process Regression, ANN – Artificial Neural Network, SVR – Support Vector Regression, OAT- Time-of-week and outdoor temperature model, AVE – Mid4of6 Averaging Model)

2.1 Model Description

The selected testbed model to generate the dataset is the office building from the *Modelica Buildings Library* (Wetter *et al.*, 2014) equipped with a variable air volume (VAV) reheat system with an implementation of the control sequence *VAV 2A2-21232*. It consists of a five-zone layout shaped after one floor of the DOE Medium Office standard prototype building (Goel *et al.*, 2014). Because this model only simulates HVAC load, we simulate all other non-controllable loads (e.g., plug loads, lighting loads, etc.) as proportional to building occupancy and scaled accordingly. The occupancy was obtained from a separate agent-based stochastic occupancy simulator for office buildings (Chen, Hong and Luo, 2018). The same stochastic occupancy data is also used to model the internal thermal load that the HVAC simulation uses to balance thermal loads in the building. The simulation timestep is selected as 15 minutes given that it is a common resolution for the evaluation of demand resources (IRC (ISO/RTO Council), 2018). For demand flexibility we generate *load shedding events* through a +2°F global thermostat reset for 1-hour duration for every hour during the occupied hours (Vindel *et al.*,

2021). The evaluation period is 90 summer days, not including weekends or holidays. The selected period ensures the HVAC system is operating in cooling mode and the power consumption is engaged by both the fans and chiller.

2.2 Baseline Models

We select five candidate baseline models commonly used for building energy forecasting for demand flexibility M&V. Averaging-based models are commonly used in practice of which we select the Mid4of6 model (AVE) for this evaluation (KEMA, 2011). This model averages the load of the six most recent valid days, excluding the days with the highest and lowest energy consumption. The rest of the models evaluated are regression-based models. The simplest of these models is a piecewise linear model (OAT) that uses a time-of-the-week indicator and outdoor air temperature to predict electric load (Mathieu, Callaway and Kiliccote, 2011). The three remaining models are supervised machine learning models: artificial neural network (ANN, feed-forward, 4 hidden layers, 10-neurons per layer, sigmoid activation), support vector regression (SVR, radial basis function kernel, ε =0.1), and Gaussian process regression (GPR, squared-exponential kernel). The input features for all machine learning models are the same: outdoor air temperature, direct normal irradiance, global horizontal irradiance, relative humidity, time-of-day indicator. The test/train split for all regression-based models was obtained by generating two random datasets from geographically close weather files within the same climate zones as shown in Table 1 (ASHRAE, 2019). Note that a comparison between these machine learning models is difficult because of the variety of architectures and initialization parameters. However, for this work, our aim is not to optimize these model design decisions but to perform an initial comparison of the outputs and strengths of each model.

Table 1: Test/Train Split Weather Data for Regression-based Models

Climate Zone	1, Very Hot	2, Hot	3, Warm	4, Mild	5, Cold
Train Data	Miami, FL	Houston, TX	Atlanta, GA	Baltimore, MD	Chicago, IL
Test Data	Kendall, FL	San Antonio, TX	Athens, GA	Arlington, VA	Aurora, IL

2.3 Evaluation Metrics

The metrics we want to calculate are related to the difference between the observed shed and the real shed. Therefore, in this context, we consider the real shed, the measurand and the observed shed as the measurement. For smart grid applications, two metrics used are the median absolute percentage error (MdAPE) and the normalized mean bias error (NMBE). The first one measures the relative error of the observed measurement, accounting for the variability in the magnitude of the real shed events. We select the median, over the more common mean of the absolute percentage error, to filter out the effect of outlier events with very low magnitude sheds. The metric NMBE, on the other hand, indicates whether a particular baseline model instance generates a biased measurement of the real shed event. Positive values for NMBE indicate that the observed shed is lower than the real shed, and the opposite for negative values. Additionally, we calculate a metric called the reliability threshold estimate (REL). This reliability-inspired metric roughly estimates the frequency of a model falling within a specified threshold (Aman, Simmhan and Prasanna, 2015). For distributed energy resources participating in reliability-based grid services (e.g., contingency reserves) a commonly used threshold is $\varepsilon = 10\%$ (Aman, Simmhan and Prasanna, 2015). A negative REL value indicates that most events were higher than the threshold, and value of -1 indicates that all events are beyond the threshold. The opposite holds for positive values of REL. To evaluate the performance of the baseline

models, independent of the simulated DF events, we will calculate the CVRMSE. Note that this metric is relative to the load (P_i) and not to the magnitude of individual events. For reference, values for CVRMSE less than 25% are considered an acceptable fit for the measurement of energy and demand savings (ASHRAE, 2014).

Table 2: Selected Evaluation Metrics

	Metric	Target	Equation
(1)	Median Absolute Percentage Error (%)	Event	$\text{MdAPE} = \text{median}\left(\left\|\frac{y_i - \hat{y}_i}{y_i}\right\| \text{ for } 1, \dots, n\right) \times 100\%$
(2)	Normalized Mean Bias Error (%)	Event	$\text{NMBE} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i} \times 100\%$
(3)	Reliability Threshold Estimate (-)	Event	$\text{REL} = \frac{1}{n}\sum_{i=1}^{n} \text{C}(y_i, \hat{y}_i)$ $\text{C}(y_i, \hat{y}_i, \varepsilon) = \begin{cases} 1, & \text{if } \lvert y_i - \hat{y}_i\rvert / y_i < \varepsilon \\ 0, & \text{if } \lvert y_i - \hat{y}_i\rvert / y_i = \varepsilon \\ -1, & \text{if } \lvert y_i - \hat{y}_i\rvert / y_i > \varepsilon \end{cases}$
(4)	Coefficient of Variance RMSE (%)	Load	$\text{CVRMSE} = \frac{1}{\bar{P}}\sqrt{\frac{1}{n}\sum_{i}^{n}\left(P_i - \hat{P}_i\right)^2} \times 100\%$

2.4 Experiment Plan

A summary of the designed experimental plan and evaluation is presented in the steps below, and applied to all the selected climate zones:

1. Simulate the building operation generating data for both nominal and load shedding events. Repeat this step for both training and testing weather conditions.
2. Record electricity consumption for HVAC system and building total. Record all input features: outdoor air temperature, direct normal irradiance, global horizontal irradiance, relative humidity, time-of-day indicator.
3. Using the nominal power consumption for the training weather, for both assessment boundaries, fit regression-based baseline models with their respective input features.
4. Calculate the CVRMSE metric (Equation (4)) using the predicted baseline load for each model on both the test and training sets.
5. Compute the measured sheds and real sheds by subtracting the average hourly difference between the true baseline consumption and observed load, as well as the predicted baseline consumption and observed load.
6. Calculate all other evaluation metrics (Equations (1)- (3)) for all combinations of baseline model and assessment boundary.

3. Results and Discussions

The discussions are divided into two sections. Section 3.1 reports and analyses the baseline model results of all the triple instances. Section 3.2 interprets the results relative to the modelled demand flexibility events. The source code for the results presented and the dataset generated is hosted in a public repo in GitHub (https://github.com/INFERLab/DFUncertain-EGICE22/).

3.1 Baseline Model Performance

The performance of all baseline model instances is summarized in Figure 2. The results shown in this section are not connected to the modeled demand flexibility events and are only a reflection of the baseline model fit using Equation (4). For all modeled instances, HVAC consumption was less than 50% of total building power consumption (44%,44%,42%,40% and 37% for climate zones 1-5 respectively). The average Coefficient of Variation of the Root Mean Square Error (CVRMSE) for the building assessment boundary, for all models, is 6.95% and 8.90% for training and testing respectively. For the HVAC boundary the average CVRMSE for training and testing are 8.22% and 11.50%. Additionally, the maximum CVRMSE value for any triple instance is 24.47%. As mentioned earlier, some guidelines suggests that CVRMSE values below 25% are considered a good model fit (ASHRAE, 2014). Hence, by this guideline, all models performed below this minimum performance threshold. In fact, on the test data, 90% of triples had CVRMSE < 15%. Another observation is that using the CVRMSE metric, the HVAC assessment boundary performed slightly worse than the building boundary. This is primarily because the metric is associated with the magnitude of the total electricity consumption. Hence the same magnitude error has different CVRMSE values at different assessment boundaries.

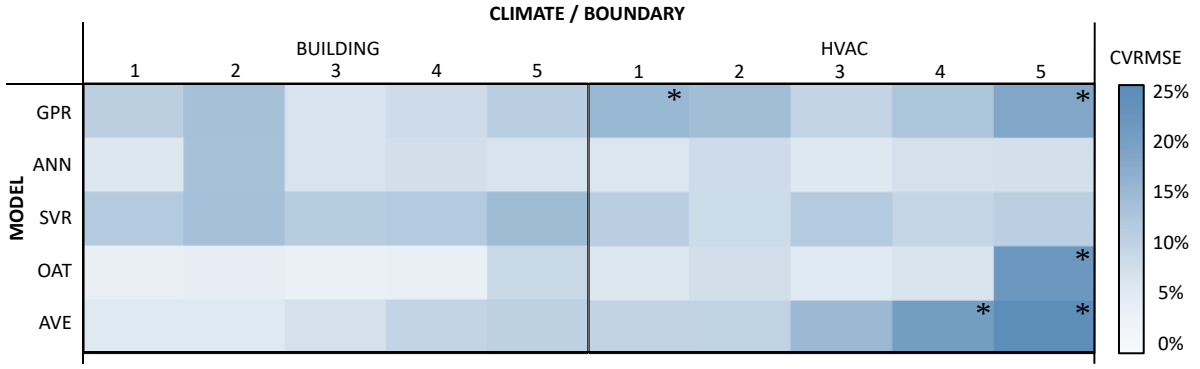

Figure 2 Baseline Model Performance Summary on Test Dataset (CVRMSE > 15% marked with *)

The performance of all models did not appear to consistently vary by climate zones. One notable exception to this finding is the AVE model at the HVAC boundary, and less so at the building boundary. The AVE model performs progressively worse when going from a hotter to a colder climate. The reason for this is that climate zones defined as “cold” have significantly more variability in daily environmental conditions than other climates that are consistently hot. Because of the inductive bias of AVE-type models, intra-day variability is expected to hinder performance. A similar decrease in performance was not observed for all other models. This is an interesting finding, particularly for the HVAC boundary, given that the power consumption is directly affected by weather conditions. For building power, the relations between environmental input features and power are less direct. The OAT model has the best performance for all climate zones at the building boundary. One explanation for this could be that this model captures the schedule-nature of all non-controllable loads by generating an independent model for each timestep. Although our dataset has a stochastic input for the non-controllable loads, these are derived from a scheduled input. This limitation is not particular to our model, but a general drawback of most building energy models. The ANN model was the best performing model using HVAC power consumption. One explanation for this is that the ANN model is able to capture the non-linear behavior of HVAC consumption better than the OAT linear model. Although the GPR model has a lower performance than the ANN and OAT model, one of the advantages of this model is that it outputs uncertainty measures over

predictions. The trade-off in performance for certain applications could be compensated by the probabilistic output of the model.

3.2 Event-specific Performance

As described earlier, we model events by generating synthetic load shed events through a 1-hour global thermostat reset. For each event, we can calculate a real shed (y_i) by subtracting the average real baseline load and the DF event load for the 1-hour period. It is essential to emphasize that the real shed value is independent of assessment boundary. When the baseline model outputs a prediction for the baseline load, at either boundary, we can calculate a measured load shed ($\hat{y}_i$) by subtracting the average predicted load and the DF event load. Finally, for each event, we calculate the relative error ($((y_i - \hat{y}_i)/\, y_i) \times 100\%$). For example, let us assume that a DF event is calculated to have a 20kW average real shed magnitude. If the baseline model underpredicts the baseline load by 4kW, meaning that the measured shed is 16kW, the relative event error is +20%. A boxplot distribution of the relative event error for all triples is shown for the testing data in Figure 3.

The choice of figure as a boxplot distribution is to show the variability in relative event error. While most relative errors are close to 0% the variance of the distribution provides a visual of the expected measurement uncertainty. Figure *3* has limits for relative event error capped at +200% and -200%. Although there are events with relative errors beyond these thresholds (only 2.25% of events), they were considered outliers when calculating the interquartile range of the boxplot distribution. These events are outside the range because they were low magnitude events and not because of significant baseline model error. Not surprisingly, the model performance discussed in Section 3.1 is a good predictor of the performance when measuring DF events. However, the baseline model performance measured by Equation (4) is not able to capture the large gap in event-specific performance driven by the choice of assessment boundary. For all models, measuring the events using an HVAC power baseline model resulted in lower error on average. More specifically, for the test data this constituted in a reduction in MdAPE from 34% to 21%, on average. Regardless of the assessment boundary, in reference to these experiments, the MdAPE values can be considered high for certain grid services. Similar conclusions have been obtained regarding the expected measurement error for demand resources (Mathieu, Callaway and Kiliccote, 2011). Nevertheless, the true magnitude of the expected relative error should be subject to a more comprehensive study than the one presented.

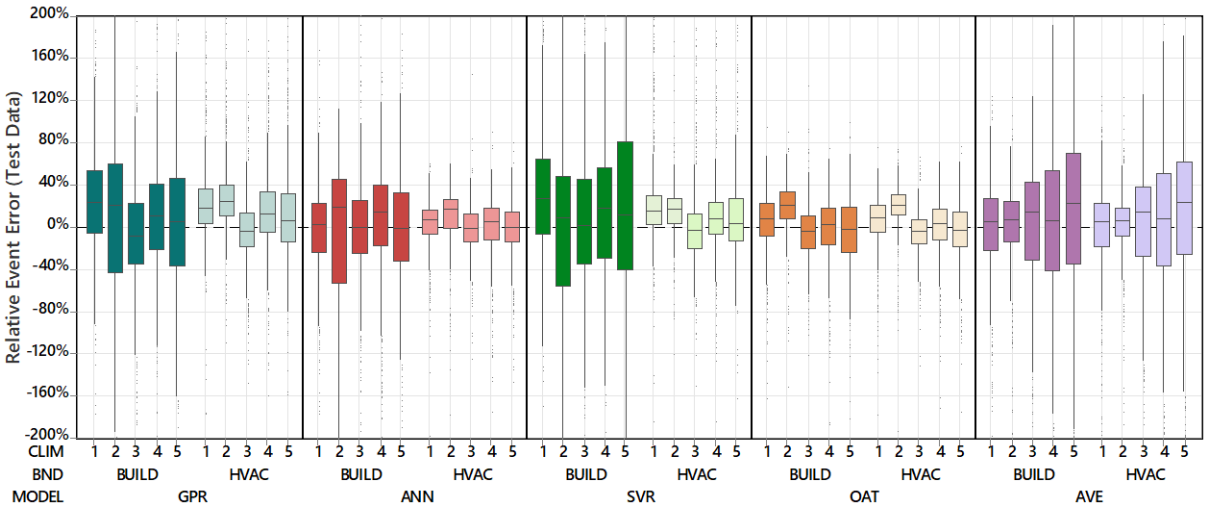

Figure 3 Boxplot distribution of relative event errors for all triple instances for Test Dataset

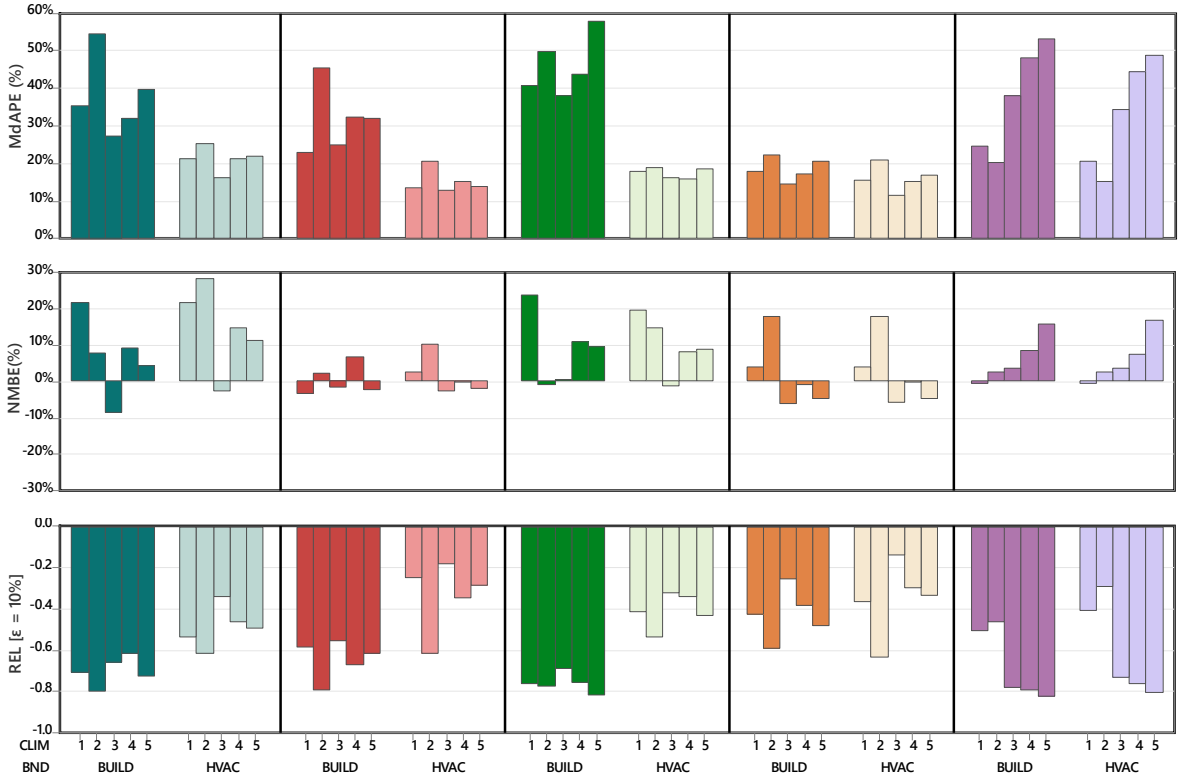

Figure 4 Event-specific metrics for Test Dataset

Figure 4 summarizes the event-specific metrics for the test dataset. As mentioned before, the MdAPE pane visually recapitulates the improvement in measurement accuracy when using the HVAC assessment boundary. Overall measurement bias is an important factor given that under/over-estimation of DF events have different implications for a grid service. In our experiments, most models exhibit low bias (i.e., NMBE close to 0%). However, for some models, we find that there is a tendency of a positive bias. In other words, the measured shed is less than the real shed. Interestingly, this conclusion agrees with existing research studies that state that baseline models tend to understate the achieved load reductions (Coughlin *et al.*, 2009; Granderson *et al.*, 2021). This could result in a systematically lower compensation for the building that provided the services. Finally, the low values for the REL metric expose existing limitations of demand flexibility as a reliability resource. For all instances, a measured event was found more likely to be beyond the 10% threshold used for the REL metric. This is represented by all triple instances having a negative REL value. This motivates the need for more accurate baseline models, or other non-baseline methods to achieve the minimum performance thresholds required by certain grid services.

4. Conclusion

In this paper we present an analysis of how the choice of baseline model and assessment boundary affect the measurement uncertainty of demand flexibility events. We generate an extensive dataset of simulated demand flexibility events by varying these two factors for five different climate zones. Lastly, we calculate performance metrics on this dataset to compare the performance of each instance. In our analysis we found several conclusions:

- Baseline error performance metrics alone, without the context of the magnitude of the demand flexibility events, are not a good indicator of the expected event measurement error.

- More accurate event measurements can be obtained by using the HVAC assessment boundary than using the building assessment boundary. This comprises of an event MdAPE reduction from 34% to 21%, on average.
- The choice of baseline model should be made considering the assessment boundary. For example, averaging models can be appropriate for hotter climates at the building boundary, while machine learning approaches can be more accurate for variable climates at the HVAC boundary.
- Although most models exhibit low bias, some models tend to underrepresent the magnitude of the real event.

We identify several avenues for future work. Given that measurement uncertainty is inherent and unavoidable for demand flexibility, we recognize the need for methods that communicate the expected uncertainty associated with specific events. This can help improve the trust of grid operators when leveraging demand resources for reliability. Furthermore, in this work we do not explore the aggregate assessment boundary. Measurement uncertainty at the aggregate level has the potential to be lower given that it can be reduced over a large population of buildings. Measurement uncertainty at this boundary has the potential to be lower than the ones explored in this work. Finally, extending this analysis to real building data is necessary because there can be added uncertainties that cannot be accounted with a simulation model.

Acknowledgements

This material is based upon work partially funded by ENGIE Lab Crigen, (CSAI)

References

Aman, S., Simmhan, Y. and Prasanna, V.K. (2015) "Holistic Measures for Evaluating Prediction Models in Smart Grids," *IEEE Transactions on Knowledge and Data Engineering*, 27(2), pp. 475–488. doi:10.1109/TKDE.2014.2327022.

Amasyali, K. and El-Gohary, N.M. (2018) "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, 81(April 2017), pp. 1192–1205. doi:10.1016/j.rser.2017.04.095.

ASHRAE (2014) "Measurement of Energy, Demand, and Water Savings," *ASHRAE Guideline 14-2014*, pp. 1–150. Available at: www.ashrae.org%0Awww.ashrae.org/technology.

ASHRAE (2019) "ANSI/ASHRAE/IES Standard 90.1-2019 Energy Standard for Buildings Except Low-Rise Residential Buildings." Atlanta, GA: ASHRAE.

Cappers, P., MacDonald, J., Goldman, C. and Ma, O. (2013) "An assessment of market and policy barriers for demand response providing ancillary services in U.S. electricity markets," *Energy Policy*, 62, pp. 1031–1039. doi:10.1016/j.enpol.2013.08.003.

Chen, Y., Hong, T. and Luo, X. (2018) "An agent-based stochastic Occupancy Simulator," *Building Simulation*, 11(1), pp. 37–49. doi:10.1007/s12273-017-0379-7.

Coughlin, K., Piette, M.A., Goldman, C. and Kiliccote, S. (2009) "Statistical analysis of baseline load models for non-residential buildings," *Energy and Buildings*, 41(4), pp. 374–381. doi:10.1016/j.enbuild.2008.11.002.

FERC (2021) *2021 Assessment of Demand Response and Advanced Metering, Federal Energy Regulatory Commission*. Available at: http://www.ferc.gov/legal/staff-reports/2013/oct-demand-response.pdf.

Goel, S., Rosenberg, M., Athalye, R., Xie, Y., Wang, W., Hart, R., Zhang, J. and Mendon, V. (2014) *Enhancements to ASHRAE Standard 90.1 Prototype Building Models*. doi:10.2172/1764628.

Goldberg, M. and Agnew, G.K. (2013) *Measurement and Verification for Demand Response National Forum of the National Action Plan on Demand Response*.

Granderson, J., Sharma, M., Crowe, E., Jump, D., Fernandes, S., Touzani, S. and Johnson, D. (2021) "Assessment of Model-Based peak electric consumption prediction for commercial buildings," *Energy and Buildings*, 245, p. 111031. doi:10.1016/j.enbuild.2021.111031.

Granderson, J. and Price, P.N. (2014) "Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models," *Energy*, 66, pp. 981–990. doi:10.1016/j.energy.2014.01.074.

IRC (ISO/RTO Council) (2018) "North American Wholesale Electricity Demand Response Program Comparison."

Ji, Y., Xu, P., Duan, P. and Lu, X. (2016) "Estimating hourly cooling load in commercial buildings using a thermal network model and electricity submetering data," *Applied Energy*, 169, pp. 309–323. doi:10.1016/j.apenergy.2016.02.036.

KEMA (2011) "PJM Empirical Analysis of Demand Response Baseline Methods." Clarke Lake, MI: KEMA Inc.

KEMA-XENERGY (2003) "Protocol Development for Demand Response Calculation — Findings and Recommendations," (February).

Lei, S., Mathieu, J.L. and Jain, R.K. (2021) "Performance of Existing Methods in Baselining Demand Response From Commercial Building HVAC Fans," *ASME Journal of Engineering for Sustainable Buildings and Cities*, 2(2), pp. 1–13. doi:10.1115/1.4050999.

Mathieu, J.L., Callaway, D.S. and Kiliccote, S. (2011) "Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices," *Energy and Buildings*, 43(12), pp. 3322–3330. doi:10.1016/j.enbuild.2011.08.020.

Motegi, N., Piette, M.A., Watson, D.S., Sezgen, O. and ten Hope, L. (2004) "Measurement and Evaluation Techniques for Automated Demand Response Demonstration," in *2004 ACEEE Summer Study on Energy Efficiency in Buildings, Pacific Grove, CA*.

Neukomm, M., Nubbe, V. and Fares, R. (2019) *Grid-Interactive Efficient Buildings*. doi:10.2172/1508212.

Nexant (2017) "California ISO Baseline Accuracy Assessment," p. 59.

Roth, A. and Reyna, J. (2019) *Grid-Interactive Efficient Buildings Technical Report Series: Whole-Building Controls, Sensors, Modeling, and Analytics*. Golden, CO (United States). doi:10.2172/1580329.

Satchwell, A.J., Piette, M.A., Khandekar, A., Granderson, J., Frick, N.M., Hledik, R., Faruqui, A., Lam, L., Ross, S., Cohen, J., Wang, K., Urigwe, D., Delurey, D., Neukomm, M. and Nemtzow, D. (2021) *A National Roadmap for Grid-Interactive Efficient Buildings*.

Schiller, S.R., Schwartz, L. and Murphy, S. (2020) *Performance Assessments of Demand Flexibility from Grid-interactive Efficient Buildings: Issues and Considerations*.

Taylor, J.A. and Mathieu, J.L. (2015) "Uncertainty in Demand Response—Identification, Estimation, and Learning," in *The Operations Research Revolution*. INFORMS, pp. 56–70. doi:10.1287/educ.2015.0137.

Vindel, E., Bergés, M., Akinci, B. and Kavvada, O. (2021) "Demand flexibility potential model for multi-zone commercial buildings using internal HVAC system states," in *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. New York, NY, USA: ACM, pp. 176–179. doi:10.1145/3486611.3486654.

Wetter, M., Zuo, W., Nouidui, T.S. and Pang, X. (2014) "Modelica Buildings library," *Journal of Building Performance Simulation*, 7(4), pp. 253–270. doi:10.1080/19401493.2013.765506.

Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y. and Livingood, W. (2021) "A review of machine learning in building load prediction," *Applied Energy*, 285(July 2020), p. 116452. doi:10.1016/j.apenergy.2021.116452.