LBNL-XXXXX

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Commercial Building Energy Baseline Modeling Software: Performance Metrics and Method Testing with Open Source Models and Implications for Proprietary Software Testing

Philip N Price, Ph.D.,
Jessica Granderson, Ph.D.,
Michael Sohn, Ph.D.
Nathan Addy
David Jump, Ph.D., P.E.

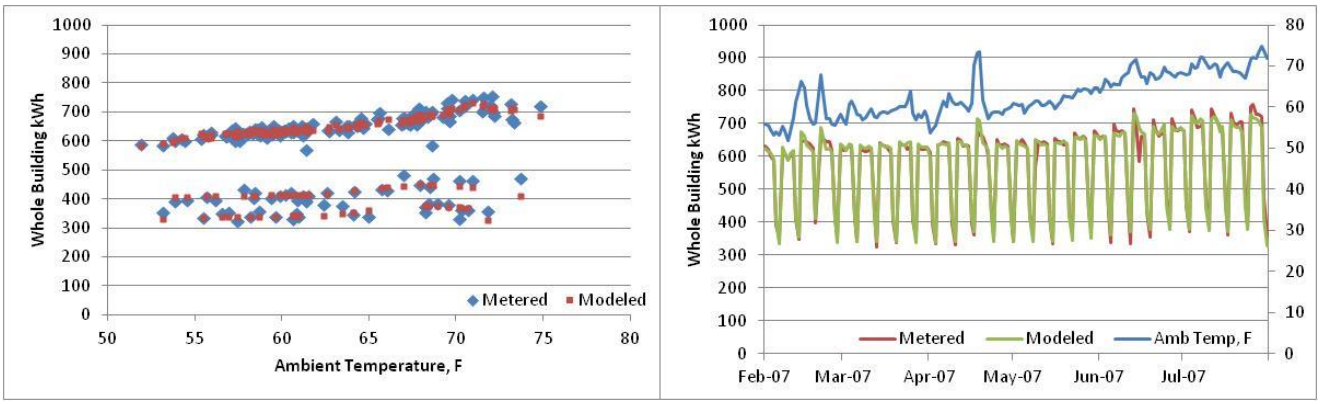
Environmental Energy Technology Division

September, 2013

Commercial Building Energy Baseline Modeling Software: Performance Metrics and Method Testing with Open Source Models and Implications for Proprietary Software Testing

Final Report

ET Project Number: ET 12PGE1311



Project Manager: **Leo Carrillo**
Pacific Gas and Electric Company

Prepared By:

Phillip N Price, Ph.D.
Jessica Granderson, Ph.D.
Michael Sohn, Ph.D.
Nathan Addy
Lawrence Berkeley National Laboratory
1 Cyclotron Road
Berkeley, CA 94720

David Jump, Ph.D., P.E.
Quantum Energy Services & Technologies, Inc. (QuEST)
2001 Addison St. Suite 300
Berkeley, CA 94704

Issued: **September 9, 2013**

Legal Notice

This report was prepared for Pacific Gas and Electric Company for use by its employees and agents. Neither Pacific Gas and Electric Company nor any of its employees and agents:

(1) makes any written or oral warranty, expressed or implied, including, but not limited to those concerning merchantability or fitness for a particular purpose;

(2) assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, process, method, or policy contained herein; or

(3) represents that its use would not infringe any privately owned rights, including, but not limited to, patents, trademarks, or copyrights.

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

ACKNOWLEDGEMENTS

This work described in this report was funded by the Pacific Gas and Electric Company and by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

This project was developed as part of Pacific Gas and Electric Company's Emerging Technology – Technology Development Support program under internal project number ET12PGE1311. Quantum Energy Services & Technologies, Inc. (QuEST) and Lawrence Berkeley National Laboratory (LBNL) conducted this evaluation for Pacific Gas and Electric Company with overall guidance and management from Leo Carrillo. For more information on this project, contact *lmcz*@pge.com.

The authors also want to acknowledge all others who assisted this project, including Portland Energy Conservation Inc., Agami Reddy, PGE's Mananya Chansanchai, Mangesh Basarkar, and Ken Gillespie, and the members of the Technical Advisory Group that PG&E organized for this project.

The TAG consisted of representatives from utilities, national laboratories, universities and other organizations. Members included:

Adam Hinge, Sustainable Energy Partnership
Amy Reardon, CPUC Energy Division
Bass Abushakra, Milwaukee School of Engineering
Carmen Best, CPUC Energy Division
Charles Middleton, PG&E
Charlie Culp, Texas A&M
Dries Berghman, FSC Group
Eric Martinez, SDG&E
Gavin Hastings, Arizona Public Service
Glenda Towns, SoCal Gas
Graham Henderson, BC Hydro
John Graminski, PG&E
Josh Bode, FSC Group
Kris Subbarao, PNNL
Leif Magnuson, PG&E
Martha Brook, CEC
Mary Anderson, TRC Solutions (formerly Heschong Mahone Group)
Mary Ann Piette, LBNL
Matt Golden, Efficiency.org
Michael Bobker, CUNY
Michael Brambley, PNNL
Nathaniel Taylor, SDG&E
Rob Rubin, SoCal Gas/SDG&E
Ryan Stroupe, PG&E
Sonny Enriquez, SCE
Tom Fenimore, Duke Energy
Tracy Narel, EnergyStar

CONTENTS

EXECUTIVE SUMMARY I
Background and Objectives i
Methodology ii
Results and Key Findings ii
Discussion x
Conclusions and Future Work xi
1. INTRODUCTION 1
2. STUDY METHODOLOGY 2
3. STUDY RESULTS 6
4. STUDY DISCUSSION AND CONCLUSIONS 16
Future Work 18
5. REFERENCES 19
APPENDIX 1: DATA 20
Load Data 20
Weather Data 23
APPENDIX 2: MODELS 24
Mean Week 24
Day-Time-Temperature Regression 24
Time-of-Week-and-Temperature Regression 24
Change Point Model 25
Cooling Degree-Days and Heating-Degree-Days Model 25
APPENDIX 3: ANALYSIS OF THE VOLUNTEER DATASET 26
One year training period, one year prediction period (12:12) 26
Screening Metrics 29
Bias as a Function of NAICS codes 29
Bias as a Function of Building Energy Consumption 30
Bias as a Function of the Goodness of Fit During Training Period .30
Bias as a function of load variability during the prediction period .31
Portfolio effects 32
APPENDIX 4: ANALYSIS OF THE REPRESENTATIVE DATA 34
One year training period, one year prediction period (12:12) 34
Portfolio Effects 37
Six-month training, 12-month prediction period (6:12) 38
Three-month training, 6-month prediction period (3:6) 39
Three-month training, 12-month prediction period (3:12) 40
APPENDIX 5: EVALUATING MODELS 41

EXECUTIVE SUMMARY

BACKGROUND AND OBJECTIVES

The overarching goal of this work is to advance the capabilities of technology evaluators in evaluating the building-level baseline modeling capabilities of Energy Management and Information System (EMIS) software. Through their customer engagement platforms and products, EMIS software products have the potential to produce whole-building energy savings through multiple strategies: building system operation improvements, equipment efficiency upgrades and replacements, and inducement of behavioral change among the occupants and operations personnel. Some offerings may also automate the quantification of whole-building energy savings, relative to a baseline period, using empirical models that relate energy consumption to key influencing parameters, such as ambient weather conditions and building operation schedule. These automated baseline models can be used to streamline the whole-building measurement and verification (M&V) process, and therefore are of critical importance in the context of multi-measure whole-building focused utility efficiency programs.

This report documents the findings of a study that was conducted to begin answering critical questions regarding quantification of savings at the whole-building level, and the use of automated and commercial software tools. To evaluate the modeling capabilities of EMIS software particular to the use case of whole-building savings estimation, four research questions were addressed:

1. What is a general methodology that can be used to evaluate *baseline model* performance, both in terms of a) overall robustness, and b) relative to other models?
2. How can that general methodology be applied to evaluate *proprietary models* that are embedded in commercial EMIS tools? How might one handle practical issues associated with data security, intellectual property, appropriate testing 'blinds', and large data sets?
3. How can buildings be pre-screened to identify those that are the most model-predictable, and therefore those whose savings can be calculated with least error?
4. What is the *state of public domain models*, that is, how well do they perform, and what are the associated implications for whole-building measurement and verification (M&V)?

Additional project objectives that were addressed as part of this study include: (1) clarification of the use cases and conditions for baseline modeling performance metrics, benchmarks and evaluation criteria, (2) providing guidance for determining customer suitability for baseline modeling, (3) describing the portfolio level effects of baseline model estimation errors, (4) informing PG&E's development of EMIS technology product specifications, and (5) providing the analytical foundation for future studies about baseline modeling and saving effects of EMIS technologies.

A final objective of this project was to demonstrate the application of the methodology, performance metrics, and test protocols with participating EMIS product vendors. The test protocols and demonstration, which will take place at a later date, will be reported separately.

The target audiences for this work are the energy efficiency program administrators, regulatory authorities, EMIS technology vendors, building science researchers, and national standards organizations.

METHODOLOGY

The evaluation methodology developed was a four-step process as follows:

1. Gather a large test data set comprised of interval data from hundreds of commercial buildings.
2. Split the data sets into model training and prediction periods. Tailor the training and prediction periods for the use case of interest.
3. For a given set of baseline models, generate predictions based on the training period data and compare those predictions to the prediction period data using a set of statistical performance metrics.
4. Assess each model's absolute performance relative to the use case, and assess each model's relative performance to other models.

Training and prediction periods were tailored for the use case of savings estimation through M&V. Training periods were selected to be 12, 6, and 3 month durations, while prediction periods were kept at 12 months because savings are reported on an annual basis. Analysis of varying training period durations provided insight about the amount of data actually needed to define a robust baseline model.

Several performance metrics were investigated in order to identify the most useful for the whole-building savings use case. These included model 'goodness-of-fit' metrics, such as R^2, monthly Mean Annual Percentage Error (MAPE) and the normalized root-mean-squared-error (n(RMSE)), and model predictability metrics, such as absolute and relative bias error.

Five open-source baseline modeling methods from the public domain literature were evaluated using this process. These models were: the mean week, day-time-temperature regression, time-of-week and temperature regression, change-point, and cooling-degree-days and heating-degree-days models. These models were selected because they were readily available and considered representative of the current state of art. They served as test models to define critical elements of the model evaluation methodology and their evaluation results provided benchmarks for comparison in evaluating proprietary models.

RESULTS AND KEY FINDINGS

Composition of the Test Data Set

Developing and evaluating empirical baseline models requires building energy use and independent variable data for each building included in the test data set. In the commonly used public domain models, ambient dry-bulb temperature and day/time information are the only variables used. Other weather variables such as humidity or incident solar energy are not commonly included in the most widely-used whole-building models, and measured occupancy is generally not available.

Two data sets were analyzed for this report. The "representative" data set initially included over 400 buildings that were randomly selected from the set of all small- to medium-sized commercial buildings in PG&E's service territory. Each individual building in the representative data set included two years of electric energy use data measured in 15-minute increments.

Each building's energy use over the period of interest was inspected. Buildings with obvious data quality issues, or which appeared to have been unoccupied for a portion of the time period, were removed. Each building included two years of electric energy use data measured in 15-minute increments. Building location, climate zone, and NAICS code[1] were known for these buildings. Ambient dry-bulb temperatures for each individual building's location were collected and included with the building energy use data. In a few cases, temperature data were not available from a weather station located within 15 miles of the building; these buildings were removed from the database. The representative data set ultimately contained data for 389 individual buildings.

An additional data set, referred to as the "Volunteer" data set, was selected from lists of PG&E customers who have participated in energy efficiency incentive programs. There were large differences between the Representative and Volunteer data sets in terms of types of businesses represented, the size of the buildings, and the energy consumption of the buildings. However, both the amounts of year-to-year energy consumption, and the statistical distributions of model errors, were almost the same in the Volunteer data set that they were in the Representative data set.

Both data sets were analyzed, but this study focuses on the analysis of the representative data set; results for the Volunteer dataset are included in an appendix.

Performance Metrics

One of the main parameters of interest is the error in the total amount of energy used during an evaluation period. Developing models from training period data and using them to estimate what energy use would have been during a future period is a key element of the M&V process. Our interest is in understanding the accuracy of these estimates. As calculated in Equation E-1, the "bias" in the total amount of energy used was one of the performance metrics of interest.

$$B = \hat{E}_{total} - E_{total} \qquad \text{(E-1)}$$

Where: $\hat{E}_{total}$ is the predicted energy use and E_{total} is the actual energy use during the prediction period.

Another performance metric of interest describes a model's ability to predict the total energy used for each individual month. A model that can predict well on a monthly basis may perform well when the duration of training or prediction periods are reduced. The *mean absolute percentage error* (MAPE) in monthly predictions is defined in Equation E-2. This metric is conceptually very similar to the normalized root mean squared error, a more common metric in the industry.

$$MAPE_{month} = \frac{\sum_{m=1}^{12} 100 \times \left| \frac{\hat{E}_m - E_m}{E_m} \right|}{12} \qquad \text{(E-2)}$$

[1] North American Industry Classification System (NAICS) codes are codes that classify government or business establishments according to a type of economic activity.

Where E_i = actual energy in month m, and $\hat{E}_i$ = predicted energy in month m.

Summary of Baseline Prediction Uncertainties

Figure E-1 shows the bias for all of the buildings in the representative dataset, for predictions made with the Mean Week model. The Mean Week model predicts the total building electricity use in Year 2 to be the same as in Year 1. In some buildings this leads to over-predicting the consumption in Year 2 (positive bias) and in some it leads to under-predicting (negative bias). As can be seen in the figure, In the dataset as a whole the statistical distribution of bias is very close to symmetric. The median and mean bias are close to zero. In most of this report we take advantage of the fact that the distribution of bias is nearly symmetric about zero to simplify the discussion of summary statistics: we usually summarize the absolute value of the bias rather than the bias itself.

However, although the bias distribution is nearly symmetric, it is not perfectly symmetric. Among larger buildings – those that consumed more than 300,000 kWh in Year 1 – more buildings were over-predicted than under-predicted: among these 195 buildings the median bias is 0.6% and the mean bias is 1.1%. That is, collectively, these buildings used about 1% less energy in 2011 than in 2010.

In principle, an overall decrease in building energy use could occur due to milder weather in 2011 than in 2010, but that is not the case here: the models that adjust for outdoor air temperature also over-predict the energy used by the larger buildings in 2011, on average. We speculate that the slight year-to-year decrease in energy use may be due to state economic conditions in 2010 and 2011.

The slight asymmetry in the bias distribution is unimportant in most contexts but may have important implications for the uncertainty in estimated baseline savings for a large portfolio of buildings, as we discuss later in this report.

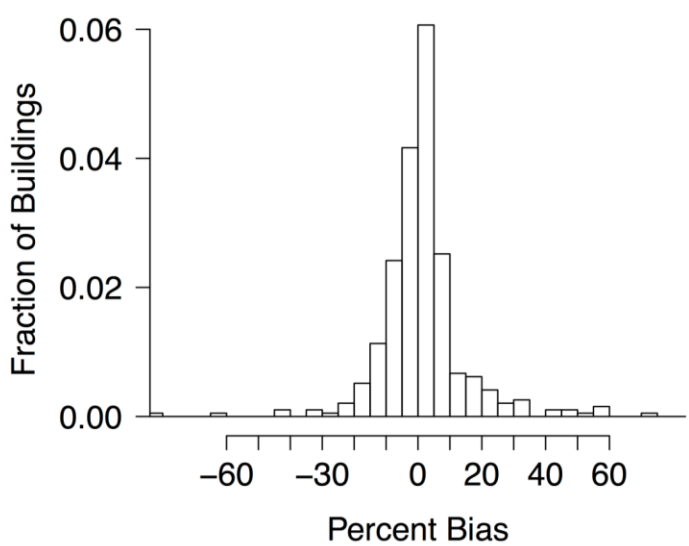


FIGURE E-1: STATISTICAL DISTRIBUTION OF PERCENT BIAS FROM THE MEAN WEEK MODEL, FOR ALL OF THE BUILDINGS IN THE REPRESENTATIVE DATASET

Table E-1 shows percentiles and the mean of the absolute percent bias for all the buildings in the representative data set, when the public domain models are fit using 12 months of training period data and then predict 12 months of energy use. Given this data set, training, and prediction period durations,

the mean bias for the public domain models was approximately 8.40%. For half the buildings in the data set, the bias was less than 4.82%. This suggests that for large representative data sets with one year pre- and post-installation data, models that exhibit mean biases much greater than 8.4% or median biases more than 4.8% would not measure up to the best public domain models, and may not be appropriate for whole building M&V in general. However, for a particular building such a model may exhibit better performance and be acceptable.

In terms of uncertainty, this table can be read to say a randomly selected building from the general population of buildings represented by the test data set has a 10% chance that the mean week model will predict its energy use within 0.82%, a 50% chance the prediction is within 4.82%, and so on.

TABLE E-1: QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE 389 BUILDINGS IN THE REPRESENTATIVE DATASET, BY MODEL. 12 MONTH TRAINING PERIOD, 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.82	2.21	4.82	9.63	19.42	8.40
Monthly CDD and HDD	0.69	2.09	4.53	10.03	19.38	8.46
Day, Time, and Temperature	0.69	2.17	4.51	9.26	19.41	8.42
Day and Change Point	0.73	2.02	4.70	9.22	18.84	8.24
Time of Week and Temperature	0.82	2.21	4.82	9.63	19.42	8.40

Mean monthly MAPE for the selected public domain models ranged from approximately 16% to 21% (see Table 4,p. 11, column "mean"). For half of the buildings in the data set, monthly MAPE was less than 14% and less than 10% for many of the models (Table 4, p.11, column "50%" [median]). This suggests that for large representative samples and one-year pre- and post- M&V conditions, models that exhibit mean monthly MAPE much greater than 20% or median monthly MAPE much greater than 10% would not measure up to the currently available public domain models.

Use Case Considerations Leading to Evaluation Criteria

For the whole-building savings estimation use case, the uncertainty in the estimation is directly proportional to the uncertainty in the baseline model's energy use prediction. In general a large signal-to-noise ratio is desired, meaning the savings should be much larger than the uncertainty. When the distribution is normal, it is common to apply a rule of thumb that the expected signal should be two times the standard deviation; that would make the signal larger than the noise in 95% of the cases. Since the distributions of bias errors are not normal, the assessment of expected savings could be compared to some user specified percentile, such as the 90% percentile bias. For the bias errors shown in Table E-1, this means that there is a 90% chance that the mean week model prediction is accurate within 19%, and a 10% chance that it is wrong by more than 19% in one direction or the other. The statistical distribution of errors is approximately symmetric – though not exactly so, as we discuss later – so there is approximately a 5% chance that the baseline prediction is more than 19% too high and approximately a 5% chance that it is more than 19% too low. In the 5% of cases in which the prediction is more than 19% too low, even a highly effective energy efficiency program that reduces the building's electricity consumption by 19% will be estimated to have provided no benefit or even to have been harmful.

This implies that the evaluation criteria for assessing model performance must consider the magnitude of savings, and the stakeholder's level of tolerance for savings uncertainty. What level of accuracy is needed, and in what percentage of buildings?

Screening Criteria

The uncertainty in whole-building savings calculation for a given building is due to the robustness of the baseline model used to determine those savings as well as the predictability of the building itself. Table E-1 shows the quantiles and the mean of the absolute percent bias for the entire test data set. With any of the models, for half of the buildings, total energy use can be predicted with less than 5% error, but some buildings are much less predictable: errors exceed 20% in almost 10% of buildings (as seen in the 90th percentile column). An energy efficiency incentives program might benefit substantially by eliminating (screening) the less predictable buildings. Several screening approaches were investigated: (1) selecting buildings by business type (as determined by NAICS code); (2) selecting buildings with little week-to-week variation during the training period; and (3) selecting buildings for which the baseline model fit well during the training period.

An analysis of the absolute bias for a well-performing public domain model for buildings with different NAICS codes showed that only two of the twelve groups of buildings have significantly reduced 95% percentiles of absolute bias, as shown in Table E-2. These included NAICS codes 62 - Health and Social Care, and 72 – Accommodation/Food, that showed model uncertainties to be higher than 9.8 and 12.4% in only 5% of the buildings. Taken as a whole, the results suggest that buildings with some NAICS codes may be more predictable than others, but that the effect is not large enough to use NAICS code as the principal screening criterion if avoiding unpredictable buildings is a priority.

TABLE E-2: QUANTILES OF ABSOLUTE BIAS IN TIME-OF-WEEK-AND-TEMPERATURE MODEL PREDICTIONS, IN BUILDINGS WITH DIFFERENT NAICS CODE PREFIXES. COMPARE TO TABLE 4.

NAICS prefix	Description	Bldgs	10%	25%	50%	75%	95%
42	Wholesale trade	14	1.69	3.41	7.47	13.66	22.1
44	Retail trade	41	0.47	2.36	4.20	8.69	15.2
45	Retail trade	12	1.08	2.24	3.17	5.70	13.3
49	Transport/Warehousing	10	2.74	3.14	5.50	11.66	16.4
51	Information	15	0.48	1.06	2.67	8.08	21.4
53	Real Estate Leasing	53	1.85	3.68	5.93	16.82	51.5
61	Education	42	0.84	1.90	3.58	7.86	15.8
62	Health and Social Care	36	0.55	1.50	3.21	5.93	9.8
71	Arts, Entertainment	30	1.25	3.26	6.80	18.22	30.3
72	Accommodation/food	63	0.64	1.87	3.85	8.70	12.4
81	Other	32	1.35	3.25	7.26	12.77	23.3
92	Public Administration	7	2.60	3.25	3.99	7.69	56.1

Second, a load-variability metric was considered that quantified the degree to which the building's load varies from week to week at the same time during the week. Buildings with highly variable energy use are thought to be difficult to model and to accurately predict energy use. The load variability metric calculated for the training period was compared to the model predictive error for the test data set to see if a correlation existed. As no correlation was found, this screening method did not prove useful in delineating which buildings would be predictable during the training period.

Lastly, the 'goodness of fit' during the training period was explored to investigate whether buildings which provide a better fit to the energy used each month of the training period tend to have more predictable total energy used during the prediction period. Model 'goodness of fit' was characterized by

the monthly MAPE metric. This method proved to be moderately useful in identifying whether a building's energy use during the prediction period was likely to be predicted accurately.

Table E-3 shows the effect of choosing only buildings whose monthly MAPE during the training period was less than 5%. The statistical distributions are similar for these models, but note that the number of buildings (N) is much smaller than that of the test data set. This screen did not lead to an overall shift in the distribution to lower bias, but it did help significantly at the high-percentile end by eliminating many of the worst fitting buildings. For these screened buildings, a 20% reduction in energy use can be reliably detected in almost all of the buildings.

TABLE E-3: QUANTILES OF ABSOLUTE PERCENT BIAS FOR BUILDINGS WHOSE MONTHLY MAPE IN THE TRAINING PERIOD WAS LESS THAN 5% FOR EACH MODEL. N SHOWS THE NUMBER OF BUILDINGS THAT PASS THE CRITERION; N DIFFERS FROM MODEL TO MODEL AND IS MUCH LOWER FOR THE MEAN WEEK MODEL.

Model	N	10%	25%	50%	75%	90%	Mean
Mean Week	62	3.44	4.53	5.60	7.85	9.98	7.38
Monthly CDD and HDD	209	3.68	4.50	6.10	8.70	12.82	8.16
Day, Time, and Temperature	202	2.70	3.58	5.40	7.78	12.08	7.03
Time of Week and Temperature	200	2.70	3.58	5.30	7.62	11.28	6.91

Public Domain Model Performance

When considering a 12 month training and 12 month prediction period, Tables E-1 and Table 4, p. 11 show that there is little difference in bias and monthly MAPE for the public domain models investigated. There are a few buildings for which the predictions are extremely poor, which lead to the average absolute bias being much worse than the median. For the monthly MAPE metric, the range in performance was slightly larger than for bias.

When the training period was reduced to 6 months, there was not a significant degradation in median error relative to 12 month training period results, as shown in Table E-4. The median rose from just under to just over 5%, and the mean rose only slightly on average. However, significantly poorer prediction performance was observed at the higher percentiles, with errors exceeding 100% for the least-predictable 10% of buildings.

TABLE E-4 QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE REPRESENTATIVE DATASET, BY MODEL. 6 MONTH TRAINING PERIOD STARTING IN FEBRUARY 2010, 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.02	2.19	5.39	11.60	108.1	9.37
Monthly CDD and HDD	0.08	2.73	6.31	15.23	242.0	13.1
Day, Time, and Temperature	0.02	2.11	5.34	10.53	110.6	9.19
Day and Change Point	0.00	2.42	5.94	11.42	107.9	9.66
Time of Week and Temperature	0.01	2.19	5.00	10.44	110.8	9.09

When the training period was reduced to three months, error rose significantly, and the models that included a time of week parameter performed significantly better than the others, as shown in Table E-5.

TABLE E-5: QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE REPRESENTATIVE DATASET, BY MODEL. 3 MONTH TRAINING PERIOD (USUALLY STARTING IN FEBRUARY 2010), 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	1.75	4.00	8.87	15.93	184.2	13.0
Monthly CDD and HDD	4.73	14.33	32.37	62.65	147.2	55.6
Day, Time, and Temperature	1.26	2.46	6.07	12.52	25.99	11.9
Day and Change Point	1.47	3.98	8.24	17.13	32.28	14.6
Time of Week and Temperature	1.15	2.48	6.12	12.75	26.10	11.9

Portfolio effects

We assessed the model performance when a collection of buildings are treated as a group. Table E-6 shows the results for groups of buildings defined by NAICS codes. The Table E-6 demonstrates the significant reduction in error at a portfolio level for the test data set.

TABLE E-6: BIAS FOR PORTFOLIOS BASED ON NAICS PREFIXES, WITH PREDICTIONS FROM THE TIME-OF-WEEK-AND-TEMPERATURE MODEL. ONLY PREFIXES WITH AT LEAST 10 CASES ARE INCLUDED, PLUS PREFIX 92 FOR COMPARISON TO THE VOLUNTEER DATA. THE YEAR 2010 WAS THE TRAINING YEAR FOR ALL MODELS.

NAICS prefix	Bldgs	Total kWh	Predicted kWh	Percent bias
42	14	7,844,788	7,696,758	-1.89
44	41	29,935,698	30,370,868	1.45
45	12	7,320,698	7,358,519	0.52
49	10	5,720,874	5,591,634	-2.26
51	15	13,770,148	13,601,572	-1.22
53	53	37,462,843	41,062,271	9.61
61	42	16,88,7745	17,403,489	3.05
62	36	20,238,549	21,001,653	3.77
71	30	7,430,195	7,573,492	1.93
72	63	23,302,962	22,971,386	-1.42
81	32	7,303,410	7,447,883	1.98
92	6	5,127,729	5,215,852	1.72

Another approach to improving the accuracy in the prediction of baseline energy use for large groups of buildings was to reduce the number of less predictable buildings in the group. This was investigated using the screening method based on monthly MAPE, but using successively more restrictive monthly MAPE thresholds. As the threshold is made stricter, fewer buildings are included in the portfolio, which leads to larger percent errors in the total predicted energy used. However, the buildings left in the portfolio tend to be more predictable. Results of this analysis are shown in Table E-7, where these competing effects cancel each other, and the percent bias of the portfolio varies only slightly as the threshold is made more restrictive and the number of buildings per group is reduced from 300 down to 3.

TABLE E-7: BIAS FOR A PORTFOLIO BASED ON BUILDINGS FOR WHICH THE FIT WAS GOOD DURING THE TRAINING PERIOD (MONTHLY MAPE < A SPECIFIED THRESHOLD), WITH PREDICTIONS FROM THE TIME-OF-WEEK-AND-TEMPERATURE MODEL. TRAINING YEAR WAS 2010 FOR ALL BUILDINGS.

MAPE threshold	Bldgs	Total kWh	Predicted kWh	Percent bias
10%	299	192,780,964	195,266,556	1.29
9%	288	189,256,859	191,507,069	1.19
8%	271	181,761,057	183,929,834	1.19
7%	254	173,359,941	175,784,108	1.40
6%	228	159,519,215	162,114,653	1.63
5%	200	146,771,988	148,758,101	1.35
4%	165	127,162,839	129,505,457	1.84
3%	110	88,809,120	90,545,326	1.95
2%	40	43,211,073	43,751,227	1.25
1%	3	5,189,497	5,250,785	1.18

Development of Testing Protocols

The evaluation methodology, and results were used as a basis for the development of software testing protocols. These protocols are written for the whole-building energy savings estimation use case and are in a generalized form so that other stakeholders may adapt them for their own regions and applications. Two testing protocols were written: a prequalifying test protocol in which proprietary or other public domain baseline modeling software may be evaluated for a target population of buildings, and a field test protocol, in which modeling software may be evaluated for a particular building. These protocols provide flexibility in the evaluation of baselining software performance depending on the need of the interested parties, and are intended as a starting point for further development. The protocols are available in the companion document: "Functional Testing Protocols for Commercial Building Efficiency Baseline Modeling Software."

Several key issues in implementing baseline model performance evaluations are addressed by the protocols, these include:

- Developing a test data set for the target building population
- Describing and assessing impacts of potential building screening methods
- Protecting intellectual property of vendor's proprietary model software
- Maintaining strict security of building owner information and data
- Assuring software test integrity
- Guidance for developing and applying performance criteria for the whole-building savings estimation use case

The first two key issues above are based on the results described previously, while the next three key issues are concerned with the practical requirements of implementing such software tests. The protocols address the vendor intellectual property issue by providing two pathways in which the evaluation may be conducted. The protocols prohibit access to building owner information and describe how data security may be maintained through the application of 'masks.'

The protocols are designed to maintain the integrity of the software test in order that the technology evaluators may rely on their results. Appropriate data 'blinds' are described so that prediction period

energy use is not shared with vendors. This focuses the evaluation on the quality of model predictions and prevents intervening with software predictions.

The protocols will be exercised in the upcoming product test demonstration. This demonstration will provide useful insight about useful testing strategies as well as feedback on the performance of selected proprietary models.

DISCUSSION

Development of a methodology to evaluate baseline model performance and proprietary tools

This work has developed and demonstrated a solid, general statistical methodology to evaluate baseline model performance as applied to the whole-building savings estimation use case. It has described considerations for building a test data set, identified evaluation metrics appropriate for the use case, and described training and prediction periods of key interest. Evaluation of results for public domain models identified threshold values of performance metrics, and provided a means to evaluate the absolute performance of proprietary methods for a population of buildings.

The evaluation methodology has been adapted for implementation in test protocols to evaluate EMIS vendor proprietary tools, so that utility program managers and other interested parties may understand how to objectively evaluate these tools. The protocols describe how to navigate the practical issues of protecting vendor intellectual property, maintaining data security, and assuring the test integrity. The protocols provide technology evaluators with tools to understand vendor software performance for a population of buildings as well as for a specific building. These protocols have yet to be exercised in a live demonstration.

Public Domain Model Performance

The bounds of performance accuracy that can be achieved when conducting *fully automated* (that is, without oversight of an energy analyst or building engineer) whole-building M&V were demonstrated. For the five public domain models and representative data set investigated, this work showed median model errors under 5% and mean errors less than 9%. For the general population of buildings represented by the test data set, these results mean that there is a 50% chance the error is less than 5%. The analysis further showed that there was a 10% chance the error would be larger than about 20% for the general population of buildings using the baseline models evaluated.

Screening method to reduce measurement and verification error, and target building recruitment

A screening method based on monthly model predictability was identified as a potentially promising means to pre-screen or target program participants. Of the three methods investigated: building use type, fifteen-minute load variability, and model fit to historic data, the latter was found to be moderately useful in filtering buildings with the most extreme modeling errors from the data set. Applying this screening method to the test data set improved the model predictions from 80% to more than 87% accuracy in 10% of the buildings.

Shorter Training Periods

This work showed that for a 12-month post-installation period, use of a six-month baseline period may generate results that are just as accurate as those based on a twelve month baseline period. This has important implications, as reducing the baseline period required for M&V helps in the scaling deployment of efficiency projects and reducing overall costs. However the analysis showed that there is

a significantly larger bias error in the higher quantiles, so applying an appropriate screening method may be warranted to assure accurate savings results.

Portfolio Effects

When buildings are aggregated into a portfolio, baseline model prediction errors tend to cancel out so that the predicted energy use error of the portfolio decreases dramatically. Definitions of portfolios of building groups included: at random, based on goodness of fit during the training period, or by building use type. Depending on how a portfolio is defined, for portfolio populations of up to 40 buildings, analysis showed that the total portfolio annual energy use can be predicted within 1.5 to 4% accuracy.

In general, the more buildings in a portfolio, the lower the baseline error for the portfolio is likely to be. However, the baseline error will not necessarily approach zero as the number of buildings is increased: phenomena that cause many buildings' energy consumption to be over- or under-estimated – such as an overall upward or downward shift in energy consumption in a large fraction of buildings due to changing economic conditions – can lead to systematic errors in any statistical model that does not account for the cause of the shift. In most years and most portfolios, such systematic errors are likely to be small, perhaps of the order of one or two percent (or less).

Collectively, these results suggest that automated baseline models and savings calculations can provide significant value in streamlining M&V calculations. They also suggest that savings can be reliably quantified at the whole building level using available short-time interval models available today. The level of confidence required and the depth of savings help inform stakeholders in selecting which modeling methods would be the most reliable.

Whole-building approaches to savings can include multi-measure savings strategies, including major system and equipment efficiency upgrades, operational improvements, and behavioral programs. This approach is expected to yield a higher level of savings, up to 20%. This work demonstrated that a small sample of public domain models is able to demonstrate savings accuracy within 5% for half of the cases, and within 20% for 90% of the cases. Note that no such accuracy prediction is available for engineering calculations, which are typically provided for single-measures which amount to 1 to 10% of whole-building energy use. Whole-building savings estimation should therefore be no more risky than engineering calculations.

Public Domain Models as a Performance Benchmark

A further result of this study showed that the predictive capability of public domain models provides a 'performance benchmark' for other models, including proprietary models. Performance metric distributions resulting from different models in different T:P scenarios may be directly compared, providing stakeholders with key insight about the relative capabilities of the models. It should be noted that model performance depends heavily on the test data set, T:P scenarios, and actual time period of the test data upon which they are evaluated, and apple-to-apples comparisons are recommended. The product test protocols describe how to obtain these comparisons.

Conclusions and Future Work

This work developed a methodology that was used to evaluate the baseline modeling accuracy of several public domain models on a test data set and that treated those models as 'black box' predictors of energy use. Relevant input parameters and evaluation metrics were identified for the whole-building savings estimation use case.

A representative data set of approximately 400 buildings was used to determine the threshold performance of the public domain models. This information provides a useful benchmark for comparison of proprietary models. The analysis of public domain models was also used to identify effective building screening methods, understand the impact of shorter training periods, and demonstrate portfolio effects for buildings grouped in different ways.

This study did not focus on identifying the best baseline models, an exercise that would ideally include a diversity of proprietary models, and models that include variables other than outside air temperature, day, and time. Furthermore, the protocol that was developed to integrate the model evaluation methodology with the blinds and protections necessary to handle proprietary models and commercial tools has not yet been applied to assess the performance of a representative set of commercial tools. That is a key next step in validating that the protocol is practical and scalable.

This study did not evaluate actual calculations of savings from applying baseline models to data from buildings in which efficiency projects were implemented. Such a study would yield important information regarding the impact on savings uncertainty from (1) duration of pre- and post-measure periods, (2) baseline model deterioration rate (when to re-baseline), (3) post-installation models. At a minimum, that investigation would require extensive data from before and after energy efficiency improvements have been implemented in each building. The volunteer data set that was collected for this project did include such data, so that the foundation of the suggested future work is already in place. This would also set the stage for a long-term study to directly compare of the uncertainty in measured approaches to the uncertainty in approaches based on engineering calculations.

1. INTRODUCTION

The overarching goal of this work is to advance the capabilities of technology evaluators in evaluating the building-level baseline modeling capabilities of Energy Management and Information System (EMIS) software. Through their customer engagement platforms and products, EMIS software products have the potential to produce whole-building energy savings through multiple strategies: building system operation improvements, equipment efficiency upgrades and replacements, and inducement of behavioral change among the occupants and operations personnel. Some offerings may also automate the quantification of whole-building energy savings, relative to a baseline period, using empirical models that relate energy consumption to key influencing parameters, such as ambient weather conditions and building operation schedule. These automated baseline models can be used to streamline the whole-building measurement and verification (M&V) process, and therefore are of critical importance in the context of multi-measure whole-building focused utility efficiency programs.

This report documents findings of a study that was conducted to begin answering the critical questions being asked in the utility programs industry regarding quantification of savings at the whole-building level, and the use of automated and commercial tools. For energy efficiency program applications one objective is to quantify and minimize the uncertainty in reported whole-building savings, which depends on baseline model robustness, building predictability, portfolio aggregation effects, and depth of savings being measured. To evaluate the modeling capabilities of EMIS, particular to the whole-building efficiency program context, this report explores four research questions:

1. What is a general methodology that can be used to evaluate *baseline mode*l performance, both in terms of a) overall robustness, and b) amongst candidate models?

2. How can the general methodology be applied to evaluate *proprietary models* that are embedded in commercial EMIS tools? How might one handle practical issues associated with data security, intellectual property, appropriate testing 'blinds', and large data sets?

3. How can buildings be pre-screened to identify those that are highly model-predictable and those that are not, in order to identify estimates of building-energy savings that have small errors/uncertainty?

4. What is the *state of public domain models*, i.e., how well do they perform, and what are the associated implications for whole-building measurement and verification?

The findings of this study can be used to (1) inform technology assessments for energy management and information system (EMIS) products and other technologies that deliver operational and/or behavioral savings; and (2) define the product specification requirements for efficiency programs that utilize energy-use modeling tools both to establish energy use baselines and to quantify energy savings. The target audience is energy efficiency program administrators, regulatory authorities, EMIS technology vendors and national standards organizations.

Additional project objectives that were addressed as part of this study include: (1) clarification of the use cases and conditions for baseline modeling performance metrics, benchmarks and evaluation criteria, (2) providing guidance for determining customer suitability for baseline modeling, (3) describing the portfolio level effects of baseline model estimation errors, (4) informing PG&E's development of EMIS technology product specifications, and (5) providing the analytical foundation for future studies about baseline modeling and saving effects of EMIS technologies. A final objective of this project was to demonstrate the application of the methodology, performance metrics, and test protocols with participating EMIS product vendors. The test protocols and demonstration, which will take place at a later date will be reported separately.

The target audiences for this work are the energy efficiency program administrators, regulatory authorities, EMIS technology vendors, and national standards organizations.

2. Study Methodology

Some key considerations for the development of an appropriate evaluation methodology included: (1) the methodology should focus on the quality of software predictions rather than on modeling algorithms; (2) models must be tested with a large number of randomly selected individual building data sets in order to maximize the statistical robustness of the results; (3) it must have general applicability to different use cases, not only whole-building energy savings estimation; and (4) the methodology must be able to evaluate proprietary models while maintaining vendor intellectual property, assure security of customer data, and maintain the evaluation integrity.

The methodology that was applied to answer each of the five research questions listed in Section 1 is provided in the following.

1. *What is a general methodology that can be used to evaluate baseline model performance, both in terms of a) overall robustness, and b) amongst candidate models?*

A four-step process leveraging large data sets and statistical goodness-of-fit metrics served as the starting point to define a methodology to evaluate baseline model performance. This process is summarized below, and illustrated in Figure 1.

1) Gather a large test data set comprised of interval data from hundreds of commercial buildings.

2) Split the test data set into model training and model prediction periods. These periods of interest can be tailored according to the specific application, or use case of interest, e.g., energy efficiency savings, demand response load reductions, or continuous energy anomaly detection. For this study, the focus was measurement and verification of energy savings at the whole-building level.
3) For a given set of baseline models, generate predictions based on the training period data, compare those predictions to the data from the prediction period, and compute statistical performance metrics based on the comparison. Again, the models of interest, and the

specific performance metrics can be tailored to according to the specific application or use case.

4) Assess relative and absolute model performance using the performance metrics that were computed in Step 3.

Although more general, this methodology shares some similarities to the approaches used in the ASHRAE Great Energy Predictor Shootouts I and II of the mid and late 1990s [Haberl 1998; Kreider 1994]. In both cases, cross-validation is used to determine model error, and compare one model's performance to another, and in both cases, coefficient of variation of the root mean squared error is included as a performance metric. However, the ASHRAE shootouts were each limited to data from just two buildings, and the cross-validation was conducted only during the model training period. The ASHRAE shootouts focused on hourly quantifications of energy use; the methods/models considered in this report can accommodates any unit of energy prediction that may be of interest. In addition, while the ASHRAE competitions considered total energy use from a sum of submetered quantities, the demonstration in this study is limited to whole building electric metering, and associated baseline models.

After establishing this *general* 4-step process, specific elements relevant to the whole-building M&V application were identified. The training and prediction periods of interest, as well as specific performance metrics, were defined purely based on a-priori knowledge of the use case. However composition of the test data set, threshold values for performance metrics, and a means of comparing one set of performance metrics to another, were elements that required analysis to appropriately define.

The analysis that was conducted to define these elements for the model evaluation methodology made use of 5 'open-source' models from the public domain literature. These models were selected because they were readily accessible, and representative of the current state of the art in engineering practice - *not* because they are unique, or were deemed to be the *best* whole-building baseline models. They were used as reference cases and a standard 'benchmark' to define critical elements of the general model evaluation methodology.

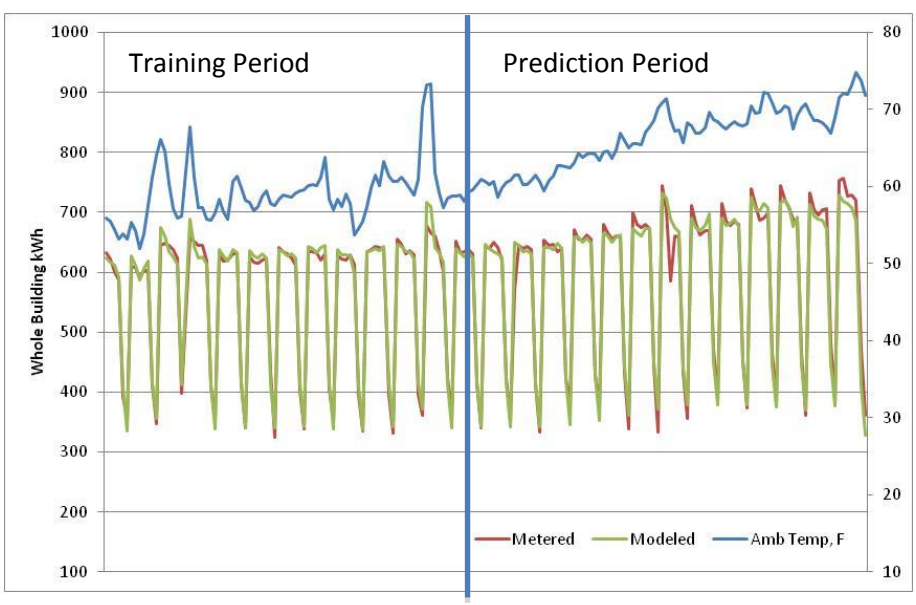


FIGURE 1: ILLUSTRATION OF MODELING TRAINING AND PREDICTION PERIODS. BASELINE MODELS ARE FIT TO DATA FROM THE TRAINING PERIOD AND PREDICTIONS ARE PROVIDED FOR THE PREDICTION PERIOD. THE PERFORMANCE OF THE MODELS ARE ASSESSED BY COMPARING MODEL PREDICTIONS TO ACTUAL METERED DATA OVER THE PREDICTION PERIOD.

2. How can the general methodology be applied to evaluate *proprietary models* that are embedded in commercial EMIS tools? How might one handle practical issues associated with data security, intellectual property, appropriate testing 'blinds', and large data sets? *How can the general methodology be applied to evaluate proprietary models that are embedded in commercial tools?*

To address issues such as customer data security and privacy, vendor software intellectual property, maintaining the test integrity, and the mechanics of working with large sets of data, flexibility in the application of the evaluation methodology is required. The test protocols must be developed to provide alternative pathways to address customer data security and vendor intellectual property, while maintaining appropriate 'blinds' to assure an unbiased outcome.

The research team outlined each testing protocol and obtained feedback from external stakeholders to define testing procedures that would be practical, and feasible. In addition, key insights from the analysis used to define the general technical methodology (under Question 1) were transferred to the development of the proprietary tool testing protocol.

3. How can buildings be pre-screened to identify those that are highly model-predictable and those that are not, in order to identify estimates of building-energy savings that have small errors/uncertainty?

The uncertainty in whole-building savings calculation for a given building is due to the robustness of the baseline model used to determine those savings, as well as the predictability of the building itself. Three potential indicators of model accuracy were investigated to determine their suitability for pre-screening. First NAICS code was considered. Of the several hundred buildings in the test data set, buildings with similar NAICS codes were grouped to determine whether certain building types were predicted with more or less accuracy than others. We examined the performance for all of the models (except the change point model, which was excluded due to a coding error).

Second, a load variability metric was considered that quantifies the degree to which the building's load varies from week to week at the same time during the week, as defined in Equation 1. For the several hundred buildings in the test data set, the predictive error was compared to the magnitude of the variability metric to determine if there was a correlation. To calculate the metric, the average load is calculated for each time interval during the week (there are 672 15-minute time intervals in a week). For each data point, the difference between the load and the average load at that time of the week is calculated and squared. The square root of the average of these squared errors defines the load variability metric, LV.

$$LV = \sqrt{\frac{\sum_{w=1}^{52} \sum_{t=1}^{672} (y_{w,t} - \bar{y}_t)^2}{(52)(672)}} \qquad (1)$$

Finally, goodness of fit during the training period was explored to determine whether buildings which provide a better fit to the energy used during each month of the training period tend to have more predictable total energy use during the prediction period. To quantify the model fit during the training period, we considered both the mean absolute percent error (MAPE), defined in Equation 3, and the normalized root mean squared error (sometimes erroneously called CV(RMSE)) in the predicted monthly energy use.

4. What is the *state of public domain models*, i.e., how well do they perform, and what are the associated implications for whole-building measurement and verification?

To assess how well public domain baseline models perform, five industry standard models were run under diverse training periods and prediction periods, a test data set of hundreds of buildings, and the evaluation methodology developed under Question 1. The mean week, day-time-temperature regression, time-of-week-and-temperature regression, change point, and cooling-degree-days and heating-degree-days models are described in Appendix 2. The combinations of training and prediction periods that were evaluated included: three, six, and twelve months of training for a 12-month prediction period; and three months of training for a 6-month and a 12-month prediction period. Distributions of model accuracy metrics (described below) were then assessed to determine how well the whole-building baseline models perform.

Distributions of model accuracy metrics are tabulated to show quantiles and the mean. Specifically, the 10th, 25th, 50th, 75th, and 90th percentiles are presented. Referring to Table 1 as an interpretive example from the actual study results, the 10th percentile error for the Mean Week model was 0.82. That means that for 10% of the buildings the model predicted the total energy used in the prediction period with an error of less than 0.82%. Similarly, the 50th percentile error was 4.82, meaning that in 50% of the buildings the error was less than 4.82%. Continuing with this example for the 90th percentile, in 90% of the buildings the error was less than 19.42%. (It follows that in 10% of the buildings the error was more than 19.42%).

Statistical distributions summarize the results. In real-world M&V cases, the true/correct value of the baseline prediction is unknown, and these statistical distributions summarize model prediction uncertainty. Model uncertainty must be expressed as a distribution in order to infer user-requested confidence/uncertainty/error ranges. In some cases, the distribution is well behaved, meaning the shape of the distribution follows a parametric distribution, such as Gaussian. In such cases, the distribution is summarized by the mean and the standard deviation. Unfortunately, the statistical distribution of errors in the present study is not well behaved (see Appendix 4). The standard deviation can be calculated, but it does not serve as a good overall summary of the distribution.

3. Study Results

Key findings from the study are presented in the following, according to each of the research questions that were addressed.

1. *What is a general methodology that can be used to evaluate baseline model performance, both in terms of a) overall robustness, and b) amongst candidate models?*

This work began with a general four-step general methodology to evaluate baseline model performance accuracy. To tailor that methodology to the application of focus, measurement and verification of whole-building energy savings, several parameters in the methodology were also defined.

Q1a. Composition of the test data set: Whole-building baseline models can include any number of independent variables that are then used to predict building load or energy use. Running and evaluating model performance therefore necessary requires *building load data*, in addition to *data for each of the independent variables*, for each building included in the test data set. In the most commonly-used industry standard models, outside air temperature, and day/time information from the interval meter time stamp are the only independent variables used. Other weather variables such as humidity are less commonly included in the most widely applied whole-building models. Variables such as measured occupancy are not generally available for the majority of buildings, or measured at the granularity of the weather and load data (typically hourly or more frequent).

The analyses conducted for this study focused on commonly used public domain models, and therefore on the composition of buildings and associated load data that should populate the test data set - outside air temperature is readily available from building location and weather feeds, whereas models that used other independent variables were not accessible to the research team.

The analyses used load data from two sources: a "volunteer" or "convenience" set of buildings that granted broad permission for their load data to be analyzed for purposes of evaluating energy efficiency, and a data set (called the "Representative" dataset) that was randomly selected from PG&E's mid-size commercial customers. Randomly selected data sets are in general preferred over volunteer data if the volunteer data do not represent the pool of buildings that are to be considered for future evaluation. In the present data, although there were large differences between the volunteer and representative datasets in terms of the types of businesses represented, the size of the buildings, and the energy consumption of the buildings, the amount of year-to-year variation in building energy consumption and the statistical distributions of model errors were almost the same That this is true of these particular datasets does not mean volunteer datasets can always be used, but at least it confirms that selection effects are not *necessarily* large with these types of data. Throughout the body of this report we focus on the Representative dataset; analysis and tables for the Volunteer dataset can be found in the Appendix 3.

The representative dataset included electricity data from about 400 buildings. We found that sample size is large enough to estimate the statistical distribution of baseline model errors for mid-sized commercial buildings *as a whole* –even a random sample 30 or 40 buildings appears to be adequate. However, we found that it is not large enough to distinguish differences in model performance between *different building types conclusively*. For example, the representative dataset included only 7 public

administration buildings and only 10 transportation and warehousing buildings. The sample sizes of these individual building types were too small to determine building type-specific differences model performance.

Gas usage data at daily intervals was received, but only for eight buildings. Of the eight buildings, only one year of therm usage data was provided. As this duration was too short for the training and prediction period combinations, this data was not analyzed. However, there is no reason the evaluation methodology cannot be carried out with gas usage data, given a long enough time period, and a representative sample of buildings within the data set.

Q1b. Training and prediction periods of interest: Given the whole-building M&V application case, a twelve month prediction period was deemed of most interest by external stakeholders. This is due to the fact that one year is the typical time period used to quantify efficiency project savings and payouts, and the fact that one year pre- and post-measure data are recommended in ASHRAE Guideline 14 [ASHRAE 2002]. Given a desire to shorten the overall M&V process, and therefore total project time, three, six, and twelve-month training periods were also of interest in evaluating model performance.

Q1c. Performance metrics: For whole-building measurement and verification (M&V) of energy efficiency measures, the main parameter of interest is the error, or uncertainty, in the total amount of energy used during an evaluation period. The error in total energy use during the prediction, or post-measure period, is referred to as the *bias*, and is defined in Equation 2 where E_{total} is the measured energy use and $\hat{E}_{total}$ is the model predicted energy use. A positive bias means the model predicted energy use higher than was measured.

$$B = \hat{E}_{total} - E_{total} \tag{2}$$

The second performance metric of interest relates to the ability to predict the total energy used for each individual month. This ability is desirable because if a model fits individual months well then it may be possible to reduce the duration of either the baseline period or the evaluation period. Additionally, if a model generally predicts well for individual months, but a few months stand out as being poorly predicted, this can help to locate problems that need attention and that might affect the efficacy or assessment of the energy conservation measure. The *Mean Absolute Percent Error (MAPE)* in the monthly energy predictions is defined in Equation 3. The MAPE metric is conceptually very similar to the coefficient of variation of the root-mean-squared error CV(RMSE) ,which is used in ASHRAE Guideline 14, which is a more common metric in the industry. Both Monthly MAPE and monthly CV(RMSE) were investigated; we found that MAPE proved marginally more useful for discriminating between buildings that have less- or more-predictable energy use.

$$MAPE_{month} = \frac{\sum_{m=1}^{12} 100 \times \left| \frac{\hat{E}_m - E_m}{E_m} \right|}{12} \tag{3}$$

Q1d. Summary of Baseline Prediction Uncertainties: Figure 2 shows the bias for all of the buildings in the representative dataset, for predictions made with the Mean Week model. The Mean Week model

predicts the total building electricity use in Year 2 to be the same as in Year 1. In some buildings this leads to over-predicting the consumption in Year 2 (positive bias) and in some it leads to under-predicting (negative bias). As can be seen in the figure, In the dataset as a whole the statistical distribution of bias is very close to symmetric. The median and mean bias are close to zero. In most of this report we take advantage of the fact that the distribution of bias is nearly symmetric about zero to simplify the discussion of summary statistics: we usually summarize the absolute value of the bias rather than the bias itself.

However, although the bias distribution is nearly symmetric, it is not perfectly symmetric. Among larger buildings - those that consumed more than 300,000 kWh in Year 1 - more buildings were over-predicted than under-predicted: among these 195 buildings the median bias is 0.6% and the mean bias is 1.1%. That is, collectively, these buildings used about 1% less energy in 2011 than in 2010.

In principle, an overall decrease in building energy use could occur due to milder weather in 2011 than in 2010, but that is not the case here: the models that adjust for outdoor air temperature also over-predict the energy used by the larger buildings in 2011, on average. We speculate that the slight year-to-year decrease in energy use may be due to state economic conditions in 2010 and 2011.

The slight asymmetry in the bias distribution is unimportant in most contexts but may have important implications for the uncertainty in estimated baseline savings for a large portfolio of buildings, as we discuss later in this report.

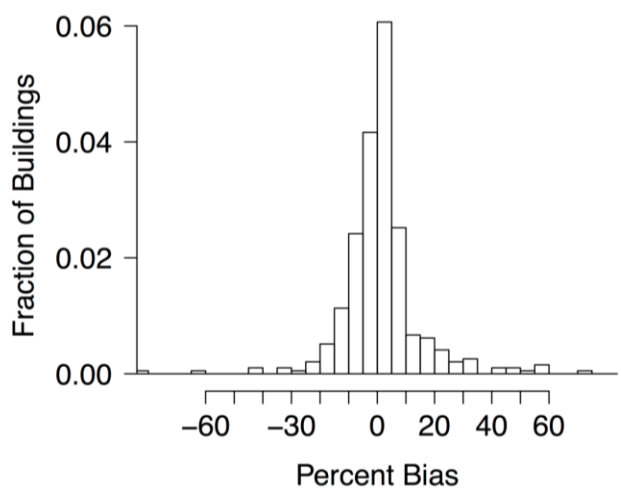


FIGURE 2: STATISTICAL DISTRIBUTION OF PERCENT BIAS FROM THE MEAN WEEK MODEL, FOR ALL OF THE BUILDINGS IN THE REPRESENTATIVE DATASET

Given a representative sample of building test data, one year of training data, and a one-year prediction period, mean absolute bias for the public domain models was approximately 8.4% (see Table 1, column "mean"). For half of the buildings in the data set, bias was less than 5% (Table 7, column "50%" [median]). This suggests that for large representative samples and one-year pre- and post- M&V conditions, models that exhibit mean biases much greater than 8% or median biases much greater than 5% would not measure up to the public domain models that are currently available, and may not be as appropriate for whole-building M&V *in general*. Of course, those models may exhibit much better

performance for specific, well-behaved *individual buildings*, with highly predictable loads. We discuss the implications of bias on model uncertainty in Question 3.

For the monthly MAPE metric, mean monthly MAPE for the public domain models ranged from approximately 16% to 21% (see Table 4, column "mean"). For half of the buildings in the data set, monthly MAPE was less than 14% and less than 10% for many of the models (Table 8, column "50%" [median]). This suggests that for large representative samples and one-year pre- and post- M&V conditions, models that exhibit mean MAPE much greater than 20% or median MAPE much greater than 10% would not measure up to the currently available public domain models.

2. How can the general methodology be applied to evaluate *proprietary models* that are embedded in commercial EMIS tools? How might one handle practical issues associated with data security, intellectual property, appropriate testing 'blinds', and large data sets? *How can the general methodology be applied to evaluate proprietary models that are embedded in commercial tools?*

The evaluation methodology's requirement to train models with training period data, and use the trained models to predict energy use for a large number of buildings, presents issues in the protection of customer data security and vendor intellectual property. The prequalifying protocol provides alternative pathways stakeholders may choose for their purposes and situations. In one path, the vendor provides their software to be run by the administrator (or its designated third party). Under this path, no customer data is shared with unauthorized parties, however vendors may be required to obtain non-disclosure agreements or other protections from the utility. On another path, the customer data is transformed to mask identifying characteristics and provided to the software vendors to generate predictions. These predictions are provided back to the test administrator for analysis. In this path, vendor intellectual property is protected and safeguards are put in place for the data.

The first pathway allows hundreds of data sets to be evaluated, as the administrator has free use of the vendor software. In the second pathway, the number of buildings in the dataset may need to be reduced due to potential data management burdens with large data sets.

The prequalifying test protocol describes how each vendor software's results may be scored to inform stakeholders how it performed for the test data set, and how its performance compared against typical public domain models. Stakeholders develop and apply the performance criteria developed from their assessment of savings risks. The protocols describe application of building screening criteria that stakeholders may employ to improve overall model performance results.

The field test protocol describes how a particular vendor's software may be tested for a particular building. This test may be desired to assure a good match of vendor modeling algorithms with a particular building and expected savings. Results of this test are whether the software meets or does not meet the minimum performance criteria required.

The test protocols are documented in the Emerging Technology Report titled: "Functional Testing Protocols for Commercial Building Efficiency Baseline Modeling Software."

3. How can buildings be pre-screened to identify those that are highly model-predictable and those that are not, in order to identify estimates of building energy savings that have small

errors/uncertainty? *How can buildings be pre-screened to identify those that are most model-predictable, and therefore those whose savings can be calculated with least error?*

Of the three screening criteria investigated, only one, monthly MAPE during the training period, proved to be moderately useful. Table 1 shows percentiles and mean of the bias for all of the buildings in the Representative dataset. Table 2 shows the results for only the buildings for which the Mean Absolute Percent Error in the monthly energy was low (less than 5%) during the training period, and Table 3 shows the results for which MAPE was less than 3% during the training period. Unexpectedly, this screen did not lead to an overall shift of the distribution towards lower bias – the 10th through 50th percentiles are actually worse than in the dataset without the screening – but it did help substantially at the bad end of the distribution by eliminating many of the worst-fitting buildings. Before applying this screen, even a 20% reduction in energy use could not be reliably detected in at least 10% of the buildings in the full dataset (Table 1), but could easily be detected in all or almost all of the buildings in the screened dataset (Tables 2 and 3). In addition to improving the performance at the 90th percentile, the screening also improves the mean absolute bias, from above 8%, to less than 7%.

Applying the screening criterion reduces errors, but also reduces the number of 'eligible' buildings. For all but one model, more than half of the buildings met the screening criterion of monthly MAPE < 5%, and the size of the dataset was kept to hundreds. For monthly MAPE < 3%, the effect was larger. This reducing effect was largest in the case of the Mean Week Model, where the monthly MAPE < 5% criterion was met for only 62 of the original 398 buildings in the data set, and the < 3% criterion was met for only 23. However, for the other models, more than half of the buildings in the dataset met the screening criterion.

TABLE 1: QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE 389 BUILDINGS IN THE REPRESENTATIVE DATASET, BY MODEL. 12 MONTH TRAINING PERIOD, 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.82	2.21	4.82	9.63	19.42	8.40
Monthly CDD and HDD	0.69	2.09	4.53	10.03	19.38	8.46
Day, Time, and Temperature	0.69	2.17	4.51	9.26	19.41	8.42
Day and Change Point	0.73	2.02	4.70	9.22	18.84	8.24
Time of Week and Temperature	0.82	2.21	4.82	9.63	19.42	8.40

TABLE 2: QUANTILES OF ABSOLUTE PERCENT BIAS FOR BUILDINGS WHOSE MONTHLY MAPE IN THE TRAINING PERIOD WAS LESS THAN 5% FOR EACH MODEL. N SHOWS THE NUMBER OF BUILDINGS THAT PASS THE CRITERION; N DIFFERS FROM MODEL TO MODEL AND IS MUCH LOWER FOR THE MEAN WEEK MODEL.

Model	N	10%	25%	50%	75%	90%	Mean
Mean Week	62	3.44	4.53	5.60	7.85	9.98	7.38
Monthly CDD and HDD	209	3.68	4.50	6.10	8.70	12.82	8.16
Day, Time, and Temperature	202	2.70	3.58	5.40	7.78	12.08	7.03
Time of Week and Temperature	200	2.70	3.58	5.30	7.62	11.28	6.91

TABLE 3: QUANTILES OF ABSOLUTE PERCENT BIAS FOR BUILDINGS WHOSE MONTHLY MAPE IN THE TRAINING PERIOD WAS LESS THAN 3% FOR EACH MODEL. N SHOWS THE NUMBER OF BUILDINGS THAT PASS THE CRITERION; N DIFFERS FROM MODEL TO MODEL AND IS MUCH LOWER FOR THE MEAN WEEK MODEL.

Model	N	10%	25%	50%	75%	90%	Mean
Mean Week	23	3.48	4.10	5.20	5.90	8.32	6.47
Monthly CDD and HDD	72	3.40	4.10	5.45	7.43	9.99	6.82
Day, Time, and Temperature	112	2.70	3.35	4.70	7.55	10.20	6.67
Time of Week and Temperature	110	2.69	3.32	4.55	7.20	10.10	6.33

4. What is the state of public domain models, i.e., how well do they perform, and what are the associated implications for whole-building measurement and verification?

The detailed analyses presented in the Appendices include many specifics related to the performance of public domain baseline models, taken over a range of conditions. Here we provide a summary framed according to some of the key questions related to whole-building M&V in the context of utility-delivered efficiency programs.

Q4.1 Relative performance of public domain models: When considering a 12-month training period and 12-month prediction period, there was relatively little difference in performance between the five public domain models. The results for the percent bias metric are shown in Table 3 above, and those for monthly MAPE are shown in Table 4 below. The median absolute bias is between 4.5 - 4.9% for all of the models, and the mean is between 8.3 - 8.5%. There are a few buildings for which the predictions are extremely poor, with errors greater than 75% (in either direction), and these lead to the average being much worse than the median.

TABLE 4: QUANTILES AND MEAN OF MONTHLY MEAN ABSOLUTE PERCENT ERROR (MAPE) FOR THE REPRESENTATIVE DATASET, BY MODEL. 12 MONTH TRAINING PERIOD, 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	5.72	8.80	13.80	23.10	38.30	21.51
Monthly CDD and HDD	4.10	5.40	8.80	16.30	32.64	16.39
Day, Time, and Temperature	3.19	5.00	8.30	15.57	31.20	15.88
Day and Change Point	4.22	6.30	10.2	17.90	33.58	17.50
Time of Week and Temperature	3.20	4.90	8.10	15.50	31.16	15.76

For the monthly MAPE metric, the range in relative performance was slightly larger than for bias, not by a large margin: the medians for the various models range from about 8 - 14%, and the means range from about 16 - 22%. Depending on the specific data set and buildings used, the values achieved for a given performance metric will differ. The results reported in Tables 3 and 4 correspond to a random sample of PG&E customer buildings; see subsection 1.a of this Study Results section for a discussion of how to compose a test data set for the purpose of general model evaluation.

Recall that this work investigated metrics that could be used to pre-screen specific buildings for *program recruitment*, to minimize the error in reported savings for a given site (see forthcoming subsection 3). It is important to note that this is in contrast to the selection of buildings to establish a test data set for model evaluation. When the MAPE-based screening filter was applied, model errors improved; in the case of the time of week and temperature model, shown in Table 1 (above), the mean error dropped from 8.4% to 6.9%, with other models based on interval data showing similar results; the screen was less

effective for the model that fit monthly total electricity use from monthly heating and cooling degree-days.

If the training period was reduced to 6 months, there was not a significant degradation in median error relative to cases in which 12 months of training data were provided. The exception was the monthly CDD and HDD model, which performed worse than the others on average, due to difficulty with the less predictable buildings. These results are shown in Table 5.

When the training period was reduced even farther, to only 3 months (Table 6), errors rose significantly. The time-of-week-and-temperature and day-time-and-temperature models consistently outperform the others.

TABLE 5: QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE REPRESENTATIVE DATASET, BY MODEL. 6 MONTH TRAINING PERIOD STARTING IN FEBRUARY 2010, 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.02	2.19	5.39	11.60	108.1	9.37
Monthly CDD and HDD	0.08	2.73	6.31	15.23	242.0	13.1
Day, Time, and Temperature	0.02	2.11	5.34	10.53	110.6	9.19
Day and Change Point	0.00	2.42	5.94	11.42	107.9	9.66
Time of Week and Temperature	0.01	2.19	5.00	10.44	110.8	9.09

TABLE 6: QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE REPRESENTATIVE DATASET, BY MODEL. 3 MONTH TRAINING PERIOD (USUALLY STARTING IN FEBRUARY 2010), 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	1.75	4.00	8.87	15.93	184.2	13.0
Monthly CDD and HDD	4.73	14.33	32.37	62.65	147.2	55.6
Day, Time, and Temperature	1.26	2.46	6.07	12.52	25.99	11.9
Day and Change Point	1.47	3.98	8.24	17.13	32.28	14.6
Time of Week and Temperature	1.15	2.48	6.12	12.75	26.1	11.9

Q4.2 Portfolio effects: The results discussed so far have focused on distributions of errors for collections of many individual buildings. However, prediction errors are much smaller when aggregated over a collection of buildings, which and are treated as a group. A portfolio of buildings will usually include some in which the prediction is too low and others in which it is too high. Although the magnitude of the error will tend to increase as buildings are added to a portfolio, the relative error will tend to decrease or remain stable/constant. (The principle is the same as with flipping coins: if you repeatedly flip a coin, the absolute difference between the number of heads and the number of tails will tend to increase, but the relative difference will tend to zero).

One approach to evaluating errors at the portfolio level is to group buildings with similar uses together. For example, retail stores and public administration buildings form separate portfolios. In all of these cases the percent bias in the prediction of the portfolio's energy use is less than the mean bias for the individual buildings of that type, because of the aggregation discussed above. Table 7 shows the

aggregation of buildings by NAICS code. The percent bias for the portfolio is often less than 2%. In comparison without aggregation, the median percent biases by NAICS code ranges from 2.7 to 7.3 (Appendix 4, Table A4-4).

TABLE 7: BIAS FOR PORTFOLIOS BASED ON NAICS PREFIXES, WITH PREDICTIONS FROM THE TIME-OF-WEEK-AND-TEMPERATURE MODEL. ONLY PREFIXES WITH AT LEAST 10 CASES ARE INCLUDED, PLUS PREFIX 92 FOR COMPARISON TO THE VOLUNTEER DATA. THE YEAR 2010 WAS THE TRAINING YEAR FOR ALL MODELS.

NAICS prefix	Bldgs	Total kWh	Predicted kWh	Percent bias
42	14	7,844,788	7,696,758	-1.89
44	41	29,935,698	30,370,868	1.45
45	12	7,320,698	7,358,519	0.52
49	10	5,720,874	5,591,634	-2.26
51	15	13,770,148	13,601,572	-1.22
53	53	37,462,843	41,062,271	9.61
61	42	16,88,7745	17,403,489	3.05
62	36	20,238,549	21,001,653	3.77
71	30	7,430,195	7,573,492	1.93
72	63	23,302,962	22,971,386	-1.42
81	32	7,303,410	7,447,883	1.98
92	6	5,127,729	5,215,852	1.72

Another approach to evaluating errors at the portfolio level is to group buildings by monthly MAPE or restrict the portfolio to a certain monthly MAPE threshold. Table 8 shows the percent bias for various levels of low monthly MAPE. The table shows that the percent bias for MAPEs less than 10% is between one and two percent. Stricter monthly MAPE thresholds means fewer buildings in the portfolio, which one would expect to lead to larger percent bias, but appears to be counter pointed by the models that are better at predicting energy use. The percent bias in the portfolio varies only slightly for portfolios ranging from 3 to 300 buildings. Table 8 indicates that when 200 buildings are aggregated together the total error is reduced to 1.35%. In contrast, without aggregation, Table 2 shows that the mean percent bias for buildings with monthly MAPE less than 5% was 6.9% for the time of week and temperature model, and the 10th percentile error was 2.7%.

TABLE 8: BIAS FOR A PORTFOLIO BASED ON BUILDINGS FOR WHICH THE FIT WAS GOOD DURING THE TRAINING PERIOD (MONTHLY MAPE < A SPECIFIED THRESHOLD), WITH PREDICTIONS FROM THE TIME-OF-WEEK-AND-TEMPERATURE MODEL. TRAINING YEAR WAS 2010 FOR ALL BUILDINGS.

MAPE threshold	Bldgs	Total kWh	Predicted kWh	Percent bias
10%	299	192,780,964	195,266,556	1.29
9%	288	189,256,859	191,507,069	1.19
8%	271	181,761,057	183,929,834	1.19
7%	254	173,359,941	175,784,108	1.40
6%	228	159,519,215	162,114,653	1.63
5%	200	146,771,988	148,758,101	1.35
4%	165	127,162,839	129,505,457	1.84
3%	110	88,809,120	90,545,326	1.95
2%	40	43,211,073	43,751,227	1.25
1%	3	5,189,497	5,250,785	1.18

Q4.3 Implications of baseline errors and portfolio effects for utility incentive payments: One method to investigate the implications of baseline model errors/uncertainties at the portfolio level is to consider how an analysis results in correct or incorrect incentive payments. To do so, we use the existing dataset as if was the data represents the results from an incentive program for a portfolio of buildings that had implemented energy conservation measures with a defined depth of savings. We used the following procedure to generate a hypothetic incentive program:

1. Create a portfolio of 20 buildings by randomly selecting buildings from the database "with replacement" (i.e., such that a given building can be sampled more than once).
2. Treat the actual metered energy used by these buildings in the prediction period as the "true baseline use."
3. Calculate the baseline projected use, using whatever model is of interest.
4. Choose a "percent savings" representing the energy savings that this portfolio will achieve.
5. Calculate the "actual savings" by multiplying the percent savings by the metered energy use in the prediction period; if the buildings had performed retrofit with the stated average effectiveness, this is how much energy they would have saved.
6. Calculate the "estimated savings" by finding the "baseline projected use" minus the "actual savings."
7. Calculate the "error in the estimated savings" by subtracting the actual savings from the estimated savings.
8. Calculate the financial implications. For example, what is the excess dollar amount paid by a utility incentive program due to the error in the savings estimate, if the incentive is 10 cents per kWh.
9. Repeat this procedure many times to produce many portfolios of 20 buildings for each chosen "percent savings".

Results of 10 portfolios of 20 buildings are summarized in Table 9 using baseline estimates computed from the time of week and temperature model. (Results for the day-time-temperature model and the change point model are very similar). Table 10 shows the mean and standard deviation of the same calculations in Table 3 but for many portfolios generated for each percent savings and for various numbers of buildings within the portfolio.

Overall, the results show that the models predict slightly higher baseline energy use than the actual in this dataset, and that this results in a greater estimated energy savings than the actual savings. This is due to the slight asymmetry of the bias distribution that was previously discussed: in California, more mid-size commercial buildings decreased their energy use in 2011 than increased it, so models that were trained on 2010 data tend to slightly over-predict the 2011 baseline. In the present dataset, the effect of the over-predicted baseline (on average) is an excess of incentive dollars paid out, for most randomly selected portfolios: the incentive payment would be 30% too high, on average, for a portfolio of buildings that attained 5% savings, or 10% too high if the buildings attained 15% savings. In a different state, a different year, or a different collection of buildings, this error could go the other direction.

TABLE 9: EFFECT OF BASELINE ERRORS ON TOTAL ESTIMATED SAVINGS FOR A PORTFOLIO OF N BUILDINGS. RESULTS ARE BASED ON HYPOTHETICAL RANDOM SELECTIONS OF BUILDINGS, WITH BASELINE PREDICTIONS FROM THE TIME OF WEEK AND TEMPERATURE MODEL.

N	Percent Savings	True Baseline Use (MWh)	Baseline Projected Use (MWh)	Actual Savings (MWh)	Estimated Savings (MWh)	Error in Estimated Savings (MWh)	Error in Estimated Savings (%)	Excess incentive dollars awarded at $0.1/kWh
20	5	8132	8123	407	397	-9	-2	-914
		18962	19821	948	1807	859	92	85,880
		10727	10885	536	695	158	29	15,822
		11676	11174	584	82	-502	-86	-50,168
		10087	10059	504	477	-28	-6	-2,754
		10205	10343	510	648	138	27	13,803
		11419	11590	571	742	171	30	17,089
		8940	8944	447	451	4	1	350
		15742	16134	787	1178	391	50	39,113
		10507	10335	525	354	-172	-33	-17,173
20	10	7454	7985	745	1276	531	71	53,081
		8308	8268	831	791	-40	-5	-4,005
		11040	11226	1104	1290	186	17	18,605
		12165	12789	1217	1841	624	51	62,402
		11954	11855	1195	1096	-99	-8	-9,910
		8636	8412	864	639	-224	-26	-22,409
		6032	6077	603	649	46	8	4,562
		13975	15400	1397	2823	1426	102	142,586
		9319	9354	932	967	35	4	3,465
		9727	9422	973	668	-305	-31	-30,455

TABLE 10: SUMMARY OF STATISTICAL DISTRIBUTIONS OF BASELINE ENERGY CONSUMPTION, PREDICTIONS, SAVINGS, AND ERRORS, FOR VARIOUS AMOUNTS OF ENERGY SAVINGS, FOR RANDOMLY SELECTED PORTFOLIOS FROM THE REPRESENTATIVE DATASET.

N	% Savings	True Baseline Use (MWh) Mean, SD	Baseline Projected Use (MWh) Mean, SD	Actual Savings (MWh) Mean,SD	Error in Estimated Savings (MWh) Mean,SD	Error in Estimated Savings (%) Mean,SD	Excess incentive dollars awarded at $0.1/kWh Mean,SD
20	5	11000, 2500	11200, 2580	540, 120	160, 360	30, 70	16000, 35000
	10			1080, 250		15, 35	
	15			1620, 380		10, 20	
40	5	21700, 3550	22000,3600	1090, 180	320, 520	30, 50	32000, 53000
	10			2190, 360		15, 25	
	15			3270, 540		10, 15	
80	5	43800, 5000	44500, 5150	2190, 260	640, 720	30, 30	64000, 71000
	10			4380, 510		15, 15	
	15			6570, 780		10, 5	

4. Study Discussion and Conclusions

Development of a methodology to evaluate baseline model performance and proprietary tools
This work has developed and demonstrated a solid, general statistical methodology to evaluate baseline model performance. The specific parameters in the general methodology were defined for use in applications focused on whole-building measurement and verification for efficiency programs. Namely, considerations for building up a test data set, performance metrics most relevant to M&V for whole-building energy savings, training and prediction periods of key interest, and threshold values for performance metrics. This work complements and extends prior research efforts such as the ASHRAE Shootouts of the 1990s [Haberl 1998; Kreider 1994] and a more recent study conducted by LBNL [Granderson 2012] .

In addition, a protocol has been developed to apply that performance evaluation methodology to assess commercial 'black box' tools. The key considerations in the development of the evaluation methodology were addressed in the protocol, while allowing some flexibility in the application of certain protocol elements. The test protocols focus on the quality of the proprietary software predictions rather than on their modeling algorithms. They require the software to be tested with a large number of randomly selected building data sets to fully populate the distributions of the results. They provide safeguards to protect vendor intellectual property, customer information and data security, and overall test integrity.

Screening to reduce measurement and verification error, and target program recruitment
Since errors in whole-building measurement and verification are due to the robustness of the baseline model *in combination with* the predictability of the building, it may be possible to determine building- or load-specific characteristics that correlate with smaller errors. These characteristics or metrics might then be used to pre-screen or target program participants. Filtering buildings for which monthly MAPE was less than 5% eliminated many of the most extreme errors for a large dataset.

In most buildings and most years, the largest source of year-to-year change in energy use is neither energy conservation measures nor year-to-year variation in weather, it is changes in characteristics of building operation and occupant behavior such as operating hours, thermostat settings, the number of occupants, the type of activities performed in the building, and so on. Surveys of building owners or occupants, or additional types of data (such as occupancy data for hotels, sales data for retail buildings, etc.) might allow better ability to screen out unpredictable buildings, as well as better models. For models fit to much less than a full year of data, and therefore to reduce the time required for M&V, the range of temperatures present in the training period has a substantial effect on the accuracy of the model in many buildings. This is an area that should be explored in more detail.

The state of public domain models and implications for whole-building savings quantification
This work showed that for a 12-month post-measure installation period, use of a six-month baseline period, i.e., six months of training data, may generate results that are just as accurate as those based on a 12-month baseline period. This has important implications, as reducing the total length of time required for M&V is key to scaling the deployment of efficiency projects in general, and reducing overall costs. Although existing M&V guidelines recommend a full 12 months of pre- and post- data, these guidelines were developed when monthly data was the standard. Improved baseline models that take advantage of increasingly available interval meter data may not require a full 12-months to develop an accurate baseline.

The analyses conducted for this study were useful in illustrating the bounds of performance accuracy that can be achieved when conducting *fully automated* whole-building measurement and verification. That is, the best performance that can be achieved without the oversight of an engineer to identify non-routine adjustments or incorporate knowledge regarding changes in building occupancy or operations. With the public domain models that were available for investigations, and the representative dataset of hundreds of buildings, this work showed median model errors of under 5% and mean errors of less than 9%. When prescreening was conducted to intentionally target participants to minimize baseline errors, the median error was actually increased slightly but the mean error was reduced to under 7%, and most of the least predictable buildings were eliminated, for a screening criterion that was satisfied by half of the buildings. Using a more restrictive screening criterion, even more of the very poorly predictable buildings were eliminated; for the best-performing model, that criterion was satisfied by about a quarter of the buildings and the mean error was reduced to under 6.5%, with 90% of the building baselines being predicted to within 10%.

As typically practiced, M&V is not fully automated, but is conducted *by an engineer* who has access to information about building occupancy, internal loads, and operations. They can therefore apply their expertise and insights to develop baseline 'adjustments' which tailor savings calculations to the particular building being evaluated. For example, in this study 20% of the buildings in a representative sample exhibited large changes in load that might be straightforward for an engineer to identify and account for, but are not easily handled in the fully-automated case.

Collectively, these results suggest that modern tools, with their automated baseline models and savings calculations can *at a minimum*, provide significant value in streamlining the M&V process, providing results that could be quickly reviewed by an engineer to determine if adjustments and further tailoring are necessary. They also suggest, that savings can be reliably quantified at the whole-building level, using the interval data-based models that are available today. Depending on the level of confidence required, and the precise depth of savings expected, these savings might be quantified in a fully automated manner (deeper savings, lower confidence), or with some engineering intervention (shallower savings, higher confidence).

Whole-building approaches to savings can include multi-measure savings strategies, including major system and equipment efficiency upgrades, operational improvements, and behavioral programs. This multi-measure approach is expected to yield a higher depth of savings, of up to 20% or more. As a point of reference, retro-commissioning (RCx) alone, saves on average 16% in commercial buildings [Mills 2009]. This work demonstrated that a small sample of public domain models is able to demonstrate savings accuracy within 20 percentage points for 90% of the cases, and within 5 percentage points for 50% of the cases. With very simple prescreening, accuracy improves by 1-2 percentage points. Note that no such accuracy prediction is available for engineering calculations, which are typically provided for single-measures that amount to 1 to 10% of whole-building energy use. Whole-building savings estimation should therefore be no more risky than engineering calculations.

When buildings are aggregated into a portfolio, errors tend to cancel out so that the percent error in the predicted energy use decreases substantially. Depending on the method of creating the portfolio (e.g. at random, or by screening on the goodness of fit during the training period, or by selecting buildings of a given business type), the total annual energy use of a portfolio of about 40 buildings can usually be predicted within 1.5 – 4% accuracy. The benefits of portfolio aggregation would not impact any

individual customer or program participant, but *are* relevant from the perspective of the utility, which may report savings at the aggregated level of many programs, or many buildings.

Public Domain Models as a Performance Benchmark

A further result of this study showed that the predictive capability of public domain models provides a 'performance benchmark' for other models, including proprietary models. Performance metric distributions resulting from different models in different T:P scenarios may be directly compared, providing stakeholders with key insight about the relative capabilities of the models. It should be noted that model performance depends heavily on the test data set, T:P scenarios, and actual time period of the test data upon which they are evaluated, and apple-to-apples comparisons are recommended. The product test protocols describe how to obtain these comparisons.

FUTURE WORK

The analyses in this study made use of freely available public domain reference models to determine: a) specific parameters to evaluate the performance of models used for whole-building measurement and verification of energy savings; b) the general state of the models that are most commonly used by today's engineers. This study did not focus on identifying the *best* whole-building baseline models, an exercise that would ideally include a diversity of proprietary models, and models that include variables other than outside air temperature, day, and time. The protocol that was developed to integrate the model evaluation methodology with the blinds and protections necessary to handle proprietary models and commercial tools has not yet been applied to assess the performance of a representative set of commercial tools. That is a key next step in validating that the protocol is practical and scalable.

This study did not evaluate actual calculations of savings from applying baseline models to data from buildings in which efficiency projects were implemented. Such a study would yield important information regarding the impact on savings uncertainty from (1) duration of pre- and post-measure periods, (2) baseline model deterioration rate (when to re-baseline), (3) post-installation models. At a minimum, that investigation would require extensive data from before and after energy efficiency improvements have been implemented in each building. The volunteer data set that was collected for this project did include such data, so that the foundation of the suggested future work is already in place. This would also set the stage for a long-term study to directly compare of the uncertainty in measured approaches to the uncertainty in approaches based on engineering calculations.

The design of pay-for-performance incentive programs, in light of the expected model performance and uncertainty is also a potential area of future work.

Finally, it is important to acknowledge that this study was constrained to whole-building site-level measurement and verification. Evaluation measurement and verification (EM&V), is a distinctly different application that includes sampling of projects, considerations of attribution, and net as opposed to gross savings. Although whole-building site-level M&V is a common step in both cases, a compelling area of future work will be to explore how these findings and statistical approaches can inform questions related to streamlining the larger EM&V process.

5. REFERENCES

ASHRAE 2002. ASHRAE Guideline 14-2002 for Measurement of Energy and Demand Savings, American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, GA.

Bonneville Power Administration, "Regression for M&V: Reference Guide," September 2011.

Cleveland, W. S. (1981) LOWESS: A program for smoothing scatterplots by robust locally weighted regression. The American Statistician, 35, 54.

Coughlin K., Piette M.A., Goldman C., and Kiliccote S., Statistical analysis of baseline load models for non-residential buildings. Energy and Buildings 41 (4), 374-381.

Granderson J., and Price P.N., Evaluation of the Predictive Accuracy of Five Whole-Building Baseline Models, Lawrence Berkeley National Laboratory report LBNL-5886E, 2012.

Mathieu JL, Price PN, Kiliccote S, and Piette MA. Quantifying Changes in Building Electricity Use, with Application to Demand Response. IEEE Transactions on Smart Grid, 2:507-518, 2011.

Mills, E. Building commissioning: A Golden Opportunity for Reducing Energy Costs and Greenhouse Gas Emissions. Report prepared for the California Energy Commission, Public Interest Energy Research. July 2009, LBNL Report No. 3645-E.

Haberl, J. S., Thamilseran, S., The Great Energy Predictor Shootout II Measuring Retrofit Savings, ASHRAE Journal, Vol.; No. 1, pp. 49-56. 1998. Also Energy systems Laboratory Report ESL-PA-98-01-01, 1998.

Kreider, J. F.; Haberl, J. S., Predicting Hourly Building Energy Use: The Great Energy Predictor Shootout – Overview and Discussion of Results, ASHRAE Transactions, Vol. 100, Pt. 2, pp. 1104-1118. 1994. Also Energy Systems Laboratory Report ESL-PA-94-06-01.

Appendix 1: Data

Load Data

We analyzed data from two sources:

- The first dataset is a convenience sample of buildings whose owners or managers agreed to have their buildings included in a previous study. We call this the “Volunteer” dataset. After eliminating buildings with various data problems (discussed below), this dataset consists of 67 buildings. The data come from years 2008, 2009, 2010, and 2011, but not all buildings have data from all years.
- The second dataset constitutes a representative random sample from PG&E’s medium and large commercial customers. We call this the “Representative” dataset. After eliminating buildings with data problems, this dataset consists of 389 buildings, each of which has data from 2010 and 2011.

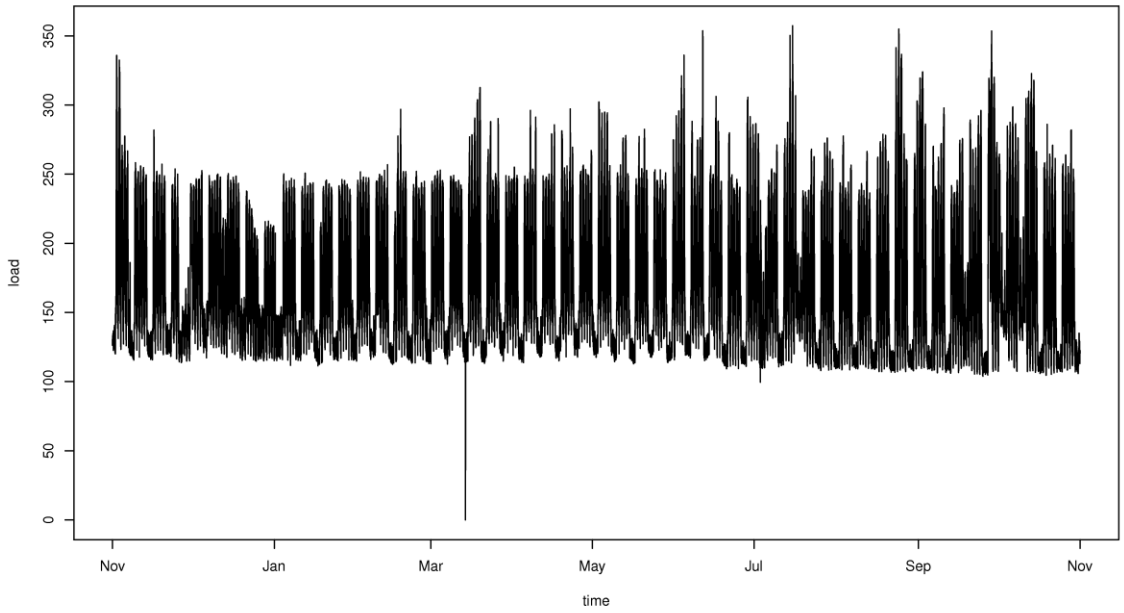


Figure A1-1: One year of load data from a building that shows "normal" behavior.

Figure A1-1 shows the load as a function of time for one of the buildings in the representative dataset. Each night the load falls to around 110 kW, and on most weekdays it peaks at about 250 kW. At this scale the weekends stand out as short intervals with low load. The two holiday weeks in December also stand out.

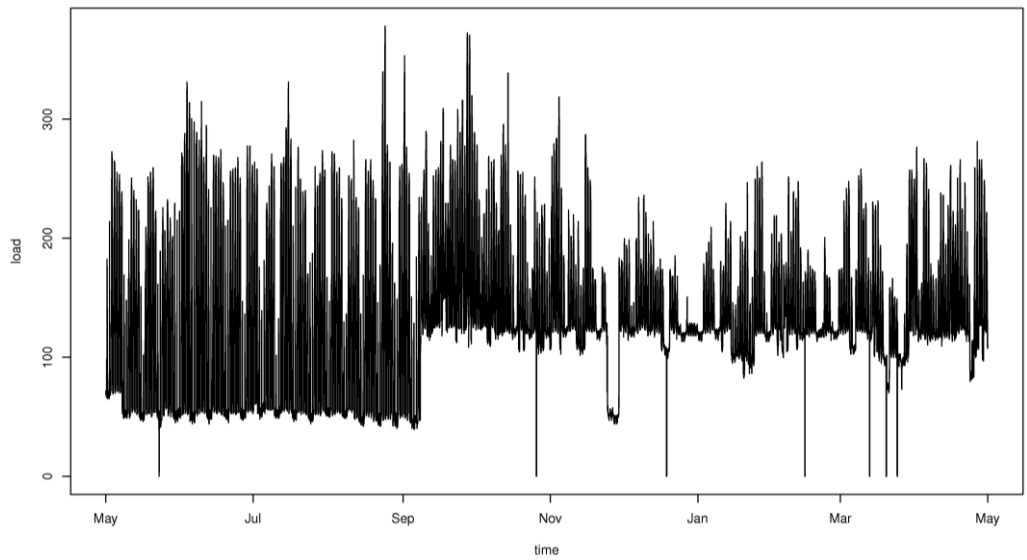


FIGURE A1-2: ONE YEAR OF LOAD DATA FROM A BUILDING WHOSE LOAD CHANGES SUBSTANTIALLY AFTER FOUR MONTHS.

Dramatic changes in load shape, such as occurred two thirds of the way through the year in the building shown in Figure A1-2, will usually lead to poor model predictions.

Figure A1-3 and Figure A1-4 show other cases in which the models being tested will probably perform poorly.

Cases such as these are within the scope of this study, so these buildings and others like them remained in the database. However, as we discuss later, we will sometimes apply a screen to remove buildings whose year-to-year total energy consumption changes substantially, and that screen will remove some buildings like this one from the test database.

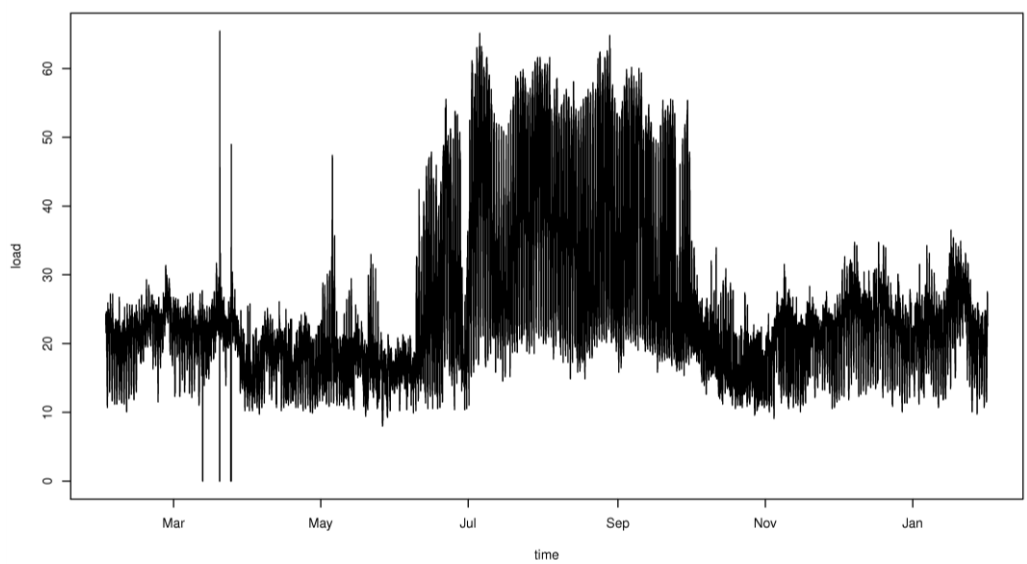


FIGURE A1-3: LOAD DATA FOR A BUILDING THAT HAS VERY HIGH SUMMERTIME LOADS, BEYOND WHAT IS PREDICTABLE FROM HIGH SUMMER TEMPERATURES.

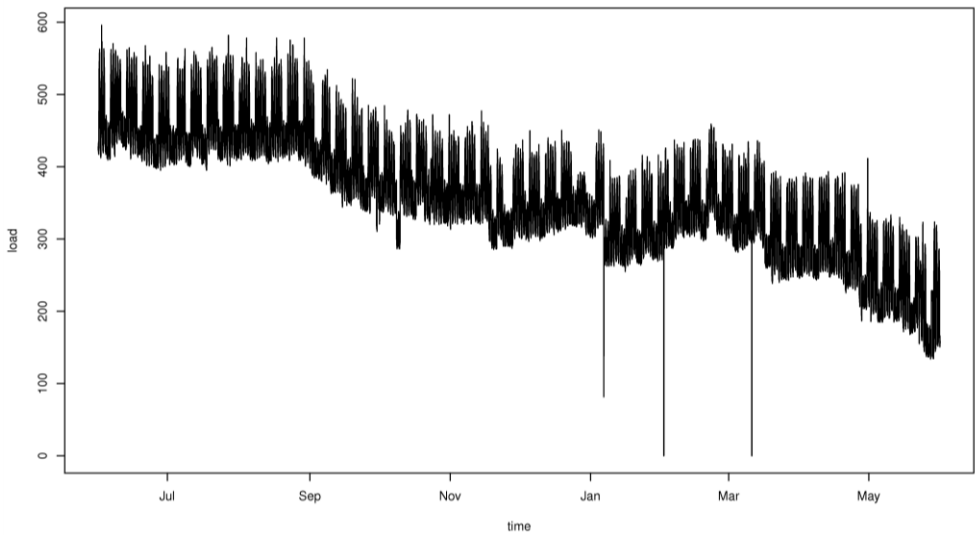


FIGURE A1-4: LOAD DATA FOR A BUILDING WITH A STRONG DOWNWARD TREND IN LOAD.

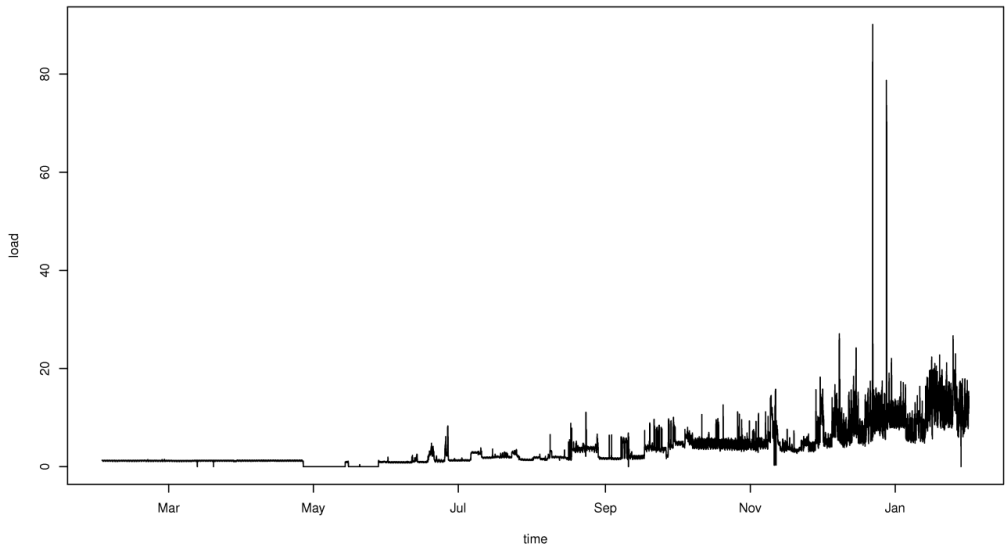


FIGURE A1-5: LOAD DATA FOR A BUILDING THAT WAS APPARENTLY UNOCCUPIED FOR MOST OF THE YEAR.

Figure A1-5 shows load data from a building that was apparently unoccupied for most of the year. Buildings such as this would not be likely be considered as participants in an ECM project. We plotted all buildings in both the Volunteer and Representative datasets, and they were visually inspected these energy use behaviors. Buildings with data quality issues, or which appeared to have been unoccupied for a portion of the time period, were removed from the database. Cases in which the data seem reliable but the building load shape changed substantially, such as Figures A1-1 through Figure A1-4, were retained.

WEATHER DATA

Outdoor air temperature (OAT) was obtained from the company Weather Underground. Temperature data were downloaded from five weather stations from each zip code that includes a building in the dataset. Temperatures from each of the stations were interpolated to the 15-minute intervals corresponding to the load data. At each time interval the median of five OAT values was calculated, and this is the temperature we use for this analysis. For about 20% of buildings there are fewer than five weather stations in the zip code; in these cases we used whatever data were available. For some of these zip codes, the few stations in the zip code have problems with bad or missing temperature values. In these cases we used temperature data from the nearest zip code that included another building in the dataset. In almost all cases this resulted in using data from a zip code within ten miles, but in about 3% of buildings the data were from ten to thirty miles away. In a few buildings we had no temperature data from within thirty miles; these buildings were excluded from the analysis.

APPENDIX 2: MODELS

MEAN WEEK

The prediction for each time of day, for each day of the week, is equal to the mean for that time of day and day of week in the training dataset. For example, the prediction for every Monday at 12:30 PM is the average load for all of the Mondays at 12:30 PM in the training data.

In this project, the quantities of interest are monthly or annual load predictions, which are generated by aggregating the Mean Week predictions for the desired time intervals. Any two months with the same number of days will therefore have very close to the same predicted energy usage: the only difference will be attributable to which days of the week are included. For instance, if one month has more weekend days and fewer weekdays than the other, and if there is a load difference between weekends and weekdays, then there will be a slight difference in the predicted energy use.

DAY-TIME-TEMPERATURE REGRESSION

In the Day-Time-Temperature model the predicted load is a sum of several terms: (1) a "day effect" that allows each day of the week to have a different predicted load; (2) an "hour effect" that allows each hour of the day to have a different predicted load; (3) an effect of temperature that is 0 for temperatures above 50F and is linear in temperature for temperatures below 50F; and (4) an effect of temperature that is 0 for temperatures below 65F and is linear in temperature for temperatures above 65F.

We define the following: i identifies the data point, day_i and $hour_i$ are the day and hour of that data point; TC_i = 0 if the temperature T exceeds 50 and is equal to 50 F -T if T < 50 F; TH=0 if T < 65 F and is equal to T-65 F if T > 65 F.

With these definitions, the model can be written as:

(A2-1)

The model is fit with ordinary regression. It can be thought of a variant of the ASHRAE five-parameter change-point model. Unlike an ASHRAE five-parameter change-point model, it has fixed points for the temperature slopes (at 50F and 65 F), and it adds time-of-day and day-of-week variation.

TIME-OF-WEEK-AND-TEMPERATURE REGRESSION

In the Time-of-Week-and-Temperature model, the predicted load is a sum of two terms: (1) a "time of week effect" that allows each time of the week to have a different predicted load from the others, and (2) a piecewise-continuous effect of temperature. The temperature effect is estimated separately for periods of the day with high and low load, to capture different temperature slopes for occupied and unoccupied building modes.

The model is described in Mathieu et al. (2011), but the determination of "occupied" and "unoccupied" periods is new to this project. For each day of the week, the 10th and 90th percentile of the load were calculated; call these L10 and L90. The first time of that day at which the load usually exceeds the L10 + 0.1*(L90-L10) is defined as the start of the "occupied" period for that day of the week, and the first time

at which it usually falls below that level later in the day is defined as the end of the “occupied” period for that day of the week.

Change Point Model

The Change-Point Model implements a six-parameter change-point model with the addition of a day-of-the-week effect.

ASHRAE Guideline 14 includes a 5-parameter change-point model, with the parameters being: the slope of the load-vs-temperature line for low temperatures, the slope of the line for high temperatures, the change point below which the temperature is low, the change point above which it is high, and the average load for temperatures that are neither low nor high. The ASHRAE model assumes that there is no relationship between temperature and load for temperatures that are above the low-temperature region but below the high-temperature region. In the present case, with months or even a year of data, there are enough data to estimate more parameters: (1) we also estimate a slope for intermediate temperatures, so this model has three slopes instead of two, and (2) at the suggestion of a member of the Technical Advisory Group, the model allows each day of the week to have a different average load in the intermediate-temperature region.

Cooling Degree-Days and Heating-Degree-Days Model

The Cooling-Degree-Day and Heating-Degree-Day (CDD-HDD) model is a model that was originally developed for analyzing monthly billing data. For each month the number of heating and cooling degree-days is calculated, and linear regression is performed to predict monthly energy usage as a function of CDD and HDD. CDD and HDD were defined with base temperatures of 55 F and 65 F, respectively.

With m identifying the month, the model can be expressed as:

(A2-2)

APPENDIX 3: ANALYSIS OF THE VOLUNTEER DATASET

ONE YEAR TRAINING PERIOD, ONE YEAR PREDICTION PERIOD (12:12)

To perform a year of training and a year of prediction it is necessary to have at least two consecutive years that are uninterrupted by retrofits. From the original dataset, 55 buildings met the requirement. If buildings had more than the minimum two years, we used the first two years, then stepped forward a year and used the following year, and so on. This yielded a total of 100 building-years that could be used for training.

Throughout this report we focus on two measures of model fit: (1) "Percent Bias" and (2) the monthly mean absolute percent error ("monthly MAPE").

Percent bias is the difference between the predicted and actual total energy usage over the entire prediction period. A percent bias of -5 means the predicted energy use in the building was 5% too low. It is often useful to take the absolute value of the percent bias for purposes of summarizing model performance; for example, if a model were to have a bias of +5% on half the buildings and -5% on the other half, its mean percent bias would be 0, but its mean absolute percent bias would be 5. The word "bias" in this case does not imply that a statistical model has a tendency to over- or under-estimate the energy: it is merely convenient shorthand for "error in the predicted total energy used."

Monthly MAPE is the error in the prediction for each month. We calculate the percent error for each month and summarize these errors by using the monthly MAPE. If six months are over-predicted by 10% and six are under-predicted by 10%, then the monthly MAPE is 10.

Table A3-1 summarizes the bias performance of the models for all of the buildings and all of the building-years. For each building, the first year of training starts with the first full month in the database and the first predicted year starts twelve months later; if the amount of data permits, the training period is shifted forward one year and the procedure is repeated.

TABLE A3-1: QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE VOLUNTEER DATASET, BY MODEL. 12 MONTH TRAINING PERIOD, 12 MONTH PREDICTION PERIOD. EXAMPLE: IN 50% OF THE CASES, THE "MEAN WEEK" MODEL HAD AN ABSOLUTE BIAS OF 4.75% OR LESS. LOWER NUMBERS ARE BETTER.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.56	2.19	4.58	12.33	23.01	9.15
Monthly CDD and HDD	0.70	1.55	4.66	12.57	23.11	9.03
Day, Time, and Temperature	0.52	1.88	4.69	12.12	22.23	8.92
Day and Change Point	0.44	1.77	4.88	12.18	21.69	8.90
Time of Week and Temperature	0.51	1.89	4.69	12.11	22.23	8.90

Table A3-1 shows that the difference between models is very small. For this set of buildings and years of data the median absolute bias is between 5.2-5.6% for all of the models. As comparison, the difference between the 25th-percentile and 75th-percentile of the buildings and years of data is approximately 10%.

The median absolute percent bias is around 5% for all of the models, and the mean is around 8.5%; there are a few buildings for which the predictions are extremely poor, and these lead to the average bias being worse than the median.

Essentially the only difference between all of the model's predictions is how (or whether) they adjust for outdoor temperature. And yet, the Mean Week model, which predicts next year's total energy use to be the same as this year, with no weather adjustment, performs as well as the others for buildings at the median predictability.

In this dataset, the Day-Time-Temperature model and the Time-of-week-Temperature model performed slightly better than the others in the most predictable building-years (as seen in the first two columns of the table), but this small advantage is not evident in the less predictable building-years.

Results can be seen in more detail in Figure A3-1, which shows the full statistical distribution for each model (except for very poor fits that are off the right side of the plot). The similar behavior of all of the models is readily apparent.

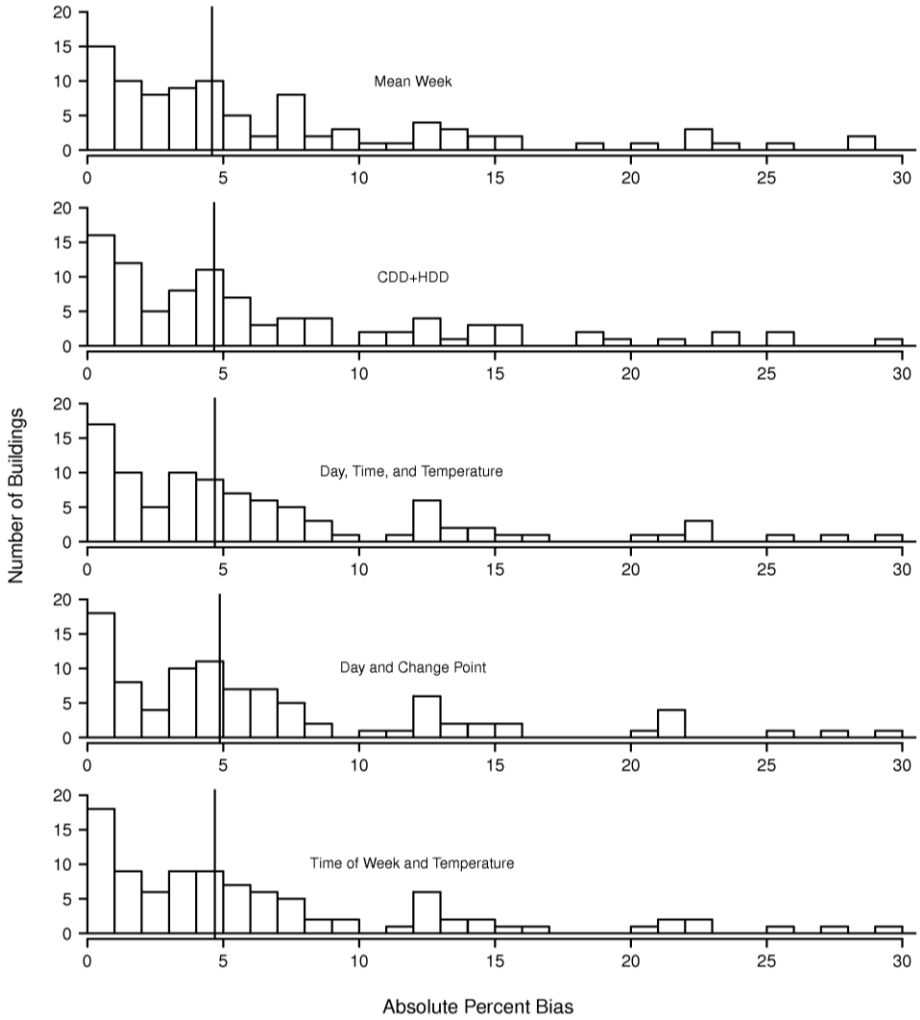


FIGURE A3-1: HISTOGRAMS SHOWING ABSOLUTE BIAS FOR EACH BUILDING-YEAR AND EACH MODEL. A FEW VERY POOR FITS ARE OFF THE RIGHT EDGE OF THE PLOTS. A VERTICAL LINE INDICATES THE MEDIAN FOR EACH MODEL.

The relative order of model performance changes from year to year, and depends on what metric one examines. In much of the work that follows we will reduce the number of plots and tables by showing results for just the Mean Week and/or Time-of-Week-and-Temperature models. The Time-of-week-and-temperature model behaves very similarly to the other models that make use of the interval temperature data for predictions, which are the day-time-temperature model and the day-and-change-point model.

TABLE A3-2: QUANTILES OF ABSOLUTE PERCENT BIAS FOR EACH MODEL. YEAR REFERS TO THE START OF THE TRAINING PERIOD

Model	Year	10%	25%	50%	75%	90%	Mean
Mean Week	2008	0.95	2.25	3.25	9.51	22.53	9.65
Monthly CDD and HDD		1.57	2.82	4.14	11.4	22.70	10.04
Day, Time, and Temperature		1.08	2.86	4.33	7.8	22.47	9.63
Day and Change Point		0.61	3.02	4.19	7.57	22.61	9.56
Time of Week and Temperature		1.05	2.73	4.32	7.81	22.46	9.61
Mean Week	2009	0.84	1.41	4.52	5.16	7.66	4.32
Monthly CDD and HDD		0.42	1.32	3.82	5.05	5.45	3.42
Day, Time, and Temperature		0.47	1.27	3.86	5.11	7.40	3.69
Day and Change Point		0.21	1.22	4.05	5.77	7.25	3.81
Time of Week and Temperature		0.41	1.28	3.87	5.12	7.36	3.70
Mean Week	2010	0.13	1.44	3.80	7.98	13.78	6.06
Monthly CDD and HDD		0.51	0.89	2.44	7.26	13.43	5.24
Day, Time, and Temperature		0.49	1.26	3.34	7.61	13.82	5.85
Day and Change Point		0.47	0.90	3.89	7.74	13.91	6.11
Time of Week and Temperature		0.53	1.27	3.34	7.66	13.85	5.87

SCREENING METRICS

BIAS AS A FUNCTION OF NAICS CODES

In the Volunteer dataset, there are only three NAICS prefixes that are represented by at least 8 building-years; stochastic variability is such that we think it is not worthwhile to try to make even preliminary conclusions based on smaller datasets.

Table A3-3 summarizes the absolute bias in predictions from the time-of-week-and-temperature model and the mean week model, respectively, for each of the three NAICS prefixes. The number of buildings represented is so sparse that we obviously cannot draw any firm conclusions.

TABLE A3-3: QUANTILES OF ABSOLUTE BIAS IN TIME-OF-WEEK-AND-TEMPERATURE MODEL PREDICTIONS, IN BUILDINGS WITH DIFFERENT NAICS CODE PREFIXES.

NAICS prefix	Description	Bldgs	Bldg-years	10th %ile	25th %ile	50th %ile	75th %ile	95th %ile
53	Real estate leasing	16	32	0.43	2.48	5.95	13.07	22.18
54	Professional/ Scientific	3	8	2.63	5.01	9.52	20.0	28.8
92	Public Administration	6	10	0.52	0.82	2.87	4.1	8.61

Bias as a Function of Building Energy Consumption

Figure A3-1 shows the bias in the time-of-week-and-temperature model prediction as a function of the building's total energy consumption, for each of the 99 building-years. Gray lines at +/- 5.5% contain half of the points. The most poorly-predicted buildings are mostly towards far the left side of the plot, but the relationship between bias (or absolute bias) and energy consumption is very weak.

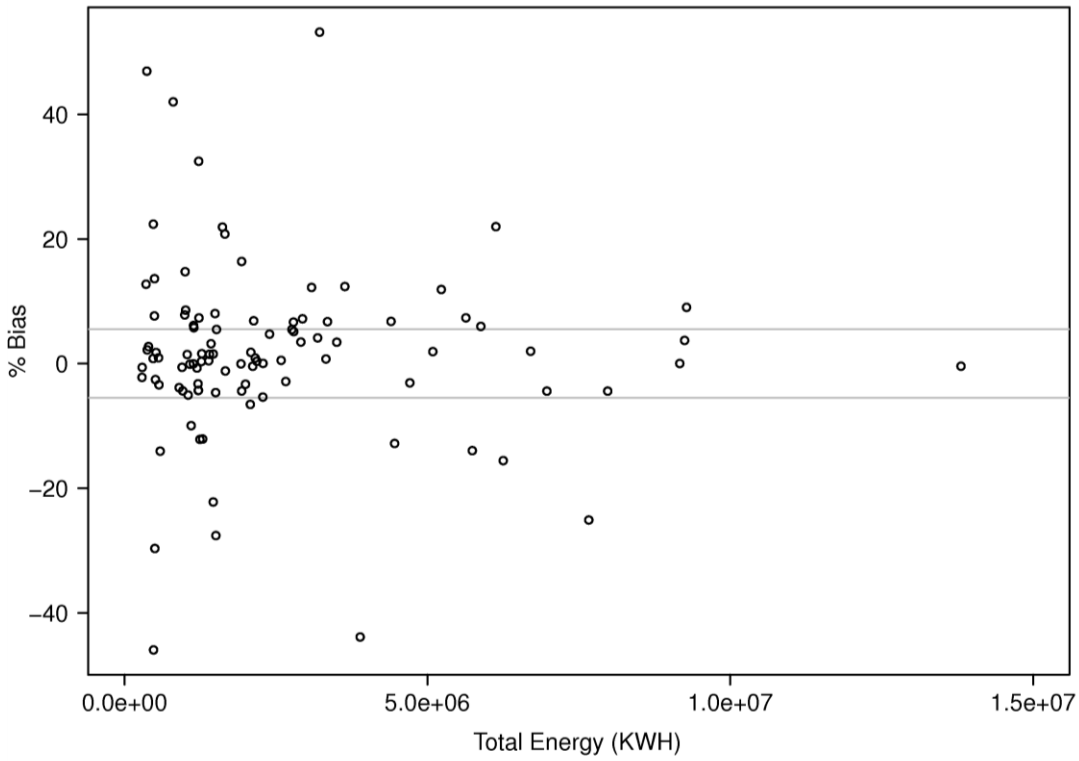


Figure A3-1: Percent Bias vs Total Annual Energy Consumption.

Bias as a Function of the Goodness of Fit During Training Period

We have already established that many buildings (or at least building-years) are not well fit by any of the models. Here we investigate whether it is possible to identify in advance – that is, using only the training data – whether the baseline in the prediction period is likely to be accurate.

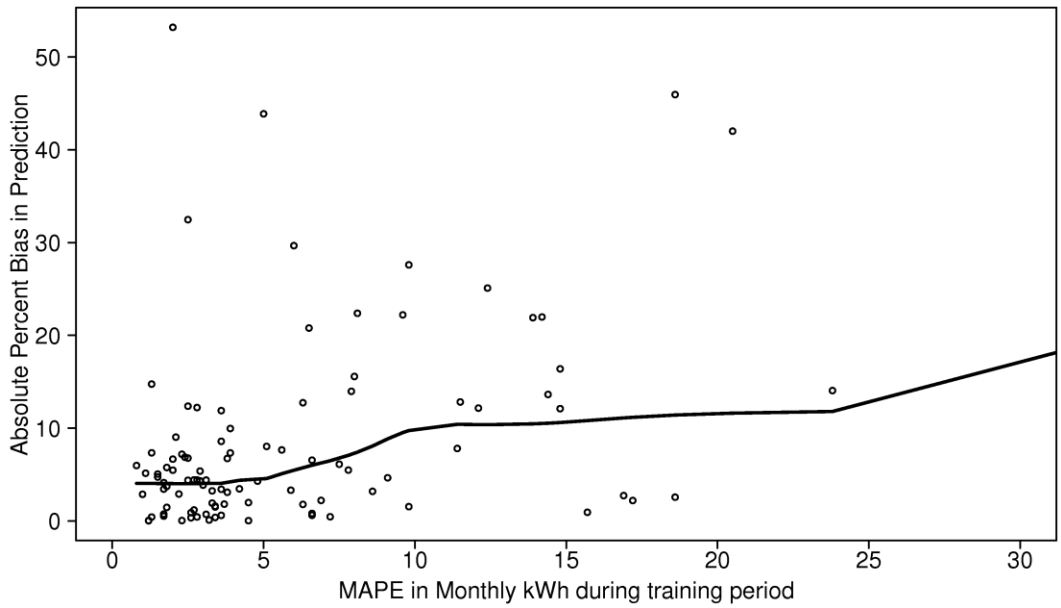


FIGURE A3-2: ABSOLUTE BIAS IN THE PREDICTION VS MEAN ABSOLUTE PERCENT ERROR IN THE MONTHLY FITS DURING THE TRAINING PERIOD, FOR THE TIME-OF-WEEK-AND-TEMPERATURE MODEL APPLIED TO THE VOLUNTEER DATA. A LOWESS CURVE (A TYPE OF LOCAL REGRESSION) IS SUPERIMPOSED.

All of the models that we tested are designed to have no "net determination bias" during the training period; that is, the sum of the predicted loads over the entire training period is equal to the sum of the actual loads. However, the sum of the predictions over a shorter time period such as a month can, and does, differ from the sum of the actual loads: some months the prediction is too high, and some months it is too low.

We speculate that if a model fits well in every month during the training period, it may continue to fit well for the prediction period, whereas if some training months are over-predicted and some are under-predicted then it may be a matter of chance whether the over- and under-predictions will cancel out in the prediction period.

Figure A3-2 shows the absolute percent bias in the prediction as a function of the mean absolute percent error in the monthly fits during the training period, for the time-of-week-and-temperature model. (Results for the other models that use interval temperature data are very similar). A "lowess" curve – a locally weighted polynomial regression (Cleveland 1981) – is superimposed; the curve is not intended to represent an intrinsic relationship between the variables, but it is useful for seeing the trend. Results are in line with expectation: if the predictions during the training period were accurate for each individual month (points toward the left side of the plot) then the bias for the prediction period tended to be low, whereas high error in the individual months was usually associated with higher bias.

BIAS AS A FUNCTION OF LOAD VARIABILITY DURING THE PREDICTION PERIOD

The "load variability" metric quantifies the amount of variation at a given time of the week throughout the year. For instance, if Monday at 12 PM sometimes has very high load and sometimes has low load, that will contribute to a large load variability number.

The value of a load variability metric during the training period was calculated for each building. The metric quantifies the root-mean-squared variation in load at a given time of week, as a percentage of the mean load. Buildings with high load variability are neither more nor less predictable than those with low load variability. For example, the correlation between load variability and absolute percent bias of the LBNL regression model is only r=0.06 (r^2 = 0.0036).

PORTFOLIO EFFECTS

The baseline prediction for any single building will over-estimate or under-estimate the energy used by the building. A portfolio of buildings will normally include some buildings for which the estimate is too high and others for which it is too low. As buildings are added to a portfolio, the total error is expected to increase more slowly than the total energy consumption, so the fractional error for a group of buildings is likely to be much lower than for a typical individual building.

We illustrate with two types of portfolios: (1) just buildings with the NAICS prefixes listed in Table A3-3, and (2) just buildings for which the MAPE of monthly load during the prediction period was less than 5. We consider only a single year at a time, and summarize results only if the portfolio contains at least 3 buildings.

TABLE A3-4: BIAS FOR PORTFOLIOS BASED ON NAICS PREFIXES, WITH PREDICTIONS FROM THE TIME-OF-WEEK-AND-TEMPERATURE MODEL. "YEAR" REFERS TO THE YEAR THE TRAINING PERIOD STARTED.

NAICS prefix	Year	Bldgs	Total kWh	Predicted kWh	Percent bias
53	2008	10	33,827,856	36,354,027	7.47
	2009	9	23,326,927	23,171,606	-0.67
	2010	6	9,734,264	9,380,903	-3.63
54	2008	3	10,709,621	8,010,115	-25.21
	2009	3	10,815,412	10,701,141	-1.06
92	2008	7	81,890,998	84,405,816	3.07
	2009	6	18,481,290	18,284,005	-1.07
	2010	5	8,385,607	8,317,940	-0.81

Table A3-4 shows the results for portfolios built around NAICS prefixes, using the time-of-week-and-temperature model. (Results for the other models that use interval temperature data are similar). Results are something of a mixed bag.

Prefix 92, composed of public administration buildings, yielded very accurate predictions for the total energy consumption, with errors under 1% in both years. Prefix 53 had very low bias in 2009, and bias about half of the percentage of a typical individual building in 2010. But the bias of the portfolio in 2008 was, as a percentage of the total, about as large as the bias in a typical building in that NAICS code, although much smaller than the mean bias in the those buildings. And in Prefix 54, the inclusion of a very mis-predicted building degraded the performance of the whole portfolio in 2008 (which, however, consists of only three buildings).

TABLE A3-5: BIAS FOR A PORTFOLIO BASED ON BUILDINGS FOR WHICH THE FIT WAS GOOD DURING THE TRAINING PERIOD (MONTHLY MAPE < 5%), WITH PREDICTIONS FROM THE TIME-OF-WEEK-AND-TEMPERATURE MODEL. "YEAR" REFERS TO THE YEAR THAT THE TRAINING PERIOD STARTED.

Year	Bldgs	Total kWh	Predicted kWh	Percent bias
2008	24	138,745,309	144,205,860	3.94
2009	19	58,264,307	58,984,509	1.24
2010	17	49,660,066	51,843,759	4.40

Table A3-5 shows the results for a portfolio composed of buildings with low Mean Absolute Percent Error of the monthly predictions during the training period (Monthly MAPE < 5%). Even with this fairly large collection of well-behaved buildings, the error in the total energy use exceeds 3 % in two of the three years.

APPENDIX 4: ANALYSIS OF THE REPRESENTATIVE DATA

ONE YEAR TRAINING PERIOD, ONE YEAR PREDICTION PERIOD (12:12)

As with the Volunteer dataset, we fit all of the models using one year of training data, and making a prediction for the following year. Unlike the Volunteer data, we cannot step forward year by year since the Representative dataset contains only two years of data (from 2010 and 2011).

TABLE A4-1: QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE 389 BUILDINGS IN THE REPRESENTATIVE DATASET, BY MODEL. 12 MONTH TRAINING PERIOD, 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.82	2.21	4.82	9.63	19.42	8.40
Monthly CDD and HDD	0.69	2.09	4.53	10.03	19.38	8.46
Day, Time, and Temperature	0.69	2.17	4.51	9.26	19.41	8.42
Day and Change Point	0.73	2.02	4.70	9.22	18.84	8.24
Time of Week and Temperature	0.82	2.21	4.82	9.63	19.42	8.40

Considering that the Volunteer dataset contains a different mix of buildings – we will look at the distribution of building types, below – and that there are often systematic differences between volunteer data and random-sample data, the similarity of the raw results in Table A3-1 and Table A4-17 is striking.

Results for the Mean Absolute Percent Error in the monthly predictions are shown in Table A4-2. Histograms of absolute mean bias are shown in Figure A4-1.

TABLE A4-2: QUANTILES AND MEAN OF MONTHLY MEAN ABSOLUTE PERCENT ERROR (MAPE) FOR THE REPRESENTATIVE DATASET, BY MODEL. 12 MONTH TRAINING PERIOD, 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	5.70	8.80	13.60	22.40	36.80	19.70
Monthly CDD and HDD	4.08	5.40	8.70	16.10	30.02	14.53
Day, Time, and Temperature	3.15	5.00	8.15	15.17	29.85	14.01
Day and Change Point	4.20	6.30	10.10	17.70	31.50	15.65
Time of Week and Temperature	3.20	4.80	8.00	15.30	29.82	13.91

As with the Volunteer data, we investigate whether there is a way to know from the training period whether the fit to the prediction period is likely to be good.

We calculated a "load variability" metric as defined in Coughlin et al.(2009), although we applied it at the 15-minute timescale of the available data rather than the one-hour timescale used by Coughlin. Load Variability quantifies the extent to which some 15-minute intervals during the training interval are higher or lower than the average. A screen based on load variability did not effectively distinguish predictable from unpredictable buildings in this dataset.

As with the Volunteer dataset, we also investigated whether the goodness of fit during the training period is predictive of how well the model will fit in the prediction period. As with the Volunteer dataset there is a useful relationship but it is not extremely strong. Figure A4-2 shows the absolute percent bias

as a function of the Mean Absolute Percent Error in the monthly fit during the training period. Buildings with lower MAPE in the training period do tend to have lower bias in the prediction. However, a low MAPE does not come close to guaranteeing a low bias: even setting a screen at MAPE < 5 will include many buildings that have a mean absolute bias during the prediction period of 10% or more.

The range of NAICS codes in the Representative dataset is substantially different from the Volunteer dataset. For example, the Volunteer dataset includes 51 buildings, of which 6 are "Public Administration"; the much larger Representative dataset, with almost 400 buildings, contains only 7 Public Administration buildings.

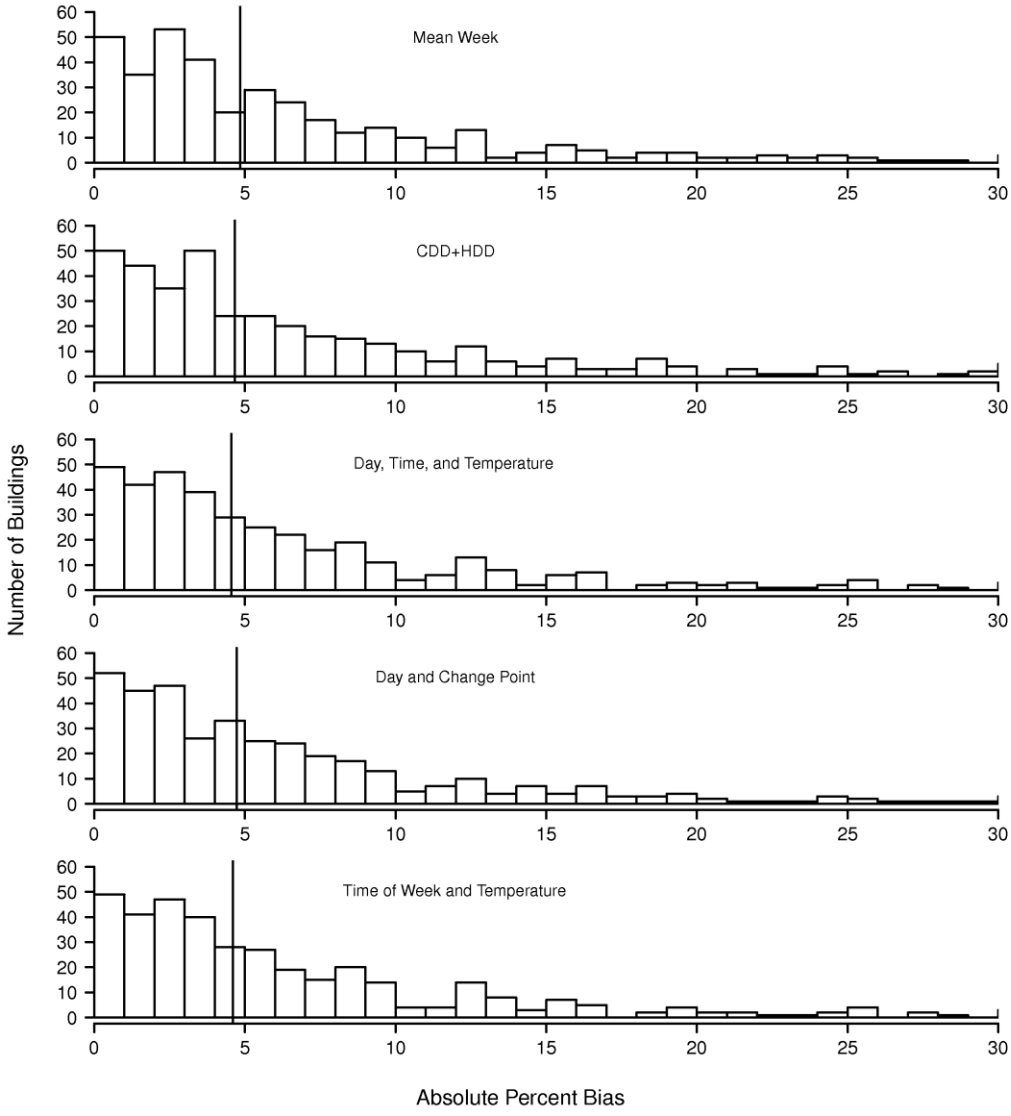


FIGURE A4-1: HISTOGRAMS SHOWING ABSOLUTE BIAS FOR EACH BUILDING-YEAR AND EACH MODEL. SOME CASES FALL OFF THE RIGHT SIDE OF THE PLOT. A VERTICAL LINE INDICATES THE MEDIAN FOR EACH MODEL.

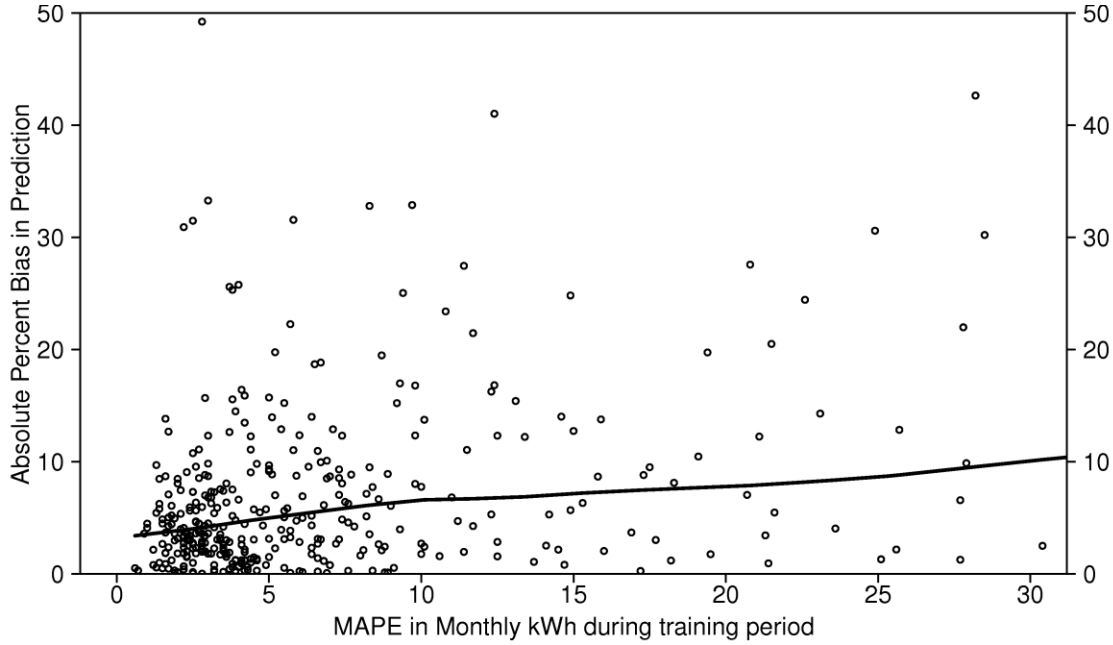


FIGURE A4-2: ABSOLUTE BIAS IN THE PREDICTION VS MEAN ABSOLUTE PERCENT ERROR IN THE MONTHLY FITS DURING THE TRAINING PERIOD, FOR THE TIME-OF-WEEK-AND-TEMPERATURE MODEL APPLIED TO THE REPRESENTATIVE DATA. SOME POINTS FALL OFF THE PLOT. A LOWESS CURVE (A TYPE OF LOCAL REGRESSION) IS SUPERIMPOSED.

TABLE A4-3: QUANTILES OF ABSOLUTE PERCENT BIAS FOR BUILDINGS WHOSE MONTHLY MAPE IN THE TRAINING PERIOD WAS LESS THAN 5%.

Model	N	10%	25%	50%	75%	90%	Mean
Mean Week	62	3.44	4.53	5.60	7.85	9.98	7.38
Monthly CDD and HDD	209	3.68	4.50	6.10	8.70	12.82	8.16
Day, Time, and Temperature	202	2.70	3.58	5.40	7.78	12.08	7.03
Time of Week and Temperature	200	2.70	3.58	5.30	7.62	11.28	6.91

The fact that the bias in the prediction period is related (albeit weekly) to the monthly MAPE in the training period suggests the possibility of screening for buildings that are likely to be predicted more accurately.

The table below shows quantiles of absolute bias in Time-of-Week-and-Temperature model predictions, in buildings with different NAICS code prefixes summarizes the bias for each of the building types that is represented by at least 10 buildings in the representative database. We also included public administration buildings because analysis of the Volunteer dataset suggested that those buildings might be especially predictable: Table A3-3 showed them to have low mean absolute bias, i.e. to be highly predictable on average. This is not the case in the representative dataset, where these buildings are approximately as predictable or unpredictable as other building types.

TABLE A4-4: QUANTILES OF ABSOLUTE BIAS IN TIME-OF-WEEK-AND-TEMPERATURE MODEL PREDICTIONS, IN BUILDINGS WITH DIFFERENT NAICS CODE PREFIXES. COMPARE TO TABLE A3-3.

NAICS prefix	Description	Bldgs	10th %ile	25th %ile	50th %ile	75th %ile	95th %ile
42	Wholesale trade	14	1.69	3.41	7.47	13.66	22.1
44	Retail trade	41	0.47	2.36	4.20	8.69	15.2
45	Retail trade	12	1.08	2.24	3.17	5.70	13.3
49	Transport/Warehousing	10	2.74	3.14	5.50	11.66	16.4
51	Information	15	0.48	1.06	2.67	8.08	21.4
53	Real Estate Leasing	53	1.85	3.68	5.93	16.82	51.5
61	Education	42	0.84	1.90	3.58	7.86	15.8
62	Health and Social Care	36	0.55	1.50	3.21	5.93	9.8
71	Arts, Entertainment	30	1.25	3.26	6.80	18.22	30.3
72	Accommodation/food	63	0.64	1.87	3.85	8.70	12.4
81	Other	32	1.35	3.25	7.26	12.77	23.3
92	Public Administration	7	2.60	3.25	3.99	7.69	56.1

PORTFOLIO EFFECTS

We quantified the error, or bias, in the prediction of the total annual energy used in various portfolios of buildings. We used the time-of-week-and-temperature model for these tables, but results are quite similar for the day-time-temperature model and the change-point-and-day model.

The overall messages of the following tables are: (1) as expected, larger portfolios of buildings lead to lower percent bias in the total energy used, and (2) screening out buildings that change dramatically from year to year, and including only buildings for which the models fit well during the training period, both tend to lead to better performance (lower percent bias) but this effect is not extremely large.

TABLE A4-5: BIAS FOR PORTFOLIOS BASED ON NAICS PREFIXES, WITH PREDICTIONS FROM THE TIME-OF-WEEK-AND-TEMPERATURE MODEL. ONLY PREFIXES WITH AT LEAST 10 CASES ARE INCLUDED, PLUS PREFIX 92 FOR COMPARISON TO THE VOLUNTEER DATA. THE YEAR 2010 WAS THE TRAINING YEAR FOR ALL MODELS.

NAICS prefix	Bldgs	Total kWh	Predicted kWh	Percent bias
42	14	7,844,788	7,696,758	-1.89
44	41	29,935,698	30,370,868	1.45
45	12	7,320,698	7,358,519	0.52
49	10	5,720,874	5,591,634	-2.26
51	15	13,770,148	13,601,572	-1.22
53	53	37,462,843	41,062,271	9.61
61	42	16,88,7745	17,403,489	3.05
62	36	20,238,549	21,001,653	3.77
71	30	7,430,195	7,573,492	1.93
72	63	23,302,962	22,971,386	-1.42
81	32	7,303,410	7,447,883	1.98
92	6	5,127,729	5,215,852	1.72

TABLE A4-6: BIAS FOR A PORTFOLIO BASED ON BUILDINGS FOR WHICH THE FIT WAS GOOD DURING THE TRAINING PERIOD (MONTHLY MAPE < A SPECIFIED THRESHOLD), WITH PREDICTIONS FROM THE TIME-OF-WEEK-AND-TEMPERATURE MODEL. TRAINING YEAR WAS 2010 FOR ALL BUILDINGS.

MAPE threshold	Bldgs	Total kWh	Predicted kWh	Percent bias
10%	299	192,780,964	195,266,556	1.29
9%	288	189,256,859	191,507,069	1.19
8%	271	181,761,057	183,929,834	1.19
7%	254	173,359,941	175,784,108	1.40
6%	228	159,519,215	162,114,653	1.63
5%	200	146,771,988	148,758,101	1.35
4%	165	127,162,839	129,505,457	1.84
3%	110	88,809,120	90,545,326	1.95
2%	40	43,211,073	43,751,227	1.25
1%	3	5,189,497	5,250,785	1.18

SIX-MONTH TRAINING, 12-MONTH PREDICTION PERIOD (6:12)

A long training period (such as a year) provides more information about a building's average behavior than does a shorter period, but this comes at a price: a long training period also increases the chance that the building's load will change substantially during the training period. Data from 8 or 9 months ago will degrade rather than improve the performance of a baseline model if the building has changed in the past 8 or 9 months.

Table A4-7 summarizes the performance of each of the baseline models in the case of a 6-month training period and a 12-month prediction period. The training period began in February, 2010 for almost all of the buildings, but in other months for some buildings that did not have data from early 2010. Comparing Table A4-7 to Table 1, which is the counterpart for the 12:12 case, we see that the 10th percentile of absolute mean bias is actually lower with only 6 months of training, and the 25th percentiles are essentially the same for the 6:12 and 12:12 cases. The performance for the median building is slightly worse in the 6:12 case than in the 12:12 case, and the upper (less well predicted) quantiles are much worse in the 6:12 case.

Surprisingly, even the "Mean Week" model, which assumes that the average load over the next year will be the same that it was during the training period, performs about as well as any of the other models.

TABLE A4-7: QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE REPRESENTATIVE DATASET, BY MODEL. 6 MONTH TRAINING PERIOD STARTING IN FEBRUARY 2010, 12 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.02	2.19	5.39	11.60	108.1	9.37
Monthly CDD and HDD	0.08	2.73	6.31	15.23	242.0	13.1
Day, Time, and Temperature	0.02	2.11	5.34	10.53	110.6	9.19
Day and Change Point	0.00	2.42	5.94	11.42	107.9	9.66
Time of Week and Temperature	0.01	2.19	5.00	10.44	110.8	9.09

TABLE A4-8: QUANTILES AND MEAN ABSOLUTE PERCENT BIAS FOR EACH MODEL, ONLY FOR THE 295 BUILDINGS WHOSE YEAR-TO-YEAR CHANGE IN MEAN POWER WAS LESS THAN 10%.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.65	1.48	3.70	6.74	12.07	5.67
Monthly CDD and HDD	0.91	2.27	4.35	8.33	15.53	8.25
Day, Time, and Temperature	0.70	1.76	3.69	7.05	10.27	5.56
Day and Change Point	0.63	1.83	4.05	7.50	11.43	6.11
Time of Week and Temperature	0.66	1.71	3.52	7.06	10.22	5.48

TABLE A4-10: QUANTILES AND MEAN OF THE MEAN ABSOLUTE PERCENT ERROR (MAPE) IN MONTHLY ENERGY USE FOR EACH MODEL, ONLY FOR THE 295 BUILDINGS WHOSE YEAR-TO-YEAR CHANGE IN MEAN POWER WAS LESS THAN 10%.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	5.42	8.40	13.50	21.60	34.18	15.8
Monthly CDD and HDD	4.72	6.30	11.50	20.90	43.56	18.0
Day, Time, and Temperature	3.20	4.80	8.55	16.75	28.48	16.1
Day and Change Point	4.52	7.40	11.90	21.10	35.04	24.9
Time of Week and Temperature	3.20	4.80	8.45	16.43	28.46	32.2

THREE-MONTH TRAINING, 6-MONTH PREDICTION PERIOD (3:6)

Currently, guidelines for measurement and verification of energy savings recommend, or in some cases require, 12 months of pre-retrofit energy data and 12 months of post-retrofit data. These requirements impose substantial burdens on participants and funders of retrofit programs, since they result in long delays in contract settlements. Also, as discussed previously, many buildings occasionally change their power usage dramatically, and the longer the baseline and prediction periods the more likely they are to include such an event.

TABLE A4-11: QUANTILES AND MEAN OF ABSOLUTE PERCENT BIAS FOR THE REPRESENTATIVE DATASET, BY MODEL. 3 MONTH TRAINING PERIOD (USUALLY STARTING IN FEBRUARY 2010), 6 MONTH PREDICTION PERIOD.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	2.63	6.56	13.91	23.63	205.8	18.6
Monthly CDD and HDD	8.44	20.96	47.51	94.02	381.3	72.2
Day, Time, and Temperature	1.35	3.27	7.49	16.21	180.1	13.9
Day and Change Point	1.80	5.28	13.93	29.76	252.0	24.2
Time of Week and Temperature	1.27	3.32	7.27	16.21	179.2	13.9

Three-month training, 12-month prediction period (3:12)

We now consider a short training period and a long prediction period.

Table A4-12: Quantiles and Mean of Absolute Percent Bias for the Representative Dataset, by Model. 3 Month Training Period (usually starting in February 2010), 12 Month Prediction Period.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	1.75	4.00	8.87	15.93	184.2	13.0
Monthly CDD and HDD	4.73	14.33	32.37	62.65	147.2	55.6
Day, Time, and Temperature	1.26	2.46	6.07	12.52	25.99	11.9
Day and Change Point	1.47	3.98	8.24	17.13	32.28	14.6
Time of Week and Temperature	1.15	2.48	6.12	12.75	26.1	11.9

Table A4-13: Quantiles and Mean of Absolute Percent Bias for the Representative Dataset, by Model. 3 Month Training Period (usually starting in February 2010), 12 Month Prediction Period, only for the 295 buildings whose mean power from 2010 to 2011 changed less than 10%.

Model	10%	25%	50%	75%	90%	Mean
Mean Week	1.75	3.85	8.38	14.44	21.61	11.2
Monthly CDD and HDD	4.43	12.13	29.32	54.40	90.88	45.7
Day, Time, and Temperature	1.13	2.05	4.91	10.50	18.18	8.8
Day and Change Point	1.30	3.42	6.83	13.88	23.04	11.4
Time of Week and Temperature	1.01	1.94	4.96	9.77	18.7	8.7

APPENDIX 5: EVALUATING MODELS

The main quantity of interest for quantifying the effectiveness of energy conservation measures (ECMs) is the reduction in the total energy used by the building over a long period such as a year. Estimating this reduction requires comparing the amount of energy the building actually used to the amount it would have used if the ECMs had not been implemented. The actual energy use can be directly measured, but the energy the building would have used cannot be directly measured so it must be estimated with a baseline model.

Conventionally, a baseline estimate is based on analyzing the building's load during the year before the ECM is implemented, and the estimate is compared to the amount of energy the building uses during the following year. (ASHRAE, 2002).

The most directly relevant question about the accuracy of the baseline model is "how well is the total energy consumption predicted over the assessment period?" In the present study, the answer is provided by the "bias": a model is 'trained' data from a 'training period' (3, 6, or 12 months) and the model is used to predict the energy used during the 'prediction period.' This boils the accuracy of the prediction down to a single number for each building, or, rather, a single number for a single combination of (building, start date of training period, duration of training period, duration of prediction period).

A less direct measure of model fit is the Mean Absolute Percent Error (MAPE) in the energy used each month. If two models are equally accurate at predicting the total energy used over the course of a year, but one is better than the other at predicting the energy used in each individual month, the latter should be preferred. One reason for preferring the model that performs better with the individual months is that it may allow assessment of energy efficiency measures on shorter timescales. Another is that a model that normally works well for individual months may help identify anomalous months that should be investigated to see if they should be excluded from the analysis (due to metering problems, short-term interruption in normal building operations, or other reasons).

We proceed according to the following principles:

1. For quantifying ECM performance, a model should perform no worse on average than existing simple regression models in a substantial majority of buildings (these are the Day-Time-and-Temperature model and the Time-of-Week-and-Temperature model in the current report). We will select one of the regression models as the "reference model" and score other models by comparing them to the reference model. *Just because a model outperforms the reference model does not mean it performs well enough to be used for a specific business case,* as we will discuss later.
    a. This should be true both for predicting total energy use and for predicting the energy use in each month.
    b. This should be true both for predicting the next year's energy use from the current year, and predicting the next year from the preceding six months.
2. No model can hope to predict the energy use, monthly or total, for a building whose behavior changes substantially for reasons that are not related to the explanatory variables available to the model. Although such occurrences are common in real-world applications, these cases should be excluded from the test dataset that is used to evaluate models.

The first principle is based on the idea that there is no reason to use an inferior model when there is a freely available alternative that performs better.

The second principle has important implications: it means the data used to evaluate the models may differ considerably from the data that the models will be called upon to analyze in real-world applications. The test dataset will have data only from well-behaved buildings – by which we mean buildings whose pattern of energy use does not change unpredictably by a large amount – but in many applications the models will be applied to some buildings that are not well-behaved. In actual use, model performance in some buildings will be worse (possibly much worse) than the performance when used for even the most unpredictable buildings in the test dataset, because the test dataset will exclude very poorly-behaved buildings.

There are statistical tests, such as the Kolmogorov-Smirnov test, that compare distributions, but no existing method or test matches our needs in this application. We believe the model evaluation procedure should meet the following criteria, which are one step forward in specificity from the principles enumerated above.

1. The model evaluation should depend on how well the model performs in terms of both bias and monthly MAPE.
2. The evaluation should depend on how well the model performs with both a year and six months of training data.
3. The evaluation should give more statistical weight to buildings that are more predictable. According to the second of the principles enunciated above, we try to exclude intrinsically unpredictable buildings from the test dataset, but the more variable of the buildings that remain will include some whose loads are not predictable from available data. The degree of fit to such buildings will be largely a matter of luck, so a model shouldn't be heavily penalized for doing slightly worse on these buildings.
4. The evaluation should quantify both the absolute performance of the model and the performance relative to an existing freely available model.
5. The procedure should be easy to understand and to apply.

We did not succeed in finding a set of screening criteria that can identify in advance a class of buildings that will yield accurate baseline predictions. The main problem is that some buildings change load behavior in a way that is inherently unpredictable, at least if temperature is the only available explanatory variable. The ten to fifteen percent of buildings that behave this way in a given year lead to a distribution of bias that has a long tail of bad performance. Such cases are not limited to certain building types or to buildings with certain types of load characteristics (such as high load variability during the training period, or poor ability to fit individual months during the training period.