Report by **SBW CONSULTING, INC.**

PG&E COMMERCIAL WHOLE BUILDING DEMONSTRATION EARLY M&V REPORT - DRAFT

Submitted to **PACIFIC GAS AND ELECTRIC COMPANY**

Submitted by **SBW CONSULTING, INC.**
2820 Northup Way, Suite 230
Bellevue, WA 98004

February 21, 2019

sbw
ENERGY + WATER + EFFICIENCY

Executive Summary

To gain an advanced understanding of challenges that may come with site-specific Normalized Metered Energy Consumption (NMEC)-based savings, PG&E contracted with SBW Consulting, Inc. to conduct early Measurement and Verification ("Early M&V") on its Commercial Whole Building (CWB) Demonstration (Demo). The Demo comprised 12 medium-sized commercial buildings: five grocery stores, six office buildings, and a library. All 12 sites had electric meters and all but one had gas meters. The participants implemented a variety of efficiency measures including retrofitting lighting and HVAC equipment and improving controls settings. The Demo required the buildings to achieve a minimum of 15% savings at the site level.

The Demo relied on International Performance Measurement and Verification Protocol (IPMVP) whole building approaches to estimate site-level savings, namely Options C and D (EVO 2016). The Option C approach prescribes regression analysis of building energy meter data with independent variables, such as weather. The Option D approach applies calibrated physics-based simulation of building energy use. Pacific Gas and Electric Company (PG&E) contracted with two vendors and a technical consultant for the Demo who each estimated whole building avoided energy use based on utility meter interval data with the Option C approach. The vendors used proprietary algorithms to model energy consumption and the technical consultant used two public domain algorithms, Time of Week and Temperature (TOWT) and Mean Week (MW). Furthermore, the contractors who implemented the efficiency measures at the Demo sites estimated normalized savings with the Option D approach, by developing physics-based models in eQUEST.

There were three key objectives of this study:

1. Establish whether savings can be reliably estimated from changes in metered energy use.
2. Provide recommendations for the best methodology to use for estimating CWB savings in the future.
3. Provide an assessment of the Demo's evaluability and recommendations for modifications to the program design to improve its evaluability prior to scaling the Demo into a full-fledged CWB program.

SBW accomplished these objectives with multiple approaches. First, we developed our own independent savings estimates for each of the 12 Demo sites, applying the IPMVP Option C approach in the Excel add-in Energy Charting and Metrics (ECAM). Next, we reviewed and compared the models and savings estimates developed for each site – a total of 123 models across 12 sites – and looked for reasons for differences between the estimates.

Additionally, SBW explored techniques to observe patterns in the interval energy use data that could indicate presence of trending or other non-routine events (NREs). Trending involves a pattern of increasing or decreasing use over the course of the baseline period which should be considered when estimating savings. NREs are changes in energy use in the building that are not due to the energy efficiency project and not accounted for in the independent variables of the regression model, such as removal of a data server room.

Finally, we examined the program materials with a focus on how well they were followed by the participating parties in the Demo. Taking into consideration all that we learned through the course of the study, we assessed how well the practices in place for the Demo support evaluation.

Key Findings

SBW found low uncertainty in its ECAM models relative to savings with the exception of gas models at three sites. **Our models indicated large electric savings (>15%) were achieved at all sites and large gas savings were achieved at five sites, all of which were non-grocery.**

SBW observed in its comparison analysis that 64 of the 88 Demo Option C estimates were within the model uncertainty intervals of the ECAM estimates. In other words, **73% of the time, the Demo Option C savings estimates were the same within the margin of error of our ECAM estimate.** Given how the energy efficiency industry has been plagued with reproducibility problems for custom savings calculations (i.e., engineering estimates), this consistency in savings estimates using different Option C models is remarkable.

The charts below (see Figures 1 and 2) present the savings estimates from each of the models at each site. Figure 1 shows the percent electric savings estimates.

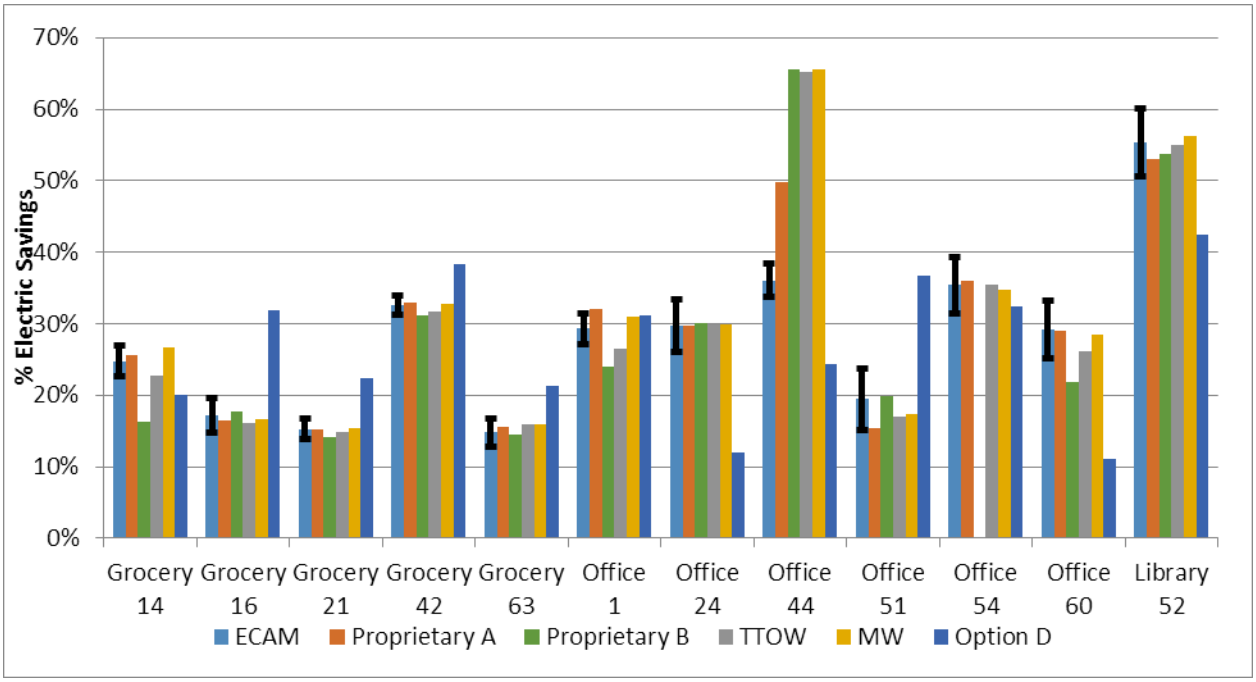


Figure 1: Percent Electric Savings Estimates

Figure 2 shows the percent gas savings estimates. We found all grocery sites and one non-grocery site had increased gas use during the reporting period. It was beyond the scope of this study to explore the reasons for the increased gas use at these sites.

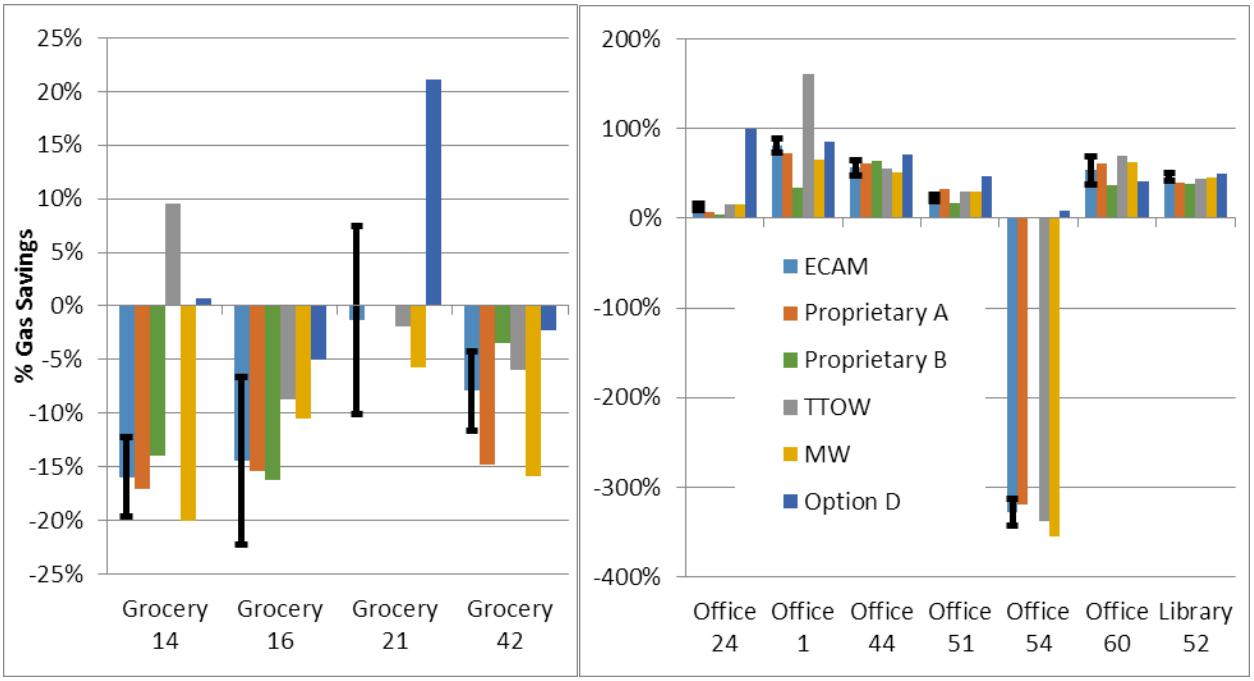


Figure 2: Percent Gas Savings Estimates

To inform how savings estimation can be improved, we examined the Demo Option C model inputs and outputs to explore reasons for differences among the 24 estimates that were significantly different. We found that most of the differences were due to differences in the models rather than in the input data, such as not accounting for day types (weekday vs. weekend) and letting the model predict negative energy use. **Notably, we observed that the proprietary algorithms did not offer advantages over the open-source, public-domain (free) algorithms for verifiable savings estimation, while the open-source algorithms offer the key evaluability advantage of transparency**. There may be other advantages to the proprietary algorithms that were beyond the scope of our study to consider.

Our review of the Demo Option C models did not reveal the modelers cleaned the interval data of erroneous data points nor adjusted for trending in the baseline period or other NREs. Further exploration of the interval data yielded groups of data anomalies that could indicate NREs. We refer to these as "series anomalies" (See Appendix 0). **If the significant series anomalies we observed in baseline and reporting period interval data are indications of NREs, it appears from our exploratory analysis across the twelve sites that the occurrences of significant NREs could be quite small.** That said, one site had a substantial NRE occur during the implementation period[1] which our analysis missed since it compared reporting period to baseline period.

Trending, a gradual change in energy use not accounted for by weather or other independent variables, in the baseline period is a specific type of ongoing NRE that we analyzed separately. We found significant baseline period trending in electric consumption at nine of the twelve

[1] Reference the Joint Study Report when it comes out.

sites, six of which would reduce savings if accounted for. Additionally, we observed half of the sites had significant baseline period trending in gas consumption, three with increasing use and three with decreasing use. However, **across all the sites with significant baseline period trending, none of the downward trends caused savings to reduce to less than 15% of annual consumption.**

We reviewed but did not make changes to the Option D models developed by the implementers for eight of the sites. **In our review of the Option D models, we found a substantial number of issues in modeling practices that raise considerable doubt as to the reliability of the savings estimates.** We believe a low level of rigor in developing the simulation models was due more to cost constraints than the capabilities of the modelers. Examples of issues we consistently observed across sites include the use of default values from DOE grocery model templates, over-simplified HVAC zoning, inaccurate scheduling, and incorrect building orientation.

We compared the Option D savings estimates of those eight sites to our ECAM estimates. There is often the assumption that engineering based approaches, such as Option D calibrated simulation, yield the more conservative (lower) savings estimate; however, **at half of the sites, the Option D estimates were larger than our ECAM estimates, often quite significantly so.** Examples of reasons for differences in the estimates include unaccounted for NREs in the ECAM models and improperly simulated energy use during periods when measures should be affecting energy use. We also observed significant bias in the Option D baseline and reporting period models of the grocery sites which in some cases added up to the difference from our savings estimates.

In our review of the materials developed for the CWB Demo, we found that generally the procedures were well-defined and followed industry best practices (or established them when there was little precedent). However, our review of the project-specific materials revealed that PG&E did little to enforce compliance with the procedures. Problem areas included inconsistent meter and weather data used by the various parties modeling each site, poor quality meter data, inadequate documentation of NREs, inconsistent baseline and reporting period dates, and insufficient reporting of goodness of fit metrics.

Recommendations

PG&E should update the CWB Policy and Procedures manual to incorporate the learnings from this and other recent studies and, most importantly, enforce compliance with its manual.

Program administrators (PA) should require the use of transparent, open-source, Option C-based algorithms for savings estimation in NMEC programs. The Option D approach is too expensive to use to develop reliable whole building savings estimates cost-effectively across a large portfolio of buildings.

Implementers should set up Option C baseline models as early in the projects as possible and begin comparing actual use to modeled use at the start of implementation, and continuing through the reporting period(s). This will enable the implementer to observe that the savings are occurring as expected and bring attention to times when they are not, either because a measure is not operating as planned or a non-routine event is occurring.

The industry needs to put more effort into developing transparent, robust automated routines for detection of and adjustment for non-routine events including baseline trending. This effort should include examining whether statistical methods and engineering methods can produce the same results.

PAs should require monitoring for non-routine events and implementers should facilitate the process of logging them, with the goal of minimizing introduction of new steps in work flow, e.g., use existing facility management systems such as work order and invoice tracking.

PAs should provide the same set of cleaned interval data to all parties involved with estimating energy use and savings at participating sites.

Implementers should select from the list of California climate zone weather stations nearest to the site that have historical weather data and document the name of the station and the date they obtained the weather data.

Implementers should ensure that their Option C models have appropriate physical relevance in addition to meeting statistical goodness of fit criteria.

Table of Contents

Introduction

The Commercial Whole Building (CWB) Demonstration ("Demo") was an invitation-only, pay-for-performance incentive trial designed to achieve 15+% energy savings in existing commercial buildings. The CWB Demo involved 12 commercial sites, all of which underwent installation of multiple energy efficiency (EE) measures. The Demo relied on International Performance Measurement and Verification Protocol (IPMVP) whole building approaches to estimate savings, namely Options C and D (EVO 2016). The Option C approach prescribes regression analysis of building energy meter data with independent variables, such as weather. The Option D approach applies calibrated simulation of building energy use. Demo specifications required a 12-month baseline period and a 12-month reporting period. Intervention periods at the 12 participating sites ranged from a little over half a year to nearly two years. PG&E contracted with two vendors and one technical consultant who estimated whole building avoided energy use[2] with the Option C approach. The two vendors used proprietary algorithms to model consumption and the technical consultant used two public domain algorithms, Mean Week (MW) and Time of Week and Temperature (TOWT). Furthermore, the contractors who implemented the efficiency measures at the Demo sites ("the implementers") also estimated normalized savings with the IPMVP Option D approach, by developing physics-based simulation models in eQUEST to simulate building energy consumption.

Savings Estimation

Option C: Create model of energy use during baseline period with baseline period weather. Use reporting period weather to create an "adjusted baseline" of what energy use would have been in the reporting period if the measures had not been implemented.
Savings = adjusted baseline energy minus actual energy.

Option D: Create simulation of energy use with efficiency measures implemented. Calibrate to reporting period energy use with reporting period weather. Replace measures with existing equipment in simulation to get baseline energy use.
Savings = simulated baseline energy use minus simulated reporting period energy use.

PG&E contracted with SBW Consulting in 2017 to conduct an Early Measurement & Verification study of the CWB Demo. The scope of the study included technical review of program materials, independent estimation of savings for each Demo site, comparison of the Demo Option C savings estimates, review of the implementer Option D savings estimates and comparison to Option C estimates, and an assessment of the evaluability of the Demo. There were three key objectives of this study:

1. Provide an assessment of the Demo's "evaluability", or how well it supports evaluation, and provide recommendations for modifications to the program design to improve its evaluability prior to scaling the Demo into a full-fledged program.

[2] Throughout this report, "avoided energy use", "estimates", and "savings estimates" are generally used interchangeably, unless otherwise stated. We only used the avoided energy use estimates for comparisons, not normalized savings. Moreover, we used our estimates prior to adjusting them for NREs and baseline trends.

2. Conduct an independent evaluation of savings resulting from the application of Option C for the Demo's 12 projects, and the application of Option D for 8 of the Demo's projects, to inform the annual EE savings claim that will be based on the 12 Option D results using code and Industry Standard Practice as required by the policy in effect at the time of program approval which preceded the passage Assembly Bill 802.
3. Provide recommendations for the best methodology to use for estimating CWB avoided energy use in the future. These recommendations will inform the recommendations to be included in the Study Process as regulatory guidelines for future program implementation.

Due to the small number of buildings in the Demo and the way they were recruited, the results of this study are not representative of the general population of buildings that may participate in a NMEC-based program. That said, the research conducted in this Early M&V study addresses misconceptions about the reliability of Option C-based and Option D-based savings estimates. Furthermore, this study yields timely recommendations for improvement of data and documentation practices that will apply across a wide variety of NMEC program implementation approaches and building types.

Methodology

This section details the methodologies the SBW engineers used to meet the objectives of the Early M&V study. The goal of this section is to document clearly what we did do, as well as what we did not do.

IPMVP Option C Model Assessment

This section describes how SBW assessed the viability of the Option C approach to whole building savings estimation. First it describes the process of establishing our independent savings estimates. Then it describes our review and comparison of the Demo Option C models and savings estimates. Finally, it details our explorations of novel statistical methods to identify non-routine events and baseline trending.

Development of Independent Option C Savings Estimates

One of the key objectives of this study was to conduct an independent estimation of Avoided Energy Use and Normalized Savings resulting from the application of Option C for the Demo's 12 projects. The SBW team produced these estimates using the open-source, Excel-based tool Energy Charting and Metrics ("ECAM"). ECAM provides a standardized and transparent means for measurement and verification (M&V) of energy savings. Its consistent, repeatable methodology for measuring savings adheres to the IPMVP (Efficiency Valuation Organization, 2016). ECAM uses methods from ASHRAE Guideline 14, Measurement and Verification of Energy and Demand Savings (ASHRAE 2014). The following steps describe the overall approach to using ECAM to define a model and estimate energy savings:

1. Synchronize, format, and address obvious outliers[3] in the raw interval data (see Appendix 0).
2. Define the confidence level to use for model uncertainty and uncertainty of savings. Per the California Energy Efficiency Evaluation Protocols, we selected a confidence level of 90%.
3. Specify the time interval for the independent and dependent variables. For this project, we developed two sets of electric models, one based on daily intervals and the other based on hourly intervals. Gas models were based exclusively on daily intervals since the meter data was provided in daily format.
4. Acquire actual weather data. For this project, we obtained local climatological data (LCD) from the National Oceanic and Atmospheric Administration's (NOAA) website https://www.ncdc.noaa.gov/cdo-web/datatools/lcd. For each site, we chose the nearest weather station that offered the historical temperature data we needed. Table 1 shows the stations we selected for each site.

Table 1 Selected Weather Stations

CWB Site ID	NOAA LCD Station Name	CWB Site ID	NOAA LCD Station Name
1	Fresno Yosemite International	44	San Jose
14	Fresno Yosemite International	51	San Jose
16	San Jose	52	San Carlos Airport
21	Napa County Airport	54	Sacramento Metropolitan Airport
24	Moffett Federal Airfield	60	Concord Buchanan Field
42	Fresno Yosemite International	63	Ukiah Municipal Airport

5. Determine the categorical variables for the model. These include Daytype, Occupancy, combinations of Daytype and Occupancy, or combinations of Daytype and Hour of Day. For gas models Daytype was the only categorical variable.
6. Specify the form of the model for each category and create the models. The model forms available in ECAM are shown in Figure 3. For this project the x-axis represents outdoor temperature and the y-axis represents energy demand (therms per day or kW). The best form is the one which results in the lowest Root Mean Squared Error (RMSE), is statistically significant, and provides physically realistic predictions (e.g. gas and electric demand never drops below zero, gas demand should not rise with outdoor air temperature).

[3] We identified outliers for further investigation, but we did not remove them from the data used to build models. We only removed energy use values of zero.

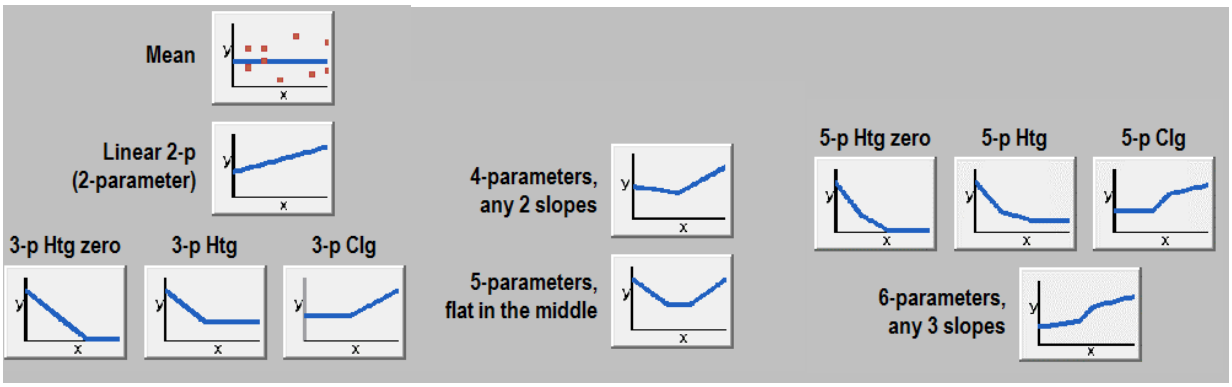


Figure 3: ECAM Model Forms

7. Estimate the savings by projecting the baseline model to reporting period conditions and subtracting the reporting period energy use from the adjusted baseline energy use. This is the IPMVP "Avoided Energy Use" type of savings. This step includes estimation of the overall uncertainty in the savings estimate. If the reporting period data covers less than a complete year, increment the "Avoided Energy Use" by proportionally extrapolating it to cover the complete year.
8. Calculate normalized savings by also creating a model of the reporting period. Then, for both the baseline and reporting period models, replacing the actual weather data in the model with California Energy Commission (CEC) Climate Zone (CZ) data and adjusting other model inputs as necessary to reflect typical conditions and modeled typical year energy use. This is the IPMVP "Normalized Savings".[4]
9. Identify non-routine events (NRE) including trending in the baseline period and assess their impact on the avoided energy use estimates (see Sections 0 and 0).

Comparison to Demo Option C models

After we delivered our independent savings estimates for the study, PG&E provided us with the input and output data from the Demo models. The Demo models comprised two proprietary algorithms ("Vendor A" and "Vendor B") and two public domain algorithms, Time of Week & Temperature (TOWT) and Mean Week (MW). The proprietary vendors did not provide their algorithms, but descriptions of the public domain algorithms are available briefly described under Public Domain models in the Glossary. The Demo electric models were hourly and gas models were daily[5]. The outputs of the Demo models did not include annualized savings estimates. For the purposes of comparison, we calculated the avoided energy use from all models (ours and the Demos') as the sum of the difference between the adjusted baseline

[4] In some cases, the CEC weather data came from a different weather station than the station used for the actual weather. We compared the differences in savings estimates based on the different weather and observed that the differences were always less than a quarter of the savings uncertainty, i.e., in the noise.

[5] Proprietary vendor B did not provide the input data it used in its models.

modeled consumption and actual reporting period metered consumption for each instance which had a value for both[6].

In all, there were 88 Demo models across the 12 sites estimating electric and gas savings. Rather than compare differences in all 88, we first identified which Demo estimates were significantly different from our estimates by checking if the Demo estimate fell within the confidence band of the ECAM savings estimate for each model at each site as demonstrated in Figure 4. We chose a confidence band equal to twice the uncertainty in the ECAM savings estimate at each site for each fuel, as we did not have uncertainty information for the Demo models. Put another way, we assumed the Demo model uncertainty was the same as the ECAM daily model uncertainty, hence the confidence band to check for overlap between the estimates was twice the confidence band of the ECAM estimate.

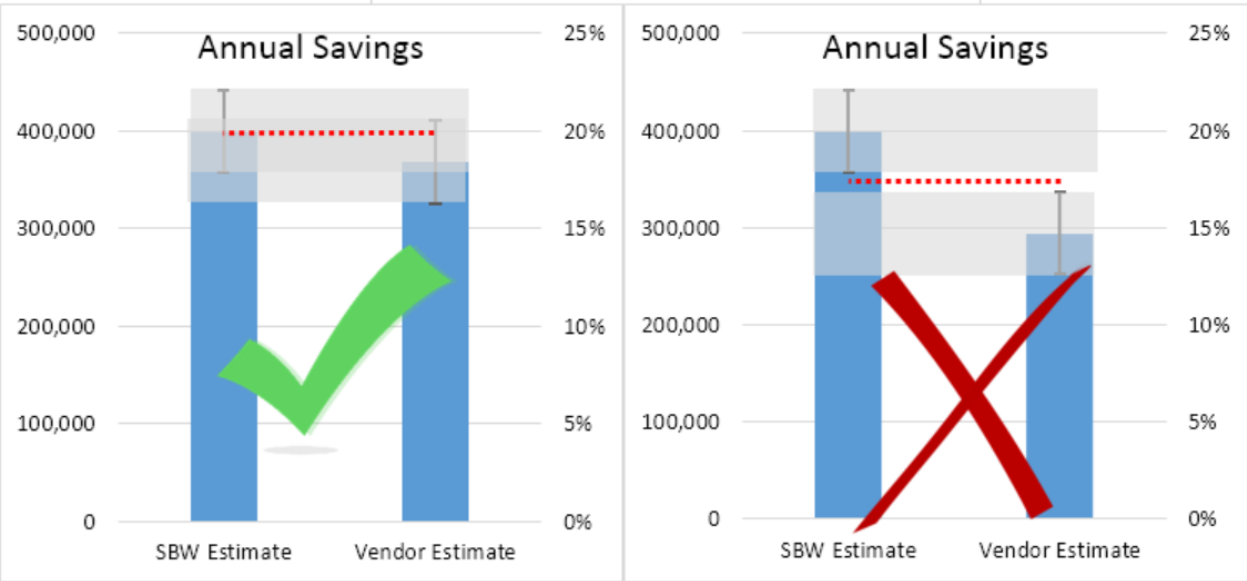


Figure 4: Confidence band comparison

For Demo estimates that fell outside the confidence band, we compared the baseline and reporting period dates, weather data, metered and modeled energy use, and model residuals to those used in our ECAM models to look for reasons for differences. We loaded the Demo data into ECAM and used the Time Series Charts, Scatter Charts, and Residual Trends features to facilitate the comparison analysis.

In addition to comparing the savings estimates calculated by the Demo models, we calculated statistics to describe how well the Demo models fit the data (see Appendix A.2 Goodness of Fit Criteria).

[6] The metered data did not have a valid value for all 8,760 hours (or 8,784 hours for leap year) in each year for each site. The team did not interpolate those missing values. We dropped the records with the invalid or missing values.

Identification and Treatment of Non-Routine Events

From the perspective of whole building, meter-based modeling, non-routine events (NRE) are changes in energy use in the building that were unassociated with program interventions and could not be accounted for with the model's independent variables (e.g., weather and occupancy). Some NREs may have large enough impact that savings estimates should be adjusted to account for them, such as removal of a large load like a data center or change in occupant type, e.g., from retail to restaurant. Ideally, all significant changes in building operations that affect energy use should be logged and reported to the program implementer. Additionally, NREs should be identified through careful examination of the meter interval data, following the methods described in Appendix 0.

Once an NRE has been identified, there are four approaches to quantifying its impact using the model (BPA 2018):

1. Look at the time series of **Residual**s for a model that includes the time period of change and estimate the magnitude of the change from the change in the residuals.
2. Use a pre-post model with a 'mini baseline' and 'mini post' period. The mini baseline is the shorter time period that exists within a baseline or reporting period and is prior to the NR change. The mini post is similar—for a NR change that is ongoing, it is the shorter time period within a baseline or post period that includes the NR change. The pre-post model uses an indicator variable for the mini post period, and the coefficient on the indicator variable is the NRE impact. This is a more robust method of looking at the time series of residuals.
3. For a change of long duration, especially one that is ongoing through the time period, e.g. baseline period: Treat the time periods around the non-routine change as a mini baseline and a mini post period, and model the change by subtracting the mini post period energy use from an adjusted baseline developed from the mini baseline period. This can be done using either a forecast or backcast approach, depending upon which mini period has better coverage for the independent variables.
4. For a temporary NR change of relatively short duration: Model the entire period (e.g. baseline period) excluding the portion of the period that includes the non-routine change. Use this model in conjunction with the independent variable(s) for the times that include the non-routine change to estimate energy use for the entire period as if the non-routine change had not occurred. Subtract this estimate from the actual energy use for the period to estimate the impact of the change.

Identification and Treatment of Long-Term Trends in Baseline Energy Usage

The SBW team reviewed the baseline period residuals in each model to identify whether the energy use had an upward or downward trend in a manner unexplained by the independent variables. If the model of the residuals described a statistically significant slope, we assumed that the trend reflected a long-term change in the energy use during the baseline period. We recalculated savings relative to the modeled rate of energy use at the end of the baseline

period, rather than relative to the modeled average rate of use throughout the baseline period and assessed the significance of the correction. This process took three steps:

1. ECAM automatically generated a linear regression model of the normalized residuals (%) versus time and determined whether it was statistically significant, as indicated by the t-statistic for the slope of the trendline greater than 1.3. Where a slope was significant, we assumed it reflected a significant upward or downward change in energy use over the whole baseline period.
2. Where a trend was significant we corrected the *adjusted* baseline energy use by a percentage equal to the value of the model of the normalized residuals at the end of the baseline period and corrected the avoided energy estimate accordingly.

IPMVP Option C and Option D Methods Comparison

The SBW team undertook this task to explore reasons for differences between savings estimated for the Commercial Whole Building Demo (CWB Demo) with the Option C and Option D approach and to determine whether Option C-based savings provides at least equivalently reliable savings estimates as Option D-based estimates for programs using savings estimated from changes in metered energy use, such as Normalized Metered Energy Consumption (NMEC) programs. To that end, we completed the following sub-tasks:

- We conducted a detailed examination of the Option D models for eight of the twelve Demo projects. The other four were subject to review by Energy Division consultants (ED) as part of the Joint Study. Our team and the Joint Study team both reviewed one of the projects.
- We performed in-depth comparisons between Option D models and SBW-developed Option C models for those eight Demo sites.

Option D Model Review

We developed a review template to check the accuracy of the Option D models against actual building conditions from the reporting period. We examined eight models and recorded results in site-specific review workbooks using the following information supplied by PG&E and the project implementer:

1. Post-reporting period verification ("VR2") reports to check modeling methodology, justification/source for model inputs, calibration process, and non-routine event handling.
2. Reporting period electric and gas utility monthly billing data to check model calibration.
3. Building plans where available
4. Control system trend data when available
5. Control screen prints showing a snapshot in time of building operation including setpoints, system/equipment status, schedules, etc.

The Option D model reviews entailed a site-specific assessment of the level of modeling rigor required to adequately capture the performance of installed measures. The level of rigor required depends on building type and measure complexity. For example, hospitals require

higher rigor level than big box retail, and rooftop units do not require as high a rigor level as a variable volume system. We determined if the appropriate complexity required modeling every single zone and system or if a simpler approach was sufficient, and what level of data was required to adequately define the model (i.e. site visits, building plans, commissioning reports, trend data, logger installation, billing data, etc.).

Model input/output checks included identification of the elements of the models that were too simplistic or more detailed than required. We looked into general building inputs (square footage, building orientation, etc.), model zoning (physical and HVAC zones), envelope (window, wall, roof, floor, infiltration), HVAC & refrigeration (system configurations, equipment capacities, efficiency, setpoints, schedules, staging, etc.), lighting (wattage, controls, schedules), and measure-related inputs (detailed check of baselines, measure inputs and modeling strategy). We also checked weather files to ensure that the weather data matched the reporting period from the most appropriate weather station and conducted a reasonableness check of our outputs and results.

We assigned each model check a disposition as follows:

- No issue found
- Model error found
- Potential issue (issue suspected that would affect model accuracy, but not enough information to determine and not necessarily an error). These could include:
  - Questionable input (not an error but the input could be improved upon)
  - Reasonable input without a source (appears reasonable but no way to confirm)
  - Building operational issue

We then assigned each issue a qualitative savings impact (small, medium, large)[7]. We asked implementers about all medium and large issues to make a final determination. Note that we did not ask about small impacts due to the large number of small impact issues found, and the time it would have taken implementers to respond to each one. We acknowledge, however, that many small impact issues taken together could have a larger impact on savings. For this reason, we included all small impact issues in our analysis. We changed the issue disposition for medium/high impact issues as needed based on implementer responses.

Option C and Option D model comparisons

For this task, the SBW team compared savings estimates and model inputs and outputs of our ECAM models and the Option D models. We ensured that all comparisons account for any differences between Option C and Option D approaches regarding period of calibration and numbers of models or simulations used in the energy savings estimates. Prior to comparison,

[7] The scope of the study did not include making modifications to the simulations and re-calibrating them so we were unable to quantify the savings impacts due to the various errors and issues.

we adjusted the simulation period and weather data in the Option D models to match the Option C models we generated in Task 3.

To aid in the comparison of Option C and Option D models, we created a single ECAM file for each site that included the following hourly data points from both the Option C model we generated in Task 3 and the implementer's Option D model:

- Outside Air Temperature (°F)
- Metered Energy (kWh or therms)
- Modeled Energy (kWh or therms)
- Residuals (kWh or therms)

We focused our comparisons on the reporting period and analyzed electricity and natural gas separately. We aligned the timestamps for the Option C and D models (retaining outlier information from the Option C model) and used ECAM's charting tools to compare weather data, metered energy consumption, and modeled energy consumption. For electricity, we compared the average daily load shape for different categorical variables (e.g. day type), and for natural gas, we compared the average consumption by day of week.

Finally, we calculated the following model statistics for each model: Net Determination Bias Error (Bias), Coefficient of Variation of the Root Mean Squared Error (CV(RMSE)), Coefficient of Determination (R^2) and Mean Absolute Percentage Error (MAPE).

Additionally, we were also tasked with providing comparisons of the estimated cost per site of using the Option C and Option D approach. PG&E supplied us with actual costs per site for Options C and D pilot model development. We used these as a starting point together with our full-scale program implementation experience to estimate total cost per site for CWB program scale-up. We estimated total cost as actual cost plus additional cost for improvements identified in the model reviews.

Program Material Technical Review

The SBW team conducted a technical review of project-specific and program-level documentation to identify shortcomings and offer suggestions for improvements and additions to the program operation. We compared the submitted data and documentation provided over the course of the study to the program requirements and specifications provided in the documents reviewed. Additionally, we compared the Demo datasets and calculated statistics to ECAM data sets and calculated statistics to determine the consistency between model input data provided by PG&E and calculation methods for statistical reliability.

Findings and Recommendations

This section details the findings and recommendations from our independent savings estimation as well as review and comparison of the Demo Option C and Option D models and savings estimates.

IPMVP Option C Models Assessment

This section presents the findings and recommendations from our assessment of the application of the Option C approach to estimating savings in the CWB Demo.

Independent Savings Estimation Findings

SBW engineers used ECAM to produce Option C models for estimating Avoided Energy Use and Normalized Savings for the 12 Demo sites. We created both hourly and daily models of electricity use. We used hourly models to get the estimated savings, and daily models for the savings precision, for the following reasons:

- Utilities are interested in the timing of savings. Daily models can only provide the savings by daytype, but hourly models can provide savings by time of day and time of week.
- The vendor and consultant models were all hourly models.
- Hourly models, since they have the most data and information, should provide the most accurate savings estimate.

At present the industry does not have a reliable way of estimating savings uncertainty (precision) using hourly models, but it does have accepted ways of estimating uncertainty using daily models[8].

Our estimates of annual electricity savings from the daily and hourly models were quite close, well within the confidence interval for savings from the daily model as can be seen in Table 2 and Table 3. As such, in the remainder of the report "SBW" or "ECAM" electricity savings estimates are from the hourly model and uncertainty from the daily model. We only had daily gas use data so all gas savings and uncertainty estimates are from daily models.

Table 2: Electric avoided energy use estimates

Building Type	Site	*Daily Model Annual Savings (kWh)*	Hourly Model Annual Savings (kWh)	Daily Model Uncertainty (kWh)	% Annual Savings (Hourly Model)
Grocery	14	*470,849*	471,480	± 20,881	25%
Grocery	16	*224,633*	224,630	± 15,484	17%
Grocery	21	*262,362*	262,351	± 11,799	15%
Grocery	42	*420,168*	421,847	± 8,772	33%
Grocery	63	*201,232*	192,578	± 12,738	15%
Office	1	*433,356*	429,747	± 15,885	29%
Office	24	*1,417,492*	1,422,582	± 89,683	30%
Office	44	*831,332*	824,052	± 27,271	36%
Office	51	*111,378*	106,805	± 11,810	19%

[8] For further information on estimating uncertainty from these types of commercial building models, see this abridged list: Reddy 2000, Lei 2011, A. Shonder 2012, Sun 2013, Baltazar 2014, Granderson 2016, Koran 2017.

Building Type	Site	*Daily Model Annual Savings (kWh)*	Hourly Model Annual Savings (kWh)	Daily Model Uncertainty (kWh)	% Annual Savings (Hourly Model)
Office	54	*286,800*	286,947	± 15,659	35%
Office	60	*132,264*	142,413	± 9,762	29%
Library	52	*171,927*	184,392	± 7,996	55%

Table 3: Gas avoided energy use estimates

Site	Building Type	Annual Savings (Therms)	Uncertainty (Therms)	% Annual Savings
14	Grocery	-6,927	± 808	-16%
16	Grocery	-3,817	± 1,030	-14%
21	Grocery	-448	± 1,456	-1%
42	Grocery	-583	± 137	-8%
24	Office	574	± 346	12%
1	Office	9,305	± 907	81%
44	Office	33,057	± 1,137	56%
51	Office	5,913	± 948	23%
54	Office	-2,262	± 173	-328%
60	Office	6,406	± 576	54%
52	Library	5,076	± 651	46%

Option C Model Comparison Findings

We compared our original independent savings estimates to the estimates provided by the Demo models. We examined a total of 88 Demo models[9] and found nearly three-quarters of the Demo savings estimates to be within our uncertainty band. We identified only 24 Demo savings estimates that were significantly different from our own savings estimates. The proprietary algorithms were significantly different for 33% of the estimates while the public domain algorithms were significantly different for 22% of the estimates.

Figure 5 and Figure 6 display comparisons of the different electric savings estimates out of each model for each site. The error bars are two times the study team's model uncertainty estimate.[10] As one can see, for most sites, most models estimate the same savings within the error band. Site 44 was an exceptional case because the vendors used interval data streams that had duplicate timestamps.

[9] There were four models for each of the 12 sites with electricity data and each of the 11 sites with gas data, except for Site 21 which did not have a gas model from Vendors 1 and 2 and Site 54 which did not have a gas model or an electric model from Vendor 2.

[10] We used two times our uncertainty estimate as a proxy for the total model uncertainty of our estimate and the vendor estimate since we did not have uncertainty estimates from the vendors. This allows for overlap in uncertainty bands by our estimate and the vendor estimates.

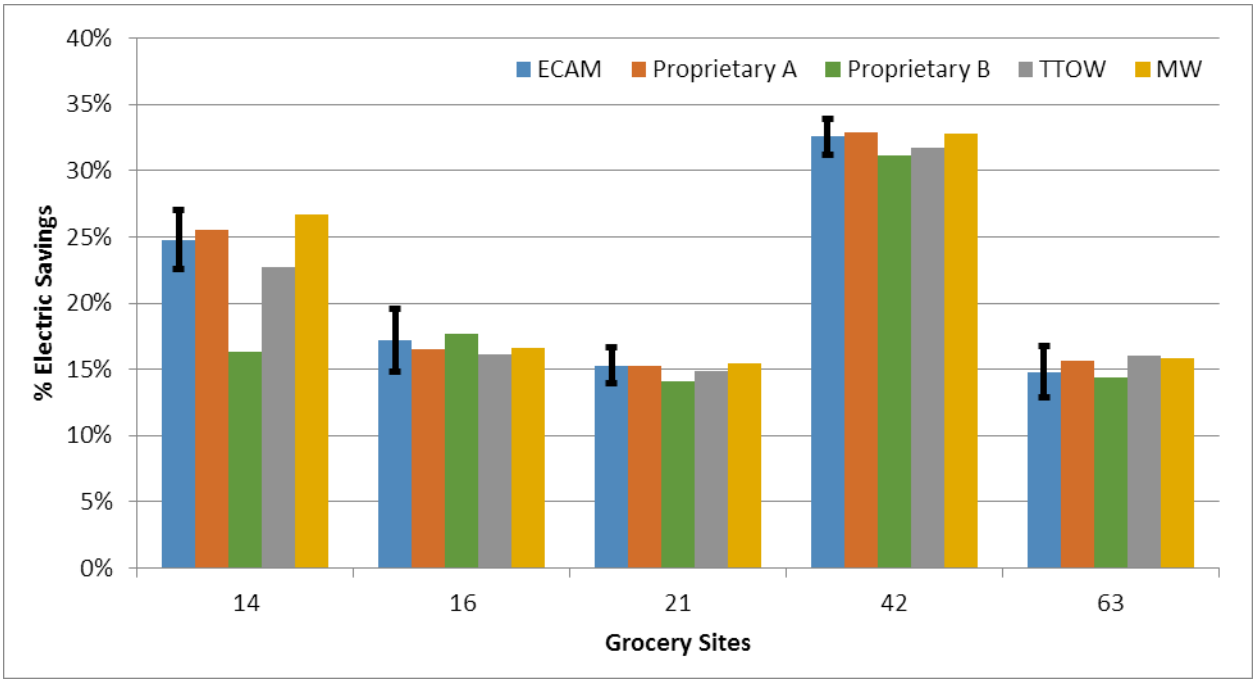


Figure 5: Electric savings estimates, grocery sites.

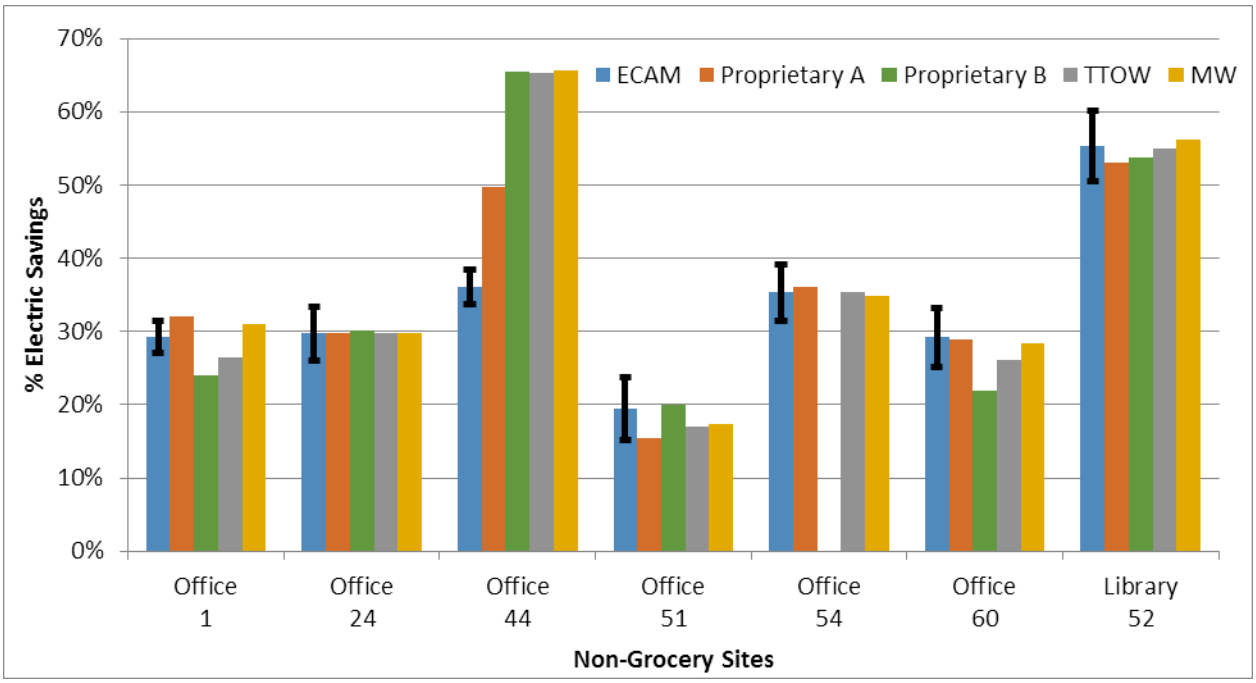


Figure 6: Electric savings estimates, non-grocery sites.

In Figure 7 and Figure 8 we present the gas savings estimates for all sites and models. The gas savings estimates varied more than the electric savings estimates with only five of the eleven sites having gas savings greater than model uncertainty. The more obvious observation is that gas savings were always negative for the grocery sites. The team did not receive gas savings estimates from the proprietary vendors for site 21. It was beyond the scope of this study to

examine reasons for negative savings. One hypothesis is that they are associated with interactive effects from refrigeration measures. The team recommends further study.

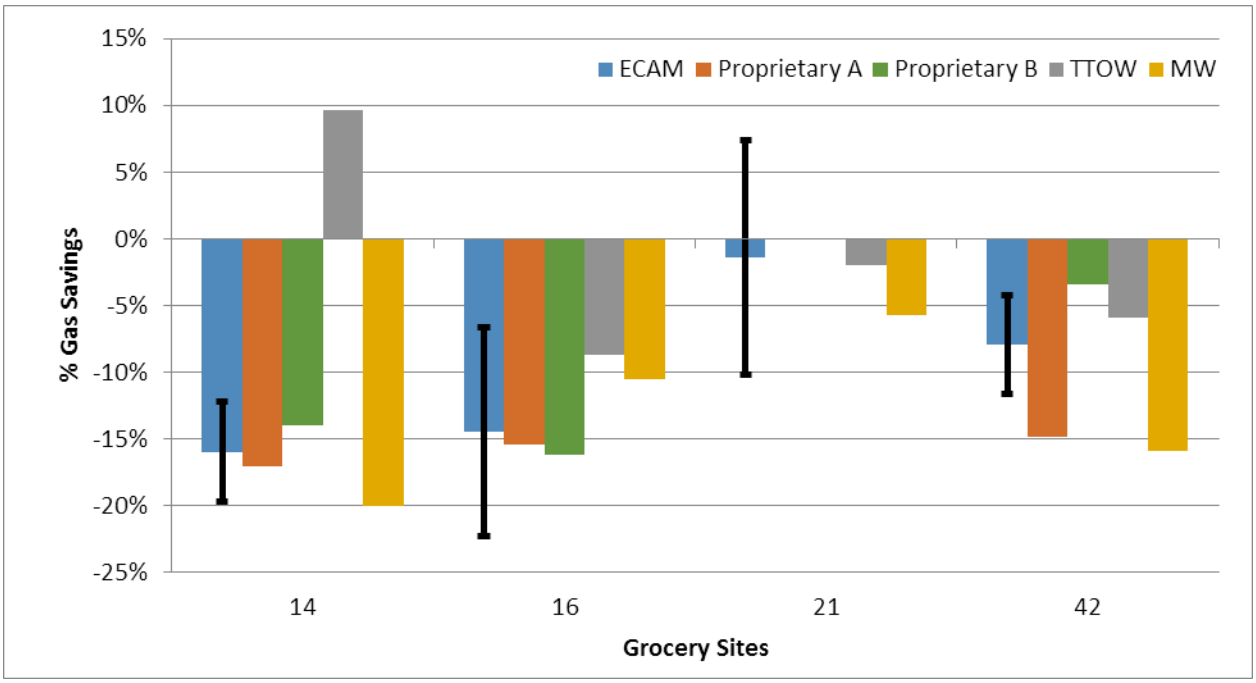


Figure 7: Gas savings estimates, grocery sites.

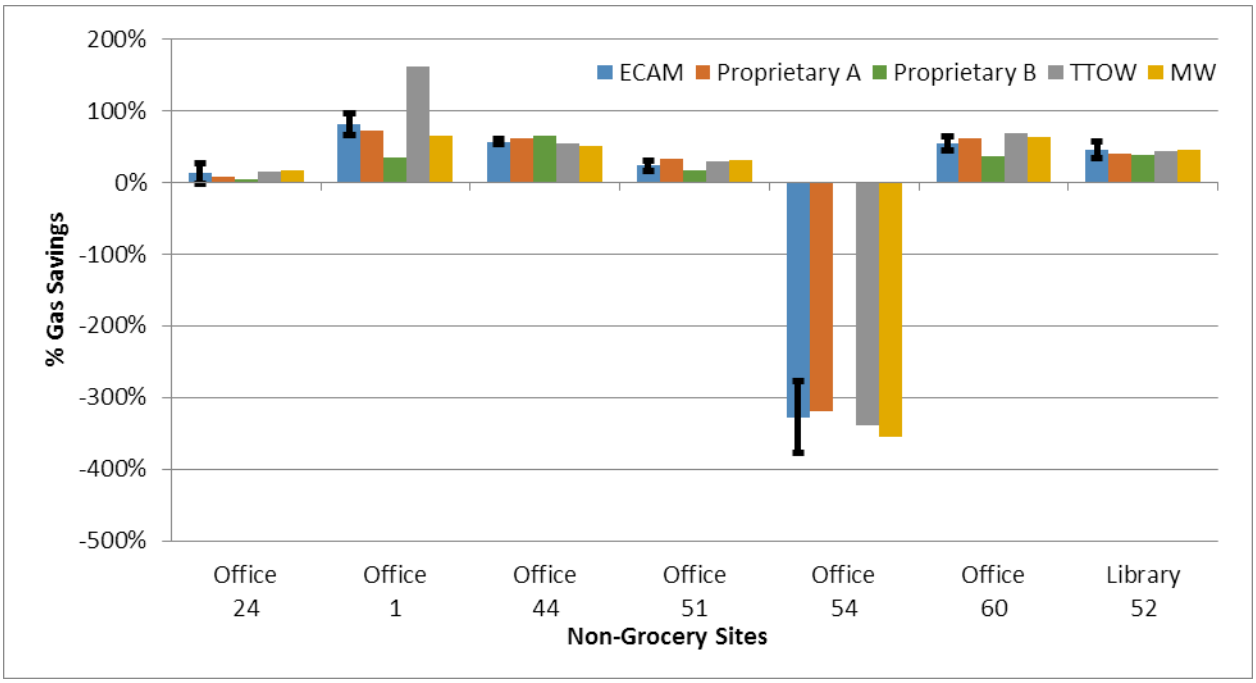


Figure 8: Gas savings estimates, non-grocery sites.

We compared the inputs and assumptions for each of the four Demo models for each site to those used in the ECAM models. We examined baseline and reporting period dates, outdoor air

temperature, as well as metered, modeled, and residual kW and therms to look for reasons for differences. Upon examining the differences evident in the model inputs and outputs, we attributed the majority of differences to the models rather than the input data. The study team categorized the observed discrepancies of savings results into the following categories:

- Model Differences
  - *Step-change in adjusted baseline model over last one and a half months of the reporting period.* In several models, the Demo models showed a step change in their modeled energy consumption for the last one and a half months of the reporting period, as illustrated with the orange line toward the end of the modeled period in Figure 9. After looking more closely at each of the models with this issue, we determined that the step change was likely due to missing weather data in the model.

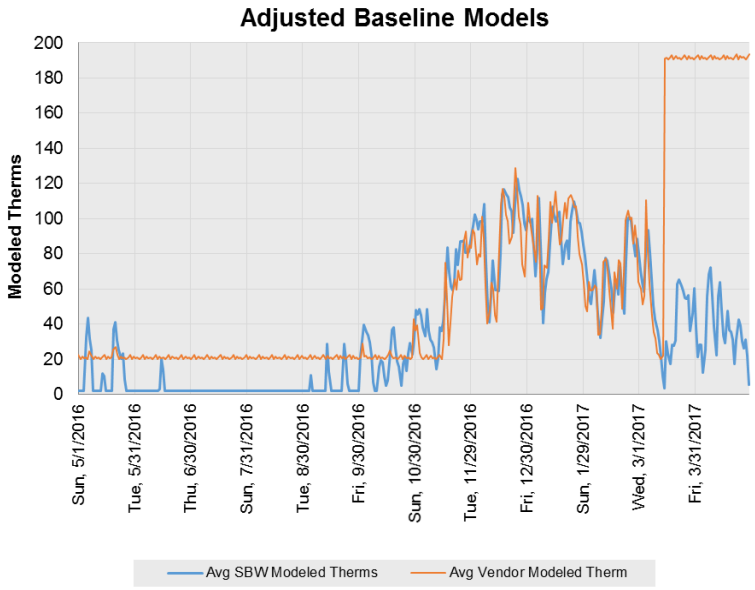


Figure 9: Example of "step change" in modeled energy use.

  - *Custom model with additional independent variable.* When developing the ECAM model, we incorporated seasonal behaviors into the model in the form of an additional independent variable. The proprietary vendors did not have such a relationship and therefore did not always model the energy use as closely as we did. At site 42, Proprietary Vendor A's gas model diverged from our gas model during shoulder season temperatures, 55°F to 65°F. At site 60, Proprietary Vendor B's electric model deviated from our electric model during cooling season temperatures, above 70°F. We observed similar vendor model errors for sites with behavior varying greatly with seasonality or by day type.
  - *Negative energy use.* The proprietary algorithms allowed modeled usage to be negative at two different sites. This practice creates a falsely high savings estimate in the model.

- *The Mean Week model does not include temperature as an independent variable.* This type of model will not perform well for energy use that is highly temperature-dependent.

- Input Inconsistencies
  - *Baseline weather discrepancies.* In some cases, our weather data did not match the weather data used by the vendors. To determine which weather data was more accurate, we re-ran our ECAM models with the weather data used by the vendor and recalculated the model statistics to compare to the original model.
  - *Discrepancies between baseline and adjusted baseline models.* We observed many models with this issue. The shape of the baseline model should not change, i.e., the relationship between baseline consumption and outdoor air temperature should not vary in the model used to compare to reporting period consumption.[11] Figure 10 exemplifies this issue at one of the sites – the gray triangles should follow the same shape as the orange diamond but diverge at temperatures greater than about 60°F.

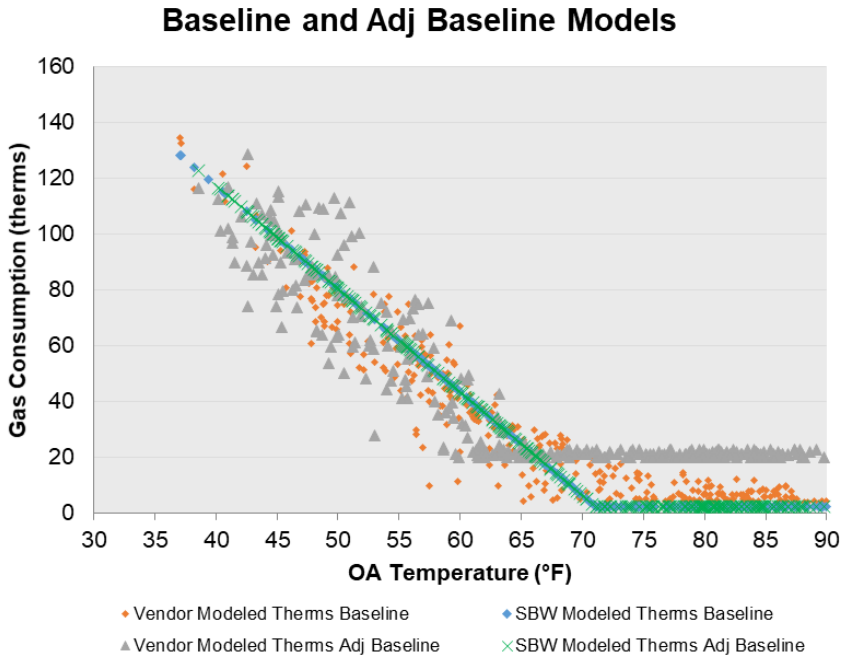


Figure 10: Example of discrepancy in adjusted baseline model

- *Differences in reporting period metered data.* We received different metered data than the vendors for several sites.

[11] Unless a non-routine event occurred, but the vendors did not document any reasons for adjustments to their models such as non-routine events.

- *Baseline study period offset.* In one model, the vendor baseline period was offset by two months from our baseline, creating a savings estimate discrepancy that did not exist during the overlapping measurement periods.

- Other:
  - *Insufficient data provided in model.* In one case, we were unable to determine why a discrepancy was observed between our savings estimate and the vendor's savings estimate. This was mainly due to missing baseline and temperature data from the vendor

After going through each of the site savings estimates and identifying and correcting for vendor model errors, we have determined that:

1. In most cases the model input variables are the same or very similar. Some sites obtained weather data from different sources causing some variation between inputs. The metered data we received for Site 44 electric was different than that received by each vendor. The Vendor 1 model for Site 44 gas assumed a different baseline period than our model and the other vendor models.
2. The majority of savings estimates received from the Demo fell within our defined savings estimate confidence band. Only 24 of the 88 Demo models were significantly different from our savings estimate. Based on our findings of reasons for the model differences, we were able to correct 22 of the 24 models such that their new savings estimate fell into our confidence interval.
3. Most of the reasons for differences were due to differences in the models and those were evenly split between proprietary and open-source algorithms, as Table 4 shows.

Table 4: Counts of Reason for Difference Types by Model Algorithm Type

Reason for Difference	Proprietary	Open-Source	Total
Model	12	12	**24**
Study Period Dates	1	0	**1**
Weather Data	1	0	**1**
Reporting Period Meter Data[12]	3	4	**7**
Insufficient Data Provided	1	0	**1**
Total	**18**	**16**	**34**

Goodness of Fit

We evaluated the performance of the Demo Option C models using the goodness of fit criteria described in Appendix 0. We were unable to evaluate the Proprietary Vendor B models using these metrics because we did not receive the baseline data for those models.

[12] There were no savings differences attributed to differences in baseline period data.

Table 5 summarizes the frequency of the Demo models exceeding the goodness of fit criteria limits.

Table 5: Count of electric models outside the defined threshold

Metric	Proprietary A		TOWT		MW	
	Electric	Gas	Electric	Gas	Electric	Gas
CV(RMSE)	0	5	1	6	1	11
R^2	1	1	1	2	6	11
Bias	12	0	11	9	7	9

Public domain models and proprietary models produce similar results across the board. With the small number of buildings in the Demo, we did not have a statistically representative sample to determine if certain vendor models were better for estimating specific building types. Option C (NMEC) models consistently produce accurate estimates of changes in whole building energy use. Proprietary models do not produce better or worse results than transparent open-source, public domain models. We recommend the use of open source models because of the transparency in the data and models to evaluate the savings estimates.

Generally, ECAM models had better fits across all sites than the Demo models. It is likely that SBW engineers put greater effort into creating the ECAM models than was put into some of the Demo models. We know that the TOWT model is an excellent model and can often provide better fits than ECAM, based on these metrics, if similar care is taken. Fits for models created in a fully-automated fashion may be somewhat poorer.

When performing statistical analyses, we found that 48 vendor models did not meet the net determination bias requirement of 0.005%. ECAM always forces the fit to have 0.000% bias. We found that 24 vendor models exceeded the CV(RMSE) requirement of 25%. None of the ECAM electric models exceeded this limit, and five of the ECAM gas models exceeded it.

Although five of the ECAM gas models did not meet the CV(RMSE) criterion, only one of the ECAM gas models was poor when looked at in a fuller context. Of the five sites with high CV(RMSE), four had very good models, with relative precision of the savings estimate under 15%. We note that CV(RMSE) can be a poor metric for utilities or sites where the energy use is highly variable but is low much of the time. Since the denominator of CV(RMSE) is average use, when average use is low CV(RMSE) is high. Hence, CV(RMSE) is often a poor metric for gas use: Since much gas use is heating-related, and seasonal, it can approach zero for much of the year. In such cases the average use will be low and CV(RMSE) can be high even with a good model.

Conclusions

1. All sites had electric savings beyond the model uncertainty and six of seven non-grocery

sites had gas savings beyond model uncertainty[13]. This success is due to a combination of the large savings achieved and the sites' energy use being statistically well-behaved, i.e., not noisy.

2. All grocery stores had increased gas consumption during the reporting period.
3. All five of the Option C (NMEC) models tested consistently produce similar estimates of changes[14] in whole building energy use. The reproducibility of the estimates of the reduction of whole building energy use following the installation of the energy efficiency measures provides strong supporting evidence for the use of this measurement approach moving forward.
4. The proprietary algorithms did not appear to provide any advantage over the transparent, open-source, public-domain algorithms for reliably estimating changes in building energy use.
5. Between the two public-domain algorithms applied in the Demo, the one that normalized to weather (Time of Week and Temperature algorithm) had better statistical fitness across all sites by most metrics.

Identification and Treatment of Non-Routine Events

The Demo Option C modelers did not provide any information about attempting to assess the presence of non-routine events occurring at any of the sites, nor any knowledge they may have had about non-routine events. However, we did receive information from the Option D implementers about NREs occurring at three sites. At two of these sites the events occurred during the respective project's reporting period, and at the third site the event occurred during the respective project's implementation period. We compared this information to the metered data and confirmed if the change in building operation was reflected in the metered data. We followed the third approach described in 0 to estimate the impact of the NRE on the electric meter at Site 51, and the fourth approach in 0 to estimate the impact of the NRE on the gas meter at Site 44. We also explored the coincidence between these two NREs and the occurrence of statistical outliers and series anomalies in the respective ECAM models. Below we describe in more detail how the reporting period NREs are reflected in the metered data, how they impact our avoided energy estimates, and how they coincide with the anomalies that we identified through statistical methods.

Site 24 Electric – Implementation Period

The Joint Study report stated that a 70 kW laboratory was removed from Site 24 during the implementation period. Since we built our model on the baseline period performance before

[13] All study team and vendor *electric* savings estimates were greater than model uncertainty. All study team gas savings estimates were greater than model uncertainty, but 9 of 24 *vendor* gas savings estimates for the non-grocery sites were less than model uncertainty, excluding the one non-grocery site which had increased gas consumption.

[14] Those changes incorporate both savings from program interventions and any non-routine events that may have occurred. None of the savings estimates included in this comparison analysis appeared to have been adjusted for non-routine events. See the Task 8 memo, which reports on our non-routine event detection analysis, as well as the full report on the CWB Early M&V study for further discussion on non-routine events.

the lab was removed, the model overestimated the kW use in the reporting period by an estimated 70 kW over the course of the year resulting in an additional 613,200 kWh of savings.

Site 44 Gas – Reporting Period

The Demo implementer reported that an NRE occurred on the gas side at Site 44 starting at the beginning of July of the reporting period due to a fault in the heating hot water control system. Figure 11 below shows the original metered and modeled gas usage over the reporting period, as well as the time-frame of the assumed NRE. The impact of this NRE is reflected in Figure 12 where gas usage clearly spiked up at the beginning of July. It dropped back to zero in the middle of July until the first week of August, when it spiked back up to anomalously high levels for the balance of the year, despite it being the cooling season. Given these observations, we can reasonably conclude that the gas usage over the last three months of the reporting period was non-routine in nature. Following the fourth approach described in Section 2.1, we estimate that this NRE had the effect of underestimating the original annual gas savings by about 10%. Juxtaposed with the uncertainty in the original savings estimate of only 3.4%, this represents a significant adjustment.

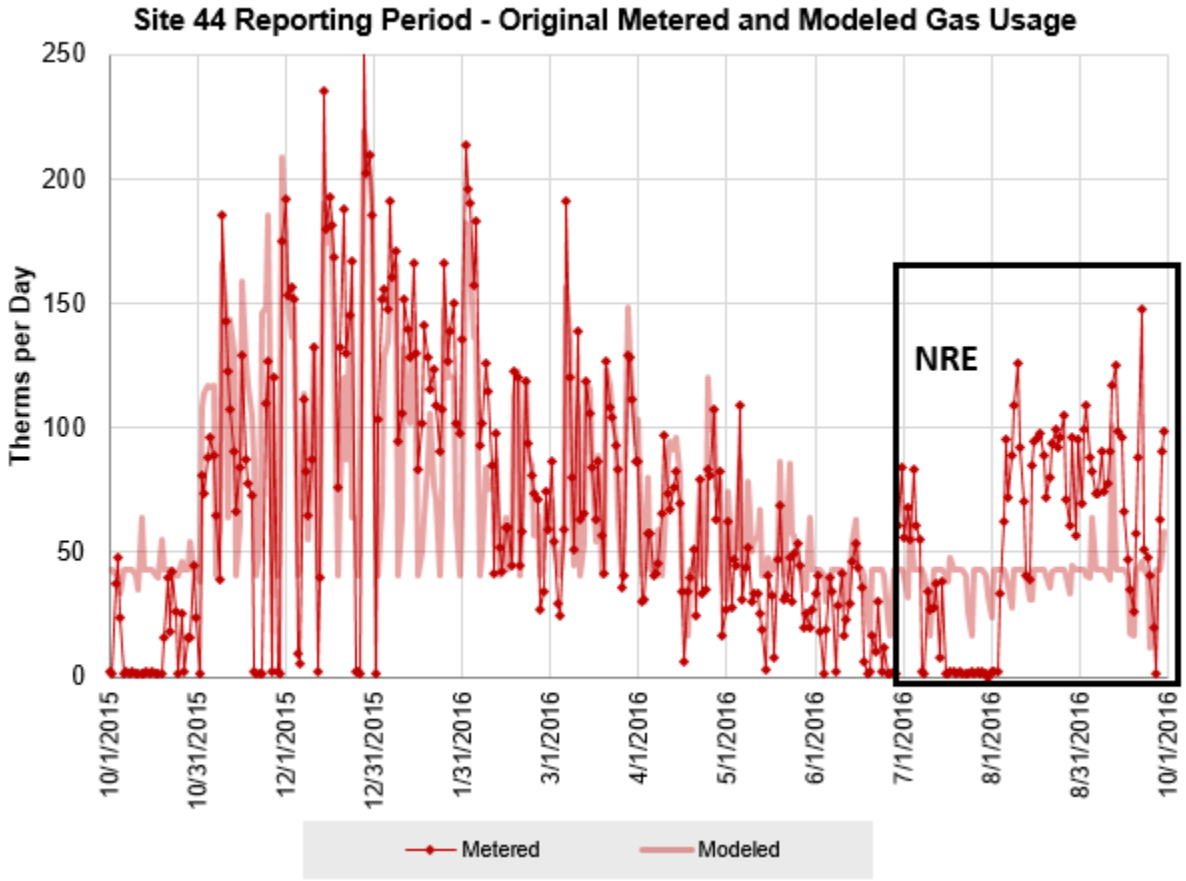


Figure 11: Site 44 - Original Reporting Period Gas Model

As shown in Figure 12 below, we found six series anomalies that coincided with the timeframe of this NRE. These six anomalies covered 23% of the NRE timeframe, a higher frequency of

occurrence than the 17% average[15] for the model shown in Figure 11. Since the result of this NRE was to create high spikes in summer gas usage, we can suspect that the series anomalies we found in the warm months are the direct result of the NRE. We do not know the causes of the series anomalies in the winter months, but we suspect that they are due to either errors in the model (e.g. building operating schedules not accurately specified), or non-routine events at the site.

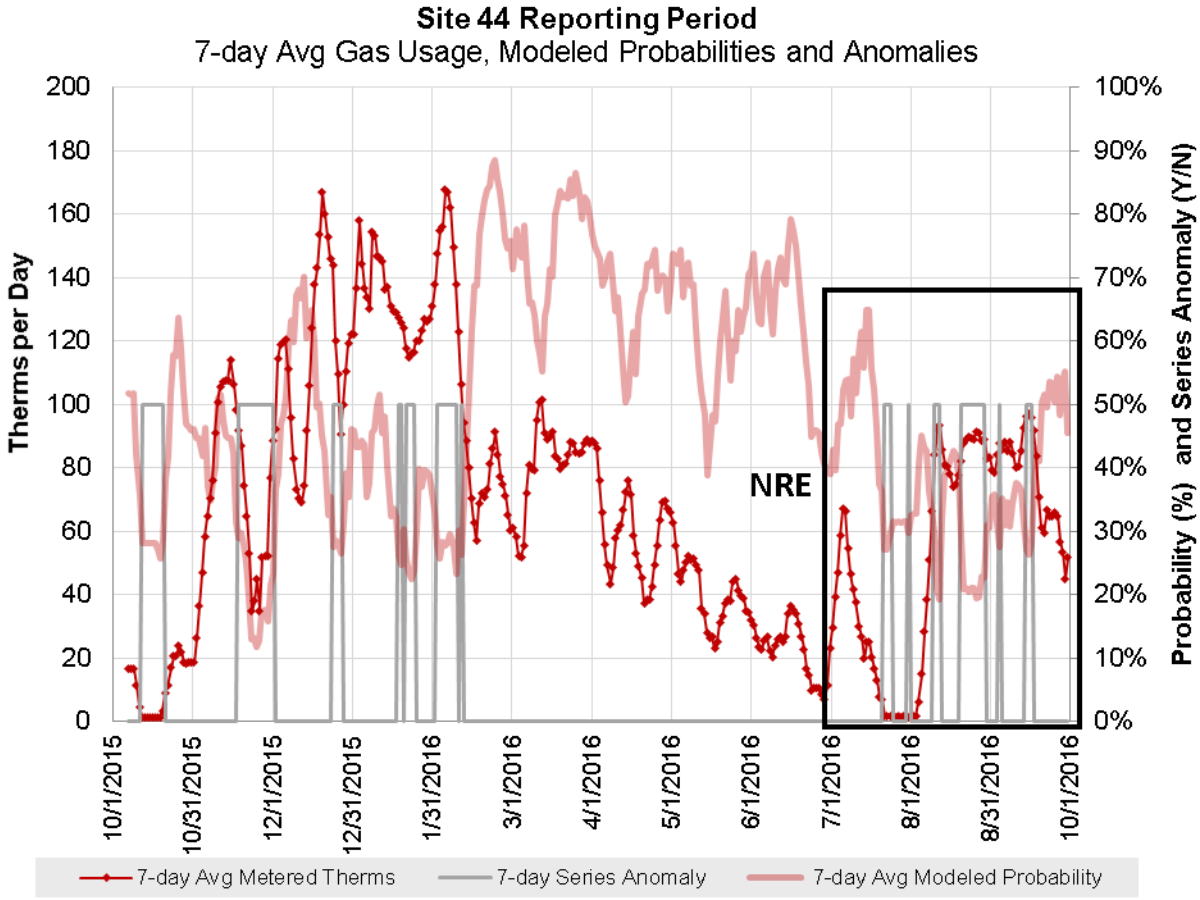


Figure 12: Site 44 – Reporting Period 7- day Avg Gas Usage, Modeled Probabilities and Anomalies

Site 51 Electric – Reporting Period

For Site 51, the Demo implementer reported that a large portion of the electric interval data spanning the first 6 months of the reporting period was not taken from the actual site meter but was instead estimated from historical usage. Coincidentally, as shown below in Figure 13, beginning near the end of May, energy usage clearly jumps up to a consistently higher level for the balance of the year. Furthermore, as shown below in Figure 14for a 10-day period, the

[15] Furthermore, the 17% average for this model is almost three times greater than the 6.5% average we found across all twelve gas models.

interval data takes a much different form starting on May 27th. Prior to that time the data consistently follows a relatively smooth sinusoidal trend of relatively low amplitude and magnitude, and after that time it consistently follows a more erratic trend of both higher amplitude and magnitude. While this phenomenon reflects errors in the raw data (i.e., estimated data was inserted in lieu of actual metered data), and are not the result of a true NRE, for discussion purposes here we refer to it as an NRE. We calculated that this NRE overestimated the original annual electrical savings by about 17%. Juxtaposed with the uncertainty in the original savings estimate of 11%, this NRE represents a significant adjustment to the savings estimate.

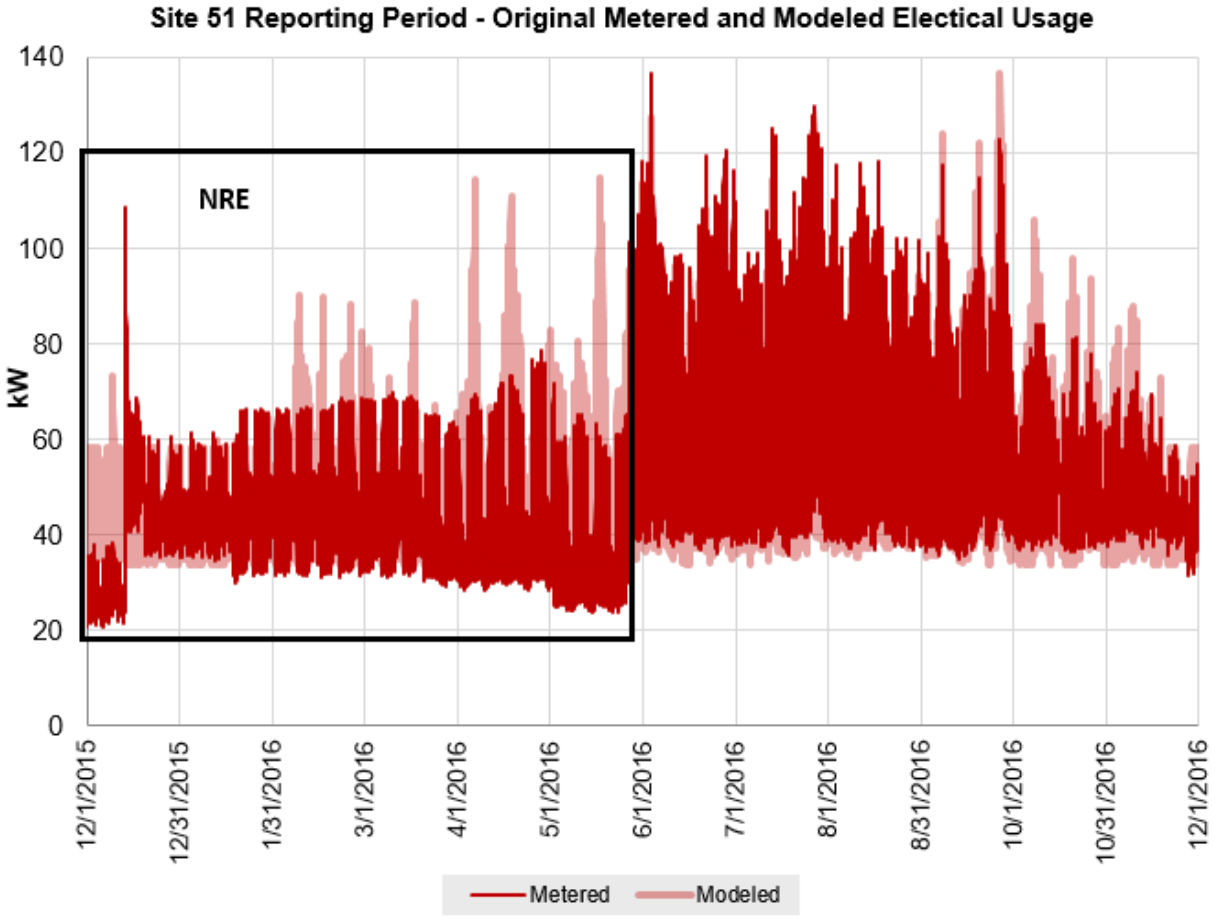


Figure 13: Site 51 - Original Reporting Period Electric Model

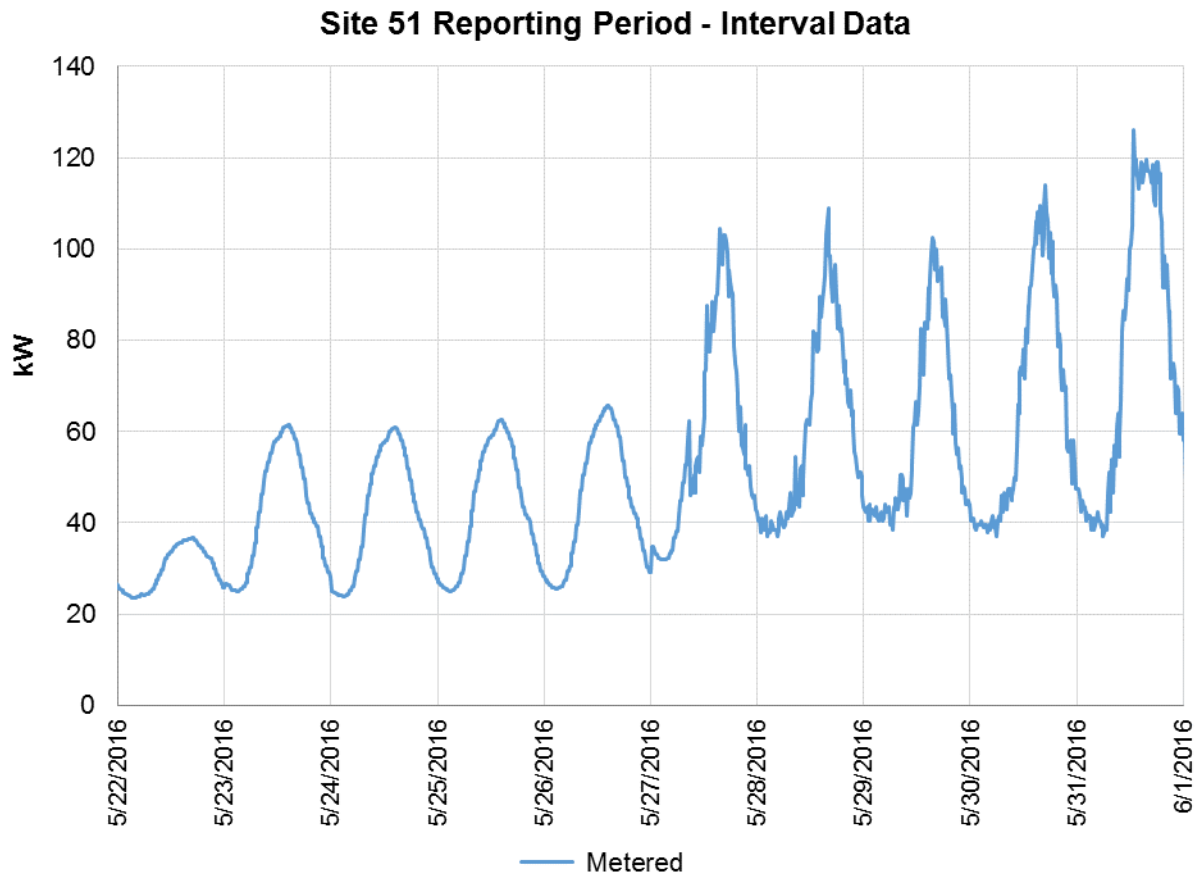


Figure 14: Site 51 - Interval Data

As shown in Figure 15 below, we found two series anomalies that coincided with the timeframe of the NRE. When cross-referenced with the trends shown above in Figure 14 it is evident that these anomalies also coincide with the two blocks of time during the NRE period when the data dropped to the lowest levels of the reporting period. These anomalies were not caused by the NRE (the insertion of estimated values in lieu of metered data) per se, but instead were caused by abrupt changes in the estimated values.

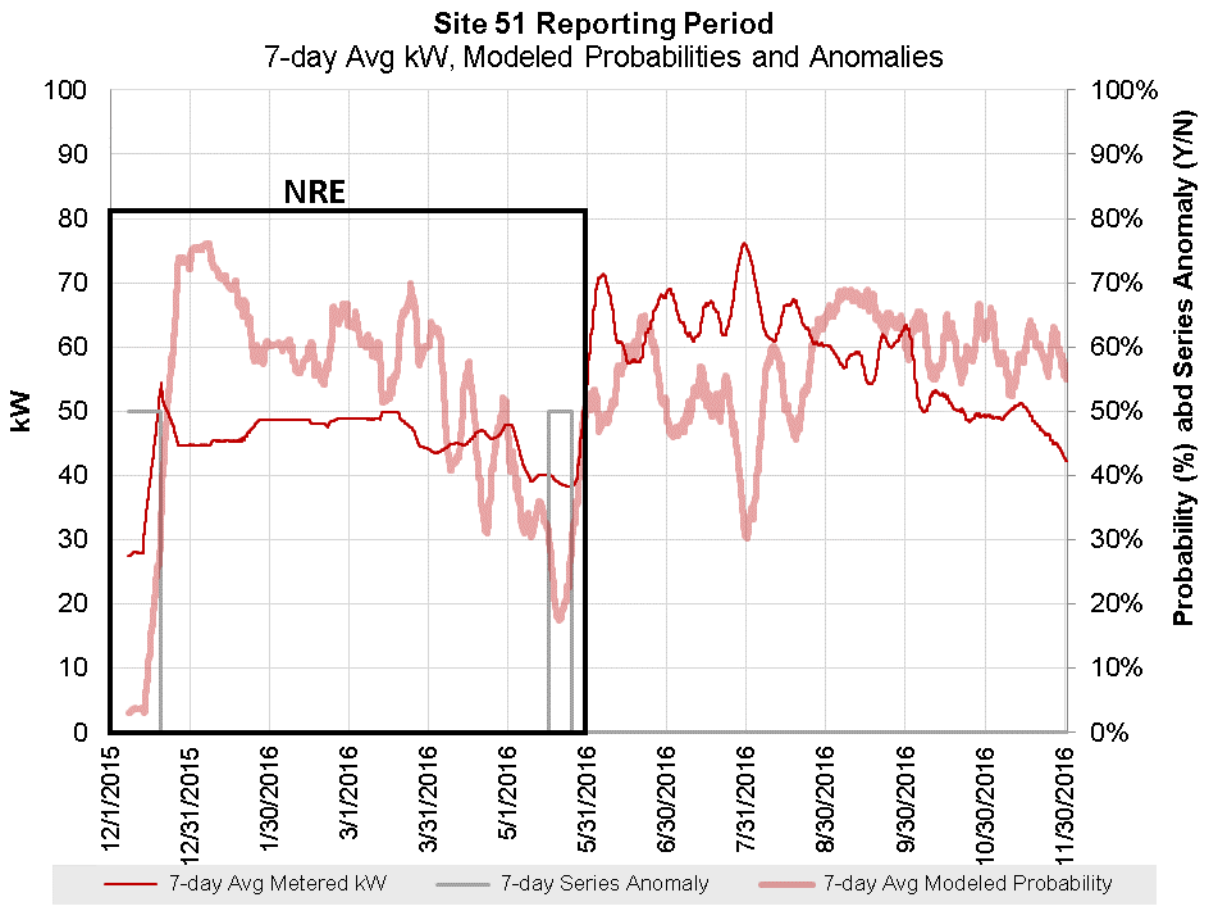


Figure 15: Site 51 – Reporting Period 7- day Avg Gas Usage, Modeled Probabilities and Anomalies

Based on the analysis at these two sites, the team concludes that data-driven, statistical approaches to detecting anomalies may be useful as a means for identifying errors in the raw data, errors in the model, and non-routine events at the site. More work is needed to further develop and test these approaches.

Table 6: NRE Adjusted Savings

	Site	Initial Annual Savings	NRE Adjustment	Final Annual Savings Adjustment
Electric	24	1,422,582	-613,200	809,382
Electric	51	106,805	18,185	124,990
Gas	44	33,057	-3,415	29,642

Identification and Treatment of Long-Term Trends in Baseline

Energy Usage

We observed significant daily baseline trending for ten of the fifteen SBW electric models and six of the twelve SBW gas models. Table 7 and Table 8 provide our adjusted savings estimates for each site after accounting for daily trending. Statistically significant baseline trends are highlighted in red font. We only adjusted the savings estimates for trending when the baseline trend was statistically significant.

Table 7: Trend adjusted electric savings

Site	Building Type	Original Baseline (kWh)	Original Savings (kWh)	Baseline Trend %	Trend-Adjusted Baseline (kWh)	Trend-Adjusted Savings (kWh)
14	Grocery	1,901,652	471,480	-3%	1,932,466	440,666
16	Grocery	1,309,249	224,630	-7%	1,355,490	178,389
21	Grocery	1,717,176	262,351	3%	1,687,959	291,567
42*	Grocery	1,294,692	421,847	-3%	1,288,001	428,537
63†	Grocery	1,300,350	192,578	-1%	1,300,350	192,578
1	Office	1,465,455	429,747	-7%	1,512,628	382,574
24	Office	4,788,733	1,422,582	8%	4,597,809	1,613,506
44	Office	2,284,886	824,052	4%	2,234,512	874,426
51	Office	549,119	106,805	-1%	549,119	106,805
54	Office	811,176	286,947	-6%	835,233	262,890
60	Office	487,197	142,413	2%	487,197	142,413
52	Library	332,907	184,392	-12%	354,201	163,098

*Site 42 had 4 meters with different trends for each model. Meters A, B, and C had statistically significant baseline trends of 1%, -7% and -7% respectively. The savings estimates were adjusted from 409,543 to 417,670 kWh for meter A, 13,279 kWh to 12,156 kWh for meter B and from -959 kWh to -1,273 kWh for meter C. In this table, site 42 summarizes the trends of all 4 meters.

†Site 63 had two meters, neither of which had statistically significant baseline trends. The overall savings for site 63 was not adjusted for trending.

Table 8: Trend adjusted gas savings

Site	Building Type	Original Baseline (therms)	Original Savings (therms)	Baseline Trend %	Trend-Adjusted Baseline (therms)	Trend-Adjusted Savings (therms)
14	Grocery	43,419	-6,927	2%	43,419	-6,927
16	Grocery	26,443	-3,817	17%	24,308	-1,682
21	Grocery	33,135	-448	2%	33,135	-448
42*	Grocery	7,386	-583	10%	7,002	-199
24	Office	4,649	574	-13%	4,649	574
1	Office	11,460	9,305	-27%	13,261	7,504
44	Office	58,519	33,057	1%	58,519	33,057
51	Office	25,592	5,913	-18%	27,777	3,729

54	Office	690	-2,262	-48%	690	-2,262
60	Office	11,934	6,406	13%	11,227	7,113
52	Library	11,066	5,076	-11%	11,662	4,480

*Site 42 had two meters with different trending for each. Meter A had a statistically significant baseline trend of 20% while meter B did not. The resulting savings estimate adjustment for meter A changed from -1,673 kWh to -1,290 kWh. In this table, site 42 accounts for the adjustment of the meter A trending.

Recommendations

Model input data should be standardized so comparisons between vendor models are consistent.

1. All parties modeling savings for any given project should use identical weather data provided by a source agreed to in advance and the source should be documented with the model.
2. Standardize preparation of interval data for use in data-driven models to minimize anomalous or missing points and provide consistency and transparency for all participating parties. This could include providing cleaned interval data which has been annotated such that all changes from the raw interval data can be traced. Additionally, the data cleaning protocol should be publicly available.
3. The Program Administrator (PG&E or third-party) should define the duration of baseline, implementation, and reporting periods, and for each work with the implementer to set the start and end dates of each period to be used for modeling and estimating savings. These dates should be stated in the project documentation.
4. Models should be checked for modeling errors and physical relevance.
5. Do not use CV(RMSE) as a pass-fail criterion for NMEC models when the use can approach zero for significant periods of time. This may include gas meters, electric meters for sites with renewable energy, or meters for net zero buildings.
6. Use transparent, open-source algorithms to facilitate evaluation of the savings estimates.
7. For buildings with weather-dependent energy consumption, use model packages that include weather (outside air temperature) as a dependent variable, such as Time of Week and Temperature (TOWT) or ECAM.
8. PG&E should examine what caused the increased gas use at grocery stores.

There are clearly many opportunities for improvements in handling data anomalies and non-routine events. In particular, detecting and adjusting for non-routine events will be critical for establishing confidence in savings from NMEC programs. Below we describe recommendations for improving detection and accounting of non-routine events.

9. Continue to explore statistical methods for auto-detection of non-routine events in the interval data like those used in this task.
10. Even after robust data-driven methods for automated detection of non-routine events have been developed and put into use, implementers must remain engaged with participating

building operators to monitor for non-routine events. This will provide corroboration of the data-driven detection, and also identify NREs when data-driven methods are unlikely to be successful, such as during implementation, or for continuous improvement programs such as SEM.We recommend the following

- Establish a checklist of important changes in building operations to monitor on an ongoing basis, including tenant or space use changes, significant equipment or operation changes such as scheduling and set points, and addition or removal or large loads such as data servers.
- Require regular reporting intervals, e.g., quarterly, in which the checklist above is completed before significant changes are forgotten.
- Provide regular access to interval data and require implementer to run the avoided energy use model early in the reporting period and at least at each reporting interval to see if reductions in energy use are as expected. If not, seek to explain why and make any necessary adjustments. This is not only helpful for detecting non-routine events, it can help identify energy efficiency measures not operating as expected.
    - For example, if modeling in ECAM, the color-coded standardized residuals on the Summary worksheet and the heat map feature can be used to visually identify anomalies and potentially non-routine events.

11. If the NRE cannot adequately be accounted for with statistical methods, careful consideration should be given as to whether it is appropriate to estimate the non-routine adjustments with engineering calculations.

- Statistical methods can determine the range of uncertainty around the estimate while engineering calculations generally do not.
- Further study is needed to understand whether statistical approaches and engineering calculations produce similar savings estimates. Until then, we advise against mixing statistical and engineering estimates when statistical methods can be used.

12. Ideally the interval data that is provided by the PA to program vendors, implementers, and evaluation consultants should cover the full period of time being investigated, and should be complete, without empty intervals, and without duplicates. The problem of anomalous spikes, both zeros and non-zeros, should continue to be investigated. Anomalous intervals that cannot be validated should be clearly identified and tagged as such before the data is released. Furthermore, data provided to vendors and implementers should be retained with project documentation to ensure the same data can be provided to evaluators, who will be looking to reproduce the results from the vendors and/or implementers. Ultimately, it would be most helpful to evaluators if they were given sufficient confidence in the completeness, validity, and consistency of the data that the need for guessing about its quality is minimized.

We recommend further study of trending in the baseline period, specifically the following:

1. For a more thorough assessment we would first check whether the trend was due to non-routine events or seasonality of energy use. If more than one year of pre-program data was

available, we would model different years to see if the trend repeated each year, indicating it is seasonal. Also, we would review the treatment of holidays in the models. If holidays were included as part of the same model as other days, we would check to see if the trend persisted with holidays excluded. If the trend of the residuals was not due to seasonality effects, we would then assume that it reflected a real trend in the baseline energy usage.

2. Once the industry has reached consensus the treatment of trending, efforts should be put into development and testing of automated trending detection and savings adjustment for sites with significant trending.
3. Interviews with building operators my help the analysist to understand the reason for the observed trend.

IPMVP Option C and Option D Methods Comparison

This section describes the findings from our review of the Option D models as well as our comparison of the Option C (ECAM) models to the Option D models. The goal of this section is to provide program administrators and implementers with the advantage and disadvantages of the Normalized Metered Energy Consumption (Option C) based savings estimation approach as compared to calibrated simulation (Option D).

Option D Model Review Findings

SBW carefully reviewed the implementer's Option D models for eight of the Demo sites. We analyzed issues by various breakdowns, including by site, end-use, issue type (scheduling, controls, efficiency, etc.), and qualitative level of impact to savings. We were able to make the following general observations.

By Site/Building Type:

- We identified more issues at grocery sites.
- We discovered a larger concentration of medium/high impact issues at non-grocery sites.
- Site 21 (a grocery) in particular had a larger number of issues, 34 in all.
- We considered only a couple issues to have high impact on savings. One was due to the use of DEER grocery store template model inputs instead of site-specific inputs. The other issue stemmed from insufficient documentation and poor modeling of three rooftop units that accounted 23% of the cooling tonnage in the model.

Figure 16 shows the distribution of issues by building type, savings impact, and fuel.

	Count of Issues			
Building Type Savings Impact	Electric Only	Gas Only	Both Fuels	All
Grocery	**53**	**7**	**39**	**99**
(1) Low	42	4	34	80
(2) Medium	11	3	4	18
(3) High			1	1
Office	**16**	**13**	**33**	**62**
(1) Low	11	7	20	38
(2) Medium	5	6	12	23
(3) High			1	1

Figure 16: Qualitative Savings Impact by Building Type and Fuel.

By End-Use:

- The HVAC end-use had the most issues, and the highest number of issues in each category (low, medium, high impact).
- Envelope and Refrigeration end-uses also had a significant number of issues.
- Fewer issues existed for lighting, domestic hot water, plug load, and exterior use end uses.
- Very few issues significantly affected multiple end-uses.

By Issue Type:

- Most issues had to do with controls setpoints
- We also found a significant number of issues surrounding scheduling, equipment capacity, building envelope inputs, and wattage inputs (i.e. lighting or plug load wattage).

Figure 17 shows the frequency of issues by end use and issue type.

End Use Type of Issue	Count of Issues
HVAC	**72**
Controls Setpoint	24
Capacity	11
Efficiency	8
Schedule	7
Orientation	5
Weather	5
System Type	4
Controls Routine	2
Envelope-Infiltration	2
Muptiple	2
Other	1
Quantity	1
Envelope	**26**
Envelope-Construction	15
Muptiple	7
Envelope-Infiltration	3
Zoning	1
Refrig	**24**
Muptiple	11
Quantity	7
Controls Routine	3
Other	2
System Type	1

End Use Type of Issue	Count of Issues
Plug/Process	**11**
Schedule	6
Wattage	5
Lighting	**10**
Wattage	5
Schedule	3
Zoning	1
Controls Setpoint	1
DHW	**8**
Capacity	7
Controls Setpoint	1
Exterior Use	**7**
Wattage	6
Schedule	1
Multiple	**3**
Muptiple	2
Envelope-Area	1

Figure 17: Count of Issues by End Use and Issue Type

In many cases across building type, end use and issue type, the documentation was not sufficient to adequately evaluate aspects of the model without further input from the project implementers. Throughout the course of our reviews, we noted patterns of issues occurring across multiple sites. Table 9 lists the most significant of these patterns.

Table 9: Project-Wide Patterns Found

Observation	Impact to Model
Building orientation was incorrect for several sites (the building was modeled facing the wrong direction relative to actual orientation).	Reduces accuracy of solar load calculations.
In many cases, model simulation period did not match reporting period.	This results in a misalignment between simulation days and actual weather days.

Observation	Impact to Model
Grocery models were based on DOE grocery model templates. Many template values were not updated to reflect actual building conditions.	Use of templates introduces inaccuracies into the model, since the model then represents template systems and controls instead of the systems and controls present in the actual building.
Weather issues were found at four sites. Some calibration weather files did not match reporting period dates. One site used an incorrect weather station. Normalized weather source was inconsistent (some used CZ2010, others used TMY).	This results in a misalignment between simulation days and actual weather days. Results in inaccurate weather-based loads. Introduces unnecessary differences between models.
Model baselines were based on calibrated baselines for grocery sites, but non-calibrated baselines for non-grocery sites.	Creates a bias between the Option D results within the pilot.

Nearly all the Option D models inaccurately modeled the daily energy consumption profile. This is not surprising, considering that Option D models are calibrated to monthly data instead of smaller interval data. As Figure 18 demonstrates, the modeled reporting (yellow line) period use matches the modeled baseline use (blue line) during unoccupied hours, yielding no savings. On the other hand, the modeled baseline is substantially higher than metered use (gray line[16]) during occupied periods, overestimating savings. The week depicted in the chart is typical of the rest of the year for this site[17]. In several sites, the modeled demand was higher than actual during peak periods and lower than actual during non-peak periods. This is likely attributable to inaccuracies in the modeling of the HVAC systems, especially considering that we found the most issues in the HVAC end use during the detailed model reviews.

[16] ECAM closely models the baseline period and was left off the chart to reduce clutter.

[17] Also observe that the Option D (eQUEST) model results are shifted by a day, i.e., the weekend occupancy pattern occurs on Sunday and Monday. eQUEST does not have a provision to account for leap day. This week is from April 2016, after leap day that year.

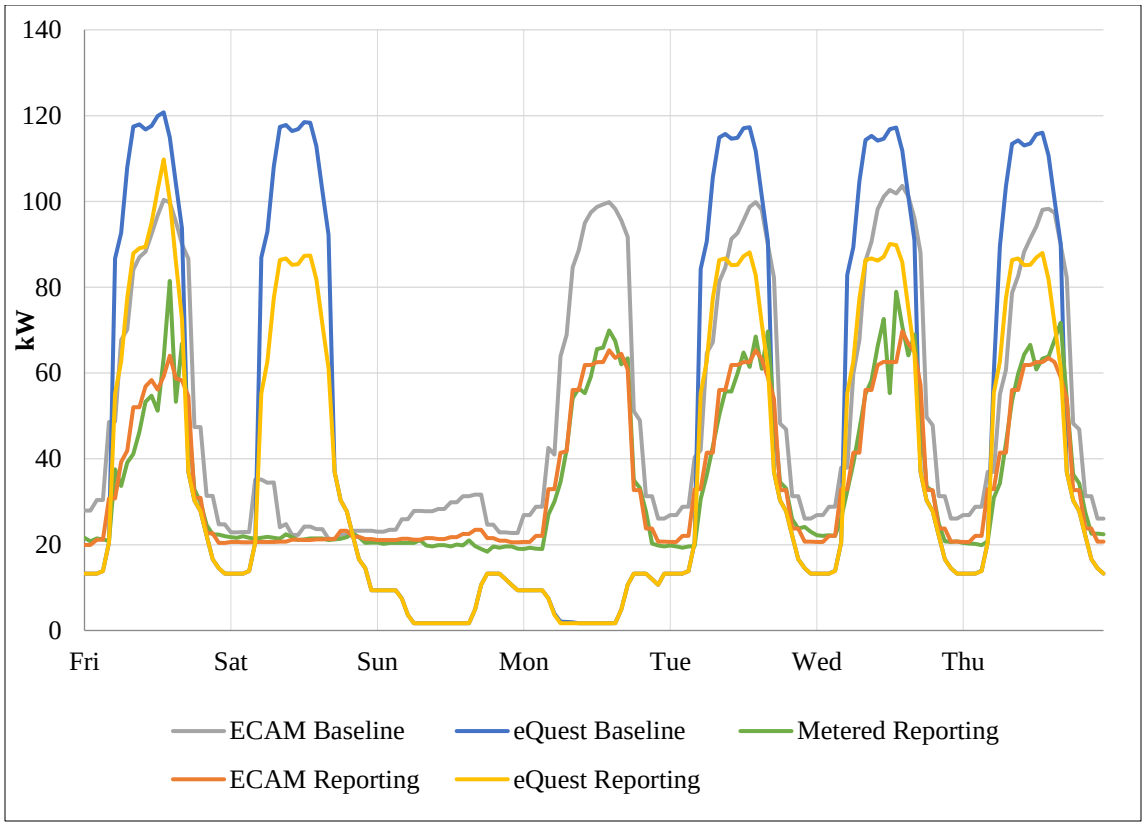


Figure 18: Comparison of Hourly Modeled and Metered kW

Option C and Option D Comparison Findings

The purpose of comparing the Option D models to the SBW-developed Option C models was to identify reasons for differences in the savings estimates with a focus on looking for opportunities for improvements in modeling, documentation, and evaluability. Below, we summarize our observations from the comparisons.

1. Non-routine events can have a large impact on estimated savings. The implementers inadequately documented their occurrences such that the study team could not confidently account for them in the Option C models.
2. Option D is significantly more expensive than Option C for comparable level of rigor as shown in Table 10. See Appendix 0 for a description of how we derived the costs estimates.

Table 10: Estimated Costs of Options C and D

	Estimated Pilot Cost/Site	Estimated Evaluation Cost/Site
Option C	$7,600	$7,600
Option D	$20,000	$24,088[1]

[1] Cost estimate based on estimated pilot cost per site plus the increased level of rigor that we recommend based on issues found with the models. The improvement costs were estimated using an assumed cost per improvement type based on past experience.

3. There were inconsistencies in actual energy consumption that the Option C and Option D models were built upon. In some cases, the metered data provided to us did not match the data used by the implementers. Differences also stemmed from inaccurately modeling the building schedule and from differences in how eQuest assigns the day type (weekday, weekend, holiday, etc.).
4. Option D's back-casted baseline energy consumption can be quite different than Option C's adjusted baseline energy consumption. After ensuring that both models used the reporting period weather data, the Option D baseline energy use was frequently substantially different from the Option C adjusted baseline energy use, particularly for the non-grocery sites. Since both models used the same weather data this should have been an apt comparison.
5. At three of the grocery sites, 16, 42, and 63, the study team observed that the implementer savings estimates may be artificially high due to bias. For grocery stores, the implementers calibrated both baseline and reporting period models to actual conditions, targeting less than 5% bias annually. At these three sites, the baseline period consumption biased high while the reporting period consumption biased low, yielding larger differences in consumption between the baseline and reporting periods. Since the study team's Option C models had zero bias annually, there was no impact on its savings estimates from bias. The impact of the bias errors in the implementer's estimates alone are sufficient to explain the differences between the study team savings estimates and the implementers estimates, though other offsetting factors may also be at play.
6. The *number* of issues found in an Option D model does <u>not</u> correlate with the *magnitude* of difference in estimated savings from the Option C model. For example, Site 51 had the lowest number of issues found yet among the highest percentage difference in the savings.
7. Option D models may be the better alternative to an Option C model for estimating whole building savings if the baseline needs to be adjusted from existing conditions, such as a codes and standards baseline.
8. Non-Routine Event Detection: With the application of statistical analysis to interval data, Option C models calibrated to hourly data facilitate the detection of non-routine events better than Option D models calibrated to longer time intervals, except for NREs that occur during the construction period and persist in the reporting period. NREs may be detected during the calibration of an Option D model but could be missed if only calibrating to monthly data. In either case, NRE detection depends largely on the type and magnitude of the NRE, and how well the model is calibrated or fits the data.

   In both options, NREs can affect the savings in either direction. An NRE that increases energy consumption in the baseline period would potentially result in an over-estimate of the savings if the NRE is not taken into account. Likewise, an NRE that increases energy consumption in the reporting period would potentially result in an under-estimate of the

savings. The latter is the case for the Site 44 Option C model. An NRE caused the gas consumption to increase during the last three months of the reporting period. The post period Option C model included the increased gas consumption from the NRE, thereby reducing the savings estimate substantially. The Option D implementer dealt with this by shifting the reporting period three months to exclude the NRE. While this method did exclude the NRE from the reporting period, it is possible that the shifted reporting period overlapped with the implementation period. Any measures still undergoing installation or commissioning during this period would have affected the energy consumption, and therefore the shifted post period may not accurately represent the true post period energy consumption.

9. Non-Routine Event Impact Quantification: Regarding which Option better facilitates quantification of non-routine events, our findings are inconclusive. The answer to this question requires additional research and improved documentation practices (please refer to our recommendations regarding documentation). Option C is much less expensive than Option D and provided more consistent results across the four vendors included in this study. The cost of estimating non-routine adjustments are not considered here as our findings are inconclusive.

Recommendations

Based on our detailed Option D review findings and comparison of Options C and D, we recommend the following. We include in this list of recommendations suggested improvements for both simulation-based savings and meter-based savings.

1. Program administrators should emphasize to NMEC program implementers the importance of documentation of changes in participating buildings that affect energy consumption, both from program interventions and non-routine events. We recommend that program administrators require a simple regular check-in, perhaps quarterly, between the implementer and participant throughout the reporting period to monitor for unexpected changes in the participating buildings and their energy use. The check-in would be a time to update a log of changes in the building and compare the actual energy use to the adjusted baseline to see if savings are in line with expectations.
2. Program administrators should ensure that meter and weather data are from consistent sources across projects
    a. Standardize method of delivering clean utility data. The utility data should include all relevant meters for the participating buildings (electric, gas, renewable or other) and accurately represent the energy use of the whole building.
    b. Establish the source for TMY-type weather, e.g. all CZ2010 or all TMY3, etc.
3. Program administrators and implementers should require consistent modeling methodology across all projects and that the models are fully documented. This includes firm baseline and reporting period start and end dates. Goodness of fit criteria should be appropriate for the type of meter being modeled and the modeling interval. Goodness of fit criteria should be by a criterion for savings precision, at least when the industry has a

consensus approach for estimating uncertainty. The data sources and dates of data acquisition should be documented. Modelers should state modifications and exceptions made to models to account for non-routine events in the building or missing data.

4. In cases where Option D is used, implementers and technical reviewers should prioritize simulation efforts on the problem areas we identified to reduce potential for issues in the future. We found most issues with grocery sites, HVAC inputs, and inputs related to controls setpoints. Furthermore, we recommend emphasizing a reliance on actual building conditions whenever possible instead of relying on template values, typical values, or defaults. This should greatly improve model accuracy, and by extension, savings reliability. Actual conditions can be gleaned from as-built drawings, trend data, nameplate information, control screens snapshots, on-site observations, and interview of knowledgeable site personnel.

Table 11 lists our recommendations for the CWB program based on the findings of our detailed model reviews and comparisons of Option C and Option D.

Table 11. Recommendations from Detailed Model Reviews

Recommendation	Notes
Focus Option D modeling efforts on problem areas to reduce potential for issues in the future.	Both modelers and technical reviewers should focus on the areas we identified. We found most issues with grocery sites, HVAC inputs, and inputs related to controls setpoints.
Ensure consistent modeling methodology across all projects.	This includes the baseline methodology and general model inputs such as correct simulation period and building orientation.
We recommend emphasizing a reliance on actual building conditions whenever possible instead of relying on template values, typical values, or defaults.	This should greatly improve model accuracy (and by extension, savings certainty). Actual conditions can be gleaned from as-built drawings, trend data, nameplate information, control screens snapshots, on-site observations, and interview of knowledgeable site personnel. Additionally, load profiles can be generated from interval data to inform schedules and daytypes.
Standardize method of delivering utility data for calibration. Ensure that it includes all relevant meters (electric, gas, renewable or other) and that it accurately represents the energy use of the whole building during the reporting period.	
Ensure that meter and weather data is from consistent sources across projects (Option C/D meter data from same source and standard TMY-type weather uses the same source, i.e., all CZ2010 or all TMY3, etc.).	
Do not calibrate Option D simulations to both reporting period and baseline conditions, unless the criterion for model bias is much more stringent, i.e. $\ll$1%.	Calibrating to both the reporting period and the baseline period can result in a large bias in the estimated savings. Calibrating to only reporting period conditions leaves an implicit assumption that baseline bias is the same as the reporting period bias, and the bias nets to zero when estimating savings.

Program Material Technical Review and Evaluability Assessment

The focus of this section is on our findings from review of program and project-specific materials and to provide recommendations for improving evaluability of NMEC programs. We also included findings and suggested improvements from our review of the Option D models.

Technical Program Documentation

The following summarizes key strengths and areas for improvement in documentation at the project and program levels.

Option C (Project Level):

- Savings estimates built on actual metered energy use of the building which has been through quality assurance for billing purposes.
- Critical shortcomings of the PG&E CWB Demo Option C documentation and data included inadequate reporting of non-routine events, lack of clarity in weather data used for analysis, and inadequate documentation of sources for various model inputs.
- Improvements that should be made to the project-level documentation prior to a full program launch for Option C include a data quality check from PG&E for gaps and erroneous values, better documentation of statistics for comparison to the actual results, clear documentation of the weather data used for analysis, and a quality check to ensure that vendors have included all required documentation when uploading information to ESFT.

Option D (Project Level):

- Key strengths of the project-level documentation provided by the Option D implementers were that they provided the eQuest model input/output files and some supporting documentation for the inputs used.
- Critical shortcomings of the project-level documentation for Option D included lack of documentation for model input sources and reasoning behind modeling procedures, and inconsistent procedures used (procedures manual not followed).
- Improvements that should be made to the project-level documentation prior to a full program launch for Option D include a detailed check of implementer reporting to ensure that all model inputs are adequately documented and that modeling methodologies are explained clearly.

Program Level:

- Key strengths of the program-level documentation include the requirement for a data quality check from PG&E and a requirement to provide the actual data (including actual weather data) used in the analysis. Shortcomings of the program-level documentation include a lack of adherence to these requirements, an inadequate explanation of the data cleaning process, and poor instruction on how to calculate key statistics for comparison to the actual results.

- Improvements that should be made to the program-level documentation include clearly stating statistical calculations and responsible parties, explanation of the data cleaning process, and a process for ensuring that required quality checks are performed.

Documents Reviewed

Program Documentation:

- Baselining Field Test Procedure Work Plan (1/12/2015)
- Commercial Whole Building Program – Fact Sheet (September 2013)
- Commercial Whole Building Program: Data and File Specifications, and Workflow Requirements (7/27/2015)
- Commercial Whole Building Performance – Procedures Manual, Version 1.3 (11/6/2015)
- PG&E Commercial Whole Building Demonstration: Study Process Overview and Plan, Version 1.4 (5/18/2016)
- Commercial Whole Building Performance – Program Manual, Version 1.2 (3/31/2014)
- Commercial Whole Building Program: M&V Analytics Output Procedures (4/14/2016)
- Commercial Whole Building Program: Field Test Plan and Procedure (9/24/2015)
- Commercial Whole Building Performance Measure Codes: Various Measures/End uses, Revision #0 (8/25/2015

Option C Documentation:

- Vendor Option C models
- Vendor model descriptions

Option D Documentation:

- Implementer Project VR2 report
- Implementer Option D VR2 level eQUEST models
- Available supporting documentation (Billing data, trend data, as-built drawings, etc.)

Recommendations

Table 12 provides the findings and recommendations for improving data and documentation, to better support evaluation of NMEC programs.

Table 12: NMEC Findings and Recommendations

Finding	Recommendation
The documentation provided by PG&E included very few reports non-routine events. In some cases, neither the implementers nor Option C modelers included apparent non-routine events in the reports. This may have been because the events were not identified by the Option C modelers during the performance period, or because the implementer was unable to determine why the NRE occurred. There was no indication of removal or treatment of non-routine events by any of the Option C modelers.	Non-Routine Event identification should be well-defined such that vendors and implementers can consistently and reliable identify, characterize and adjust for NREs. Detection of NREs could occur in multiple ways: customer identification during the survey, detection while preparing meter data Demo Option C modeling, and implementer investigation with customer following detection from meter data. Vendors should have a process to establish a model that excludes the effects of NREs in their savings estimates. If vendors treat the NREs differently in their model predictions, the treatment should be documented.
Some Demo Option C models have seemingly obvious irregularities, such as large jumps in estimates of energy use. These irregularities were not addressed by the vendors or technical consultant.	Vendors should check their models for obvious irregularities before submitting. If irregularities are observed, the vendor should include a discussion of irregularities in the model as acknowledgement of the model issue.
The energy use data provided by PG&E did not always match the data used by the Demo Option C modelers.	PAs should ensure that the same data is provided to all participating parties.
According to the Data and File Specifications, and Workflow Requirements, vendors should provide: the actual weather data used for modeling and NREs and anomaly flag history, actual electric/gas data measured by the meter, the baseline electric/gas use prediction, the performance period electric/gas use prediction and estimated savings, and the baseline model and energy savings stats. Some Demo Option C models did not provide the weather data used for modeling or the baseline period estimates, this was an issue with the proprietary models more than the open source models.	PAs should ensure that vendors have provided all required documentation.
According to the Procedures Manual, PG&E should conduct a data quality check for significant gaps and erroneous values in the energy use and independent variable data. In many cases, we observed repeated timestamps in the dataset which had to be removed for modeling. We also noticed large gaps in the text files received from PG&E detailing the 15-minute interval data from the billing meters associated with the demo sites. PG&E may have conducted a quality check, but it is unclear if PG&E made changes to the data before distributing to the vendors.	The interval meter data quality check should be described in more detail including whether the data was adjusted or cleaned by the PA when an irregularity was observed in the data, and how the irregularity was handled.
The documentation does not explain the data cleaning process. Each vendor is likely to treat data gaps and outliers differently.	A procedure for data cleaning should be included in the documentation so there is consistency across the vendor model inputs. Analysts should exclude zero values and clear outliers from the development of the baseline and performance period models. Non-zero outliers can be identified using scatterplots developed for different parts of the model.

Finding	Recommendation
The weather station used in the Option C models was not documented and frequently difficult to determine. Consequently, SBW used different weather data than that of the Demo Option C models, challenging reproducibility of results.	The PA should include the weather station data with the delivery of the meter data, so all parties have the same input data for modeling.
Each of the Demo Option C modelers provided their own calculated values for the statistics: normalized root mean squared error (n(RMSE)), mean absolute percentage error (MAPE), and absolute percent bias error (APBE). Several of the documents explain the calculation method for each of these values and indicate that the Option C modeler is responsible for their calculation. The Procedures Manual indicates that the baseline model predictions should be validated against the original baseline model data using the n(RMSE) and R2 metrics, and the net determination bias error of the model should also be calculated. The Procedures Manual does not include a description of how to calculate the net determination bias or R2 value, and it is unclear who should be calculating the values.	Statistical calculations and the responsible parties should be clearly stated in the program manual. Statistics that provide goodness of fit metrics are: net determination bias error, R2, and CV(RMSE).
The Procedures Manual called for PG&E to define the baseline and post period dates before sending the data to the vendors for modeling. Despite the definition of these dates, we found a vendor model that the baseline period to begin two months after the defined PG&E baseline period.	Vendors should use the baseline and post period dates defined by the PA or its implementer for their models.

Option D findings and recommendations

1. The implementers did not provide sources for all inputs used in the models. *Implementers should provide a source for the most influential inputs or justification for assumptions made. This will greatly improve the evaluability of the model.*
2. Implementers did not fully explain modeling methodology in many cases (i.e. why a particular system or measure was modeled in the way it was). *There should be a section in the VR2 report that explains the overall modeling methodology (i.e. how was the model built and why was it built that way). Implementers should explain in detail any model workarounds that were used to simulate non-conventional systems or conditions. This will greatly improve the evaluability of the model.*
3. In some cases, implementers did not explain what changes were made between the implementation and post-implementation (VR2) models. *Implementers should provide a log of changes made at different levels of model development (calibrated baseline to calibrated reporting period model) so that the history of the model can be fully understood by evaluators.*

Appendices

A. Technical Modeling Details

This appendix provides more detail on Model Anomalies, Data Cleaning, and Goodness of Fit Criteria.

Statistical Identification and Causes of Model Anomalies

While developing our models in ECAM, we observed many data points to be outliers, unexpected zeros, or missing based on looking at time-series charts of the raw meter data or model residuals. This section describes how we systematically characterized anomalous data points. By analyzing the variation in the residuals of the ECAM model, we identified points in time where the actual (measured) energy consumption falls outside the confidence interval of the predicted (modeled) energy consumption, for an input confidence level that points outside the interval are outliers. (This is separate from the confidence interval used for savings; the confidence that a point is an outlier must be much higher.)

Figure B.1 shows, for site 1, a model of weekday hours from 8AM to 5PM during the reporting period. The prediction interval for points is shown as the narrower light blue lines. At the 90% confidence level, 90% of the points will fall within the interval bounded by those lines. A prediction interval to exclude outliers is shown as the dark red lines. At the 99.9% confidence level, 99.9% of the points will fall within that. In other words, the points outside of that interval, highlighted in bright green, are only 0.1% likely to belong to this set of data.[18]

[18] Note that using the assumptions associated with ordinary linear regression results in discontinuities in the intervals at the lower change point of 71 ºF. Other approaches to defining confidence intervals could eliminate these discontinuities, but are usually more complex. See oaktrust.library.tamu.edu/bitstream/handle/1969.1/152310/ESL-IC-14-09-11a.pdf and http://www.iepec.org/2017-proceedings/polopoly_fs/1.3718217.1502901133!/fileserver/file/796649/filename/024.pdf for further information.

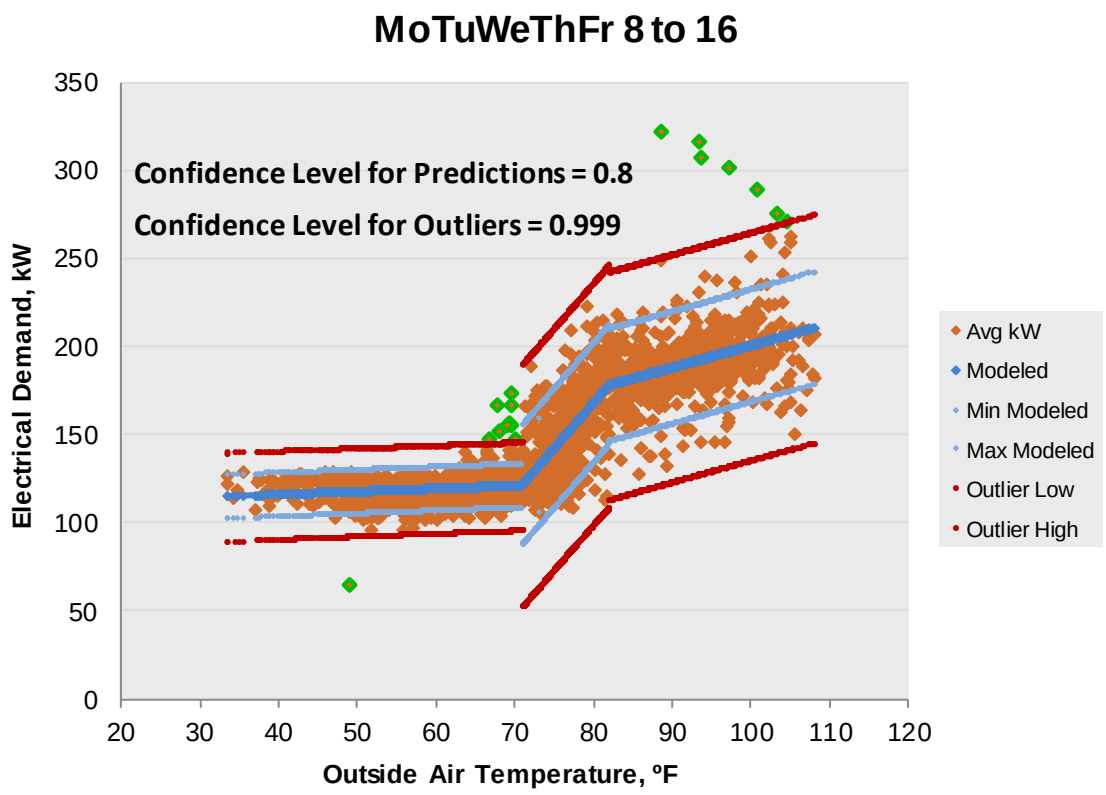


Figure B.1: Confidence Intervals for Model Predictions and Outliers

These outlier points can be thought of as *anomalies in the model*. Anomalies can be caused by errors in the raw data, errors in the model, or non-routine events at the site.

We identified two types of anomalies in the models. An "outlier" occurs when the probability of the t-score of a residual falls outside a prescribed confidence interval. A "series anomaly" occurs when the time-series rolling average of the probabilities drops to a relatively low level compared to the overall historical trend. We used the following methods to identify these anomalies.

- **Tracking of model residuals and their t-scores**

  We visually tracked residuals and t-scores using heat maps and charts intrinsic to ECAM. Heat maps are tables of data where each entry is highlighted using a color scale based on the relative magnitudes of the entries. ECAM calculates the t-score of a residual by dividing each residual by the standard error (RMSE) of the predictions using "Student's" t-distribution instead of the normal distribution. The standard error is calculated within the local temperature range (i.e. left or right of a change point) of the sub-model (e.g. daytype or time of day) from which it was derived. (These t-scores are also known as standardized residuals, although statistics references don't seem to be consistent in terminology between t-scores, standardized residuals, and studentized residuals.) Since the overall average of the residuals in an ECAM model is effectively zero[19], the anomalies of a model can be found among those points for which the absolute values of the t-scores are relatively high.

- **Tracking of probabilities of t-scores**

  An "outlier" occurs when a point falls outside a prescribed confidence interval. ECAM provides a tool for identifying outliers in this manner by calculating the probability of a given t-score in the context of the respective distribution of residuals in the corresponding local temperature range of the sub-model from which it was derived and comparing it to a user-prescribed confidence level. (This user-prescribed confidence level is distinguished from what is described above in Section 2.1.1, which applies to the model and savings uncertainty.) We selected confidence levels of 98% and 99%, so any point with a t-score probability that fell below 2% or 1%, respectively, we identified as an outlier.

  We also applied a slightly different approach based on the statistical convention known as "Chauvenet's Criterion", which defines an outlier as any point for which the t-score probability falls below 1/(2N), where 'N' is the number of points that are used to generate the regression model[20]. Using Chauvenet's Criterion, rather than assuming one confidence level across all sub-models uniformly, we calculated a unique confidence level for each local temperature range of each sub-model.

- **Tracking of probabilities over time**

  A "series anomaly" occurs when the rolling average probability drops to a relatively low level compared to the overall historical trend. We calculated the rolling-average probabilities over three different time windows (24-hr, 48-hr, and 168-hr for electric models; 3-day, 7-day, and 28 day for gas models) to identify "series anomalies" by three

[19] Per ASHRAE, the net determination bias error for an Option C model should be no more than 0.005%. In ECAM, it is typically 0.000% (i.e. zero to three significant digits). While some of the residuals are positive and others are negative, the overall sum and average is effectively zero.

[20] Also referred to as degrees of freedom.

confidence levels (70%, 80% and 90%)[21]. We calculated the hours of the year that each of these confidence levels were not reached for each metric.

- **Observation of model versus meter scatter charts**

  We also used scatter charts of modeled versus metered data to visually identify outliers. A perfect model without any outliers would be revealed in such a chart if all the points fell precisely on a straight line with a slope of unity. By contrast, outliers are revealed where points fall furthest from such an imaginary line.

Causes of Anomalies

The causes of the anomalies may fall into three general categories:

1. Errors in the raw data such as incorrect usage or temperature data. For example, anomalies may appear if a sub-set of the raw data that was used to create the model is from the wrong site, or was "filled-in", or approximated.
2. Faults in the model specifications. For example, if the occupancy schedule in an hourly model is incorrectly offset by one hour at the end of the day, then that hour may repeatedly show up as a point anomaly. If holidays are underspecified or over-specified in the model, then those mis-specified days may appear as point outliers in a daily model, or as 24 series anomalies in an hourly model.
3. From a modeling perspective, Non-Routine Events (NREs) at the site are changes in site energy consumption that do not originate from changes in the independent variables used in the model**.** Anomalies may be attributed to NREs if they are not the result of faults in the model specifications or errors in the raw data.

Findings

We found outliers and series anomalies in the baseline and reporting period models and calculated the hours of the year that each of the threshold metrics were surpassed. We found no correlation between the frequency of outliers and the frequency of series anomalies.

Out of the different metrics tested, we determined that Chauvenet's Criterion was most appropriate for detecting outliers, and the 7-day rolling average probability below 30% was appropriate for detecting series anomalies. The following tables show the percentage of the year for which we observed these outliers and series anomalies to occur in each model.

Table 13: Electric baseline and reporting period model anomaly metrics

Electric Meter	BL 168-hr Avg<30% %Yr	BL Chauvenet Outlier %Yr	RP 168-hr Avg<30% %Yr	RP Chauvenet Outlier %Yr
1	0.2%	0.4%	1.2%	0.7%

[21] In a normal distribution of values, *one* value has about a 30% probability of falling more than 1.0 standard deviation from the mean, about a 20% probability of falling more than 1.3 standard deviations from the mean, and about a 10% probability of falling more than 1.6 standard deviations from the mean. Here we calculated the *average* of the probabilities over three different time windows.

Electric Meter	BL 168-hr Avg<30% %Yr	BL Chauvenet Outlier %Yr	RP 168-hr Avg<30% %Yr	RP Chauvenet Outlier %Yr
14	1.4%	0.3%	0.0%	0.4%
16	2.2%	0.3%	0.2%	0.3%
21	0.0%	0.4%	1.1%	0.3%
24	15.2%	0.5%	7.3%	0.1%
44	0.1%	0.6%	2.5%	0.2%
51	0.0%	0.6%	5.8%	0.8%
52	0.0%	0.5%	0.0%	0.9%
54	3.4%	0.1%	0.4%	0.7%
60	0.0%	1.6%	0.0%	1.3%
42a	9.7%	0.3%	13.1%	0.3%
42b	0.4%	0.6%	0.9%	0.4%
42c	0.0%	1.4%	0.0%	0.2%
42d	2.4%	1.3%	0.0%	1.5%
63a	0.0%	0.9%	0.4%	0.6%
63b	0.6%	0.2%	0.0%	0.7%
Average	2.2%	0.6%	2.0%	0.6%

Table 14: Gas baseline and reporting period model anomaly metrics

Gas Meter	BL 7-day Avg<30% %Yr	BL Chauvenet Outlier %Yr	RP 7-day Avg<30% %Yr	RP Chauvenet Outlier %Yr
1	5.8%	2.7%	1.1%	0.8%
14	8.2%	1.6%	5.5%	0.8%
16	10.7%	1.4%	1.9%	0.8%
21	6.3%	0.5%	12.9%	0.5%
24	5.2%	2.5%	3.6%	1.9%
44	2.7%	1.1%	17.3%	0.8%
51	10.1%	0.8%	10.7%	0.5%
52	12.3%	0.8%	4.9%	1.4%
54	1.9%	2.5%	10.7%	0.5%
60	5.2%	1.4%	0.5%	2.7%
42a	3.6%	2.5%	8.5%	1.9%
42b	2.7%	1.9%	0.8%	0.8%
Average	6.2%	1.6%	6.5%	1.1%

Interval Data Cleaning

This section describes the anomalies that we observed in the raw interval data and the steps we took to address them. PG&E provided us with 15-minute electric consumption data from a total of 16 billing meters serving the 12 Demo sites and covering a period of time beginning more

than one year prior to the installation of the EE measures and ending more than a year after. PG&E's initial package of interval data consisted of six text files labeled by calendar year from 2012 through 2017. Within each text file each row contained the 96 sequential energy consumption values (kWh) associated with one day for one billing meter.

Anomalies

We found four types of data anomalies in the raw interval data.

Missing Dates

We expected each of the six text files to include all of the interval data for all 16 meters for a single calendar year. In fact, for every meter we found gaps covering various lengths of time. Many meters were missing the whole month of February, 2013. Every meter was missing March 31 over multiple years. Examples of gaps we found for individual meters include:

- The entire reporting period
- Most of the reporting period
- 70 consecutive days of the intervention period

Duplicate Dates

Every day for every meter was provided in duplicate.

Empty Intervals within Dates

The four intervals during the "spring-forward" hour at the start of daylight savings were always empty for every meter, as expected. However, for 5 of the 16 meters there were also a total of 141 empty intervals interspersed randomly over 9 different days.

Anomalous Zeros and Spikes

Of the 16 electric meters, only two were connected to loads that were normally expected to drop effectively to zero (e.g. exterior site or signage lighting loads). Of the other 14 meters, we found a total of 287 anomalous zeros in the interval data.

Steps Taken

PG&E provided additional text files to fill in the missing dates. However, like the first set of data, these additional files also contained duplicates of every day for every meter. Moreover, within the final set of non-duplicate data there remained many intervals that were empty or contained anomalous zeros. We deleted duplicate dates, empty intervals, and anomalous zeros. We kept the non-zero anomalies, although it is likely many should have been removed. The presence of these has little effect on savings estimates, but they cause a minor increase in the estimated uncertainty.

Goodness of Fit Criteria

ASHRAE Guideline 14 requires that net determination bias error (or simply "bias") is less than 0.005%, Coefficient of Variation-Root Mean Squared Error (CV(RMSE)) should be less than 25% when 12 months of data are used in computational savings and no uncertainty calculations are included with savings reports. (ASHRAE Guideline 14 Section 4.3.2.1.) When uncertainty is

estimated, it must be less than 50% of the annual reported savings, at a confidence level of 68%. (ASHRAE Guideline 14 Section 4.3.2.2.) There is no ASHRAE requirement for the Coefficient of Determination (R^2), but it is generally preferred that it is greater than 0.7. A high bias indicates that the model has a tendency to underestimate or overestimate the savings achieved over the modeled period.

The bias is estimated across the entire baseline period, so it is possible that a low bias is an indication that the low estimates and the high estimates balance each other out over the course of the year rather than an indication of a model that estimates well over the entire study period. The CV(RMSE) describes how well the model fits the measured data; a low CV(RMSE) indicates a good model fit. The coefficient of determination will be equal to 1 if the modeled data and the metered data are exactly the same for every timestep. A low coefficient of determination indicates that the vendor modeled data does not replicate the metered data well *or* that the metered data is not highly related to the chosen independent variable(s).

Model Cost Estimate

PG&E provided the estimated costs for producing Option C and Option D models during the Demonstration. Note that program costs per site may be smaller than these estimates due to increased efficiency of program model processes over the Demo.

We estimated Option D cost per site using an approximate range of costs of between $17,000 and $23,000 per site (provided by PG&E). We used the median within this range which is $20,000 per site.

We estimated Option C cost per site using the following subtask estimates provided by PG&E:

1. Demonstration Data Management and Reporting ($742/site)
2. Customer Proposal Review ($1,588/site)
3. Project Savings Determination ($4,408/site)

Total cost per site (rounded): $7,600

B. Stakeholder Comments

a table detailing each stakeholder comment and how it was addressed in the final report

C. References

ASHRAE 2014. ASHRAE Guideline 14-2014. Measurement of Energy, Demand, and Water Savings, American Society of Heating Refrigeration and Air Conditioning Engineers.

Chauvenet, William. A Manual of Spherical and Practical Astronomy V. II. 1863. Reprint of 1891. 5th ed. Dover, N.Y.: 1960. pp. 474–566.

EVO 2016. International Performance Measurement and Verification Protocol, Core Concepts. Efficiency Valuation Organization.

BPA 2018. Bill Koran, SBW Consulting Inc. “Potential Analytics for Non-Routine Adjustments”

Reddy, T.A., and D.E. Claridge. 2000. Uncertainty of measured energy savings from statistical baseline models. International Journal of HVAC&R Research 6(1):3–20. http://auroenergy.com/wp-content/uploads/2016/05/2000_Reddy_HVACR_Uncertainty-of-Measurement.pdf

Lei, Yafeng & Reddy, Agami & Subbarao, Kris. (2011). The Nearest Neighborhood Method to Improve Uncertainty Estimates in Statistical Building Energy Models (ML-11-001). ASHRAE Transactions. 117. http://auroenergy.com/wp-content/uploads/2016/05/2011_Subbarao_ASHRAE-Trans_NearestNeighborhood.pdf

A. Shonder, John & Im, Piljae. (2012). Bayesian Analysis of Savings from Retrofit Projects (SA-12-003). ASHRAE Transactions. 118. https://www.researchgate.net/publication/268095106_Bayesian_Analysis_of_Savings_from_Retrofit_Projects_SA-12-003

Sun, Y. and Baltazar, J.C. (2013), Analysis and Improvement on the Estimation of Building Energy Savings Uncertainty, ASHRAE Transactions.

Baltazar, J.C.; Sun, Y.; Claridge, D. (2014). Methodologies for Estimating Building Energy Savings Uncertainty: Review and Comparison. Energy Systems Laboratory (http://esl.tamu.edu); Texas A&M University https://oaktrust.library.tamu.edu/handle/1969.1/152310

Granderson J., Touzani S., Custodio C., Sohn M.D., Jump D. and Fernandes S., 2016. Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings. Applied Energy, 173, pp.296-308 http://eta-publications.lbl.gov/sites/default/files/lbnl1005818.pdf

Koran, B., Boyer, E., Khawaja, M. S., Rushton, J., Stewart, J., 2017. A Comparison of Approaches to Estimating the Time-Aggregated Uncertainty of Savings Estimated from Meter Data. International Energy Program Evaluation Conference. http://www.iepec.org/2017-proceedings/polopoly_fs/1.3718217.1502901133!/fileserver/file/796649/filename/024.pdf

J. L. Mathieu, P. N. Price, S. Kiliccote and M. A. Piette, *Quantifying Changes in Building Electricity Use, With Application to Demand Response*, in IEEE Transactions on Smart Grid, vol. 2, no. 3, pp. 507-518, Sept. 2011. http://eta-publications.lbl.gov/sites/default/files/LBNL-4944E.pdf

D. Glossary

Accuracy: An indication of how close the measured value is to the true value of the quantity in question. Accuracy is not the same as precision. (From the BPA Regression for M&V: Reference Guide, available at https://www.bpa.gov/ee/policy/imanual/pages/im-document-library.aspx)

Avoided Energy Use: Predicted energy savings from installed energy efficiency measures at CWB project sites. Savings are estimated using billing meter data (IPMVP Option C) or engineering estimates (IPMVP Option D).

Backcast Savings: Compare the adjusted reporting-period energy consumption to the actual baseline-period consumption. The estimate is the result of a model of energy consumption fit to reporting-period consumption data, and applied at baseline-period conditions. (This

description is an excerpt from the Superior Energy Performance Measurement and Verification Protocol for Industry.) This type of model and process is not explicitly described in IPMVP. However, the most common use of IPMVP Option D is for situations where a baseline does not exist, e.g. new construction, so the approach is similar to that of a backcast.

Blended Model: A model PG&E created that aggregates avoided energy use estimates from multiple IPMVP Option C sources to select the most accurate estimate for each time period. The objective for creating the blended model is to generate greater predictive accuracy and to reduce the confidence intervals around each time period estimate.

Calibrated Simulation Model: A model that conforms to the requirements of IPMVP Option D.

Coefficient of Determination: (Also known as R-Squared (R^2)): R^2 is the measure of how well future outcomes are likely to be predicted by the model. It illustrates how well the independent variables explain variation in the dependent variable. R^2 values range from 0 (indicating none of the variation in the dependent variable is associated with variation in any of the independent variables) to 1 (indicating all of the variation in the dependent variable is associated with variation in the independent variables, a "perfect fit" of the regression line to the data). It is calculated as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Coefficient of Variation of the Root-Mean Squared Error [CV(RMSE)]: A measure that describes how much variation or randomness there is between the data and the model, calculated by dividing the root-mean squared error (RMSE) by the average y-value. It is calculated as:

$$CV(RMSE) = \frac{1}{y}\sqrt{\frac{\sum(y_i - \hat{y})^2}{(n-p)}}$$

Confidence Interval: The range of values expected to contain the true value with a specific probability. The probability is referred to as the confidence level. (From the BPA Regression for M&V: Reference Guide)

Confidence Level: A population parameter used to indicate the reliability of a statistical estimate. The confidence interval expresses the assurance (probability) that given correct model selection, the true value of interest resides within the proportion expressed by the confidence interval. (From the BPA Regression for M&V: Reference Guide)

Early M&V: An estimate of energy savings used to inform an energy savings claim made by utilities to their regulators.

Evaluability: An assessment of whether a program is well-positioned to be evaluated. An evaluability assessment provides recommendations for improving readiness for an evaluation that are based on examination of data accessibility, data accuracy, data dictionaries, impact analysis techniques, technical program documentation, and qualifications for participation in the program.

Forecast Savings: Compare the actual reporting-period energy consumption to the adjusted baseline-period energy consumption. The estimate is the result of a model of energy consumption fit to baseline period consumption data, and applied at reporting-period conditions. (This description is an excerpt from the Superior Energy Performance Measurement and Verification Protocol for Industry.) This type of model and savings is called "Avoided Energy Consumption or Demand" or "Reporting Period Basis" in IPMVP.

Homoscedasticity: (Also known as Homogeneity of Variance.) Within linear regression, this means that the variance of the dependent values around the regression line is constant for all values of the independent variable. (From the BPA Regression for M&V: Reference Guide)

International Performance Measurement and Verification Protocol (IPMVP): Defines standard terms and best practices for quantifying the results of energy efficiency projects that is published by the Efficiency Valuation Organization.

IPMVP Option C: Regression-based techniques for estimating avoided energy use in whole buildings or at a sub-facility level that use pre- and post-intervention billing meter data and account for variables including outdoor air temperature and operating hours.

IPMVP Option D: A technique for estimating avoided energy use in whole buildings or at a sub-facility level that uses engineering-based calibrated simulations.

Multicollinearity: A statistical occurrence where two or more predictor variables in a multiple regression model are highly correlated (there are exact linear relationships between two or more explanatory variables). Allowing multicollinearity in a model can lead to incorrect inferences from the model. (From the BPA Regression for M&V: Reference Guide)

Net Determination Bias Error: The percentage error in the energy use predicted by the model compared to the actual energy use.

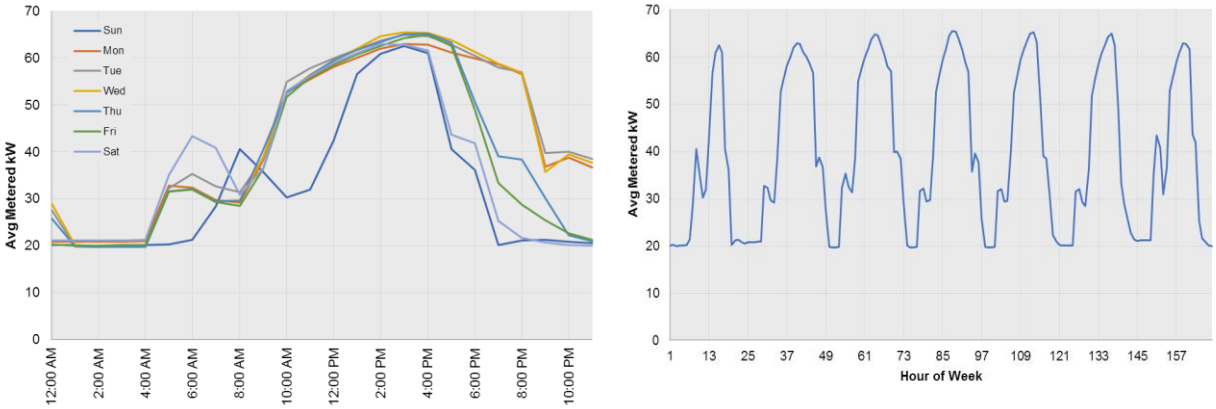
$$NBE = 100\frac{\sum_i (E_i - \widehat{E_i})}{\sum_i E_i}$$

Normalized Savings: Compare the adjusted reporting-period consumption to the adjusted baseline-period consumption. The adjusted consumption for baseline model is the result of a model of energy consumption fit to consumption data for the baseline period, and applied to the standard or "normal" conditions. The adjusted consumption for reporting period model is the result of a model of energy consumption fit to consumption data for the reporting period, and applied to the standard or "normal" conditions. This type of model and savings is called "Normalized Savings" or "Fixed Conditions Basis" in IPMVP.

Precision: The indication of the closeness of agreement among repeated measurements; a measure of the repeatability of a process. Any precision statement about a measured value must include a confidence level. A precision of 10% at 90% confidence means that we are 90% certain the measured values are drawn from samples that represent the population and that the "true" value is within ±10% of the measured value. Because precision does not account for bias or instrumentation error, it is an indicator of predicted accuracy only given the proper design of a study or experiment. (From the BPA Regression for M&V: Reference Guide)

Public Domain models:

The ***Mean Week (MW) model*** "predictions depend on day and time only. For example, the prediction for Tuesday at 3 PM is the average of all of the data for Tuesdays at 3 PM. Therefore, there is a different load profile for each day of the week, but not, for example, for each week in a month or each month in the year. This is a simplistic 'naïve' model that was intentionally included for comparative purposes." (Granderson 2016) This model can easily be shown graphically as different load shapes for each day of the week and hour of the week.



The **Time of Week & Temperature (TOWT)** model is a "linear regression-based load prediction method that includes two novel features: (1) a time-of-week indicator variable, and (2) a piecewise linear and continuous outdoor air temperature dependence derived without the use of a change-point model or assumptions about when structural changes occur." The model uses a coefficient for each hour of the week, plus two temperature relationships, one for relatively high-use hours, and one for relatively low-use hours. This "nonlinear temperature effect can be modeled with a piecewise linear and continuous temperature-dependent load model. For each facility, we divide the outdoor air temperatures experienced by that facility into" (up to) "six equally-sized temperature intervals." (J.T. Mathieu 2011) A separate coefficient for the energy use relationship to temperature applies to each of these temperature intervals, with separate temperature coefficients for the high-use and low-use hours. Essentially, the TOWT model takes a MW model and adds two temperature relationships to it.

Residual: The difference between the predicted and actual value of the dependent variable, i.e. the portion of energy use that is not explained by the model. Estimated by subtracting the predicted value (Xbar) from the actual value (Xi) in the data:

$$\hat{\varepsilon} = X_i - \bar{X}$$

(Excerpted from the BPA Regression for M&V: Reference Guide)

Root Mean Squared Error (RMSE): (Also known as the Standard Error of the Estimate.) An indicator of the scatter, or random variability, in the data, and hence is an average of how much

an actual y-value differs from the predicted y-value. It is the standard deviation of errors of prediction about the regression line. The RMSE is calculated as:

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} = \sqrt{E((\hat{\theta} - \theta)^2)}$$

Standardized Residual: (Also known as t-score.) A residual divided by its Standard Error (RMSE). This is a regression analog for a z-score, the number of standard deviations a value is away from the sample mean.

t-score: (see Standardized Residual)

t-statistic: A measure of the probability that the value (or difference between two values) is statistically valid. The calculated t-statistic can be compared to critical t-values from a t-table. The t-statistic is inversely related to the p-value; a high t-statistic (t>2) indicates a low probability that random chance has introduced an erroneous result. Within regression, the t-statistic is a measure of the significance for each coefficient (and, therefore, of each independent variable) in the model. The larger the t-statistic, the more significant the coefficient is to estimating the dependent variable. The t-statistic is calculated as:

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})}$$

Uncertainty (e.g. of Savings): The range or interval of doubt surrounding a measured or calculated value within which the true value is expected to fall within some stated degree of confidence. (See Confidence Level.) Uncertainty in regression analysis can come from multiple sources, including measurement uncertainty and regression uncertainty. (From the BPA Regression for M&V: Reference Guide). For “normalized metered energy consumption,” measurement uncertainty is assumed to be zero, since it uses data from revenue-grade meters. Even if measurement uncertainty was included, regression uncertainty would be the dominant term, by far.