San Diego Gas & Electric Multifamily HOPPs Final Report

June 25, 2021

CALMAC Study ID: SDG0344.01

res-INTEL

Residential Energy and Water Intelligence (Res-Intel)© Software

Information	Details
Sector Lead	Hal Nelson
Project Manager	Hunter Johnson
Telephone Number	(909) 542-8401
Mailing Address	4110 SE Hawthorne Blvd. #143, Portland, OR 97214
Email Address	hal.nelson@res-intel.com, hunter@res-intel.com
Report Location	TBD

Table of Contents

Executive Summary

This report summarizes the methodology and results of Res-Intel's normalized metered energy consumption (NMEC) analysis of San Diego Gas & Electric's (SDG&E) Multifamily High Opportunity Project and Programs (MF HOPPs) energy efficiency program. California Investor Owned Utilities (IOUs) are required by Assembly Bill (AB) 802 to report weather-normalized, meter-based program savings. This analysis is part of Res-Intel's work in developing methods and models to determine meter-based savings across SDG&E's portfolio.

The MF HOPPs program targets high energy use intensity (EUI) multifamily buildings built before 1980 across SDG&E territory. The program uses a direct install model to perform normal replacement of common area energy efficiency measures. This NMEC analysis focuses on 41 multifamily properties that began participating in MF HOPPs between August 2017 and December 2018. Each participating site had one or more of the following energy conservation measures (ECM) installed during this time period: light-emitting diode (LED) fixtures, high-efficiency boilers, and variable speed pool pumps.

NMEC evaluation relies heavily on the use of statistical modeling to determine a property's energy-use baseline and to evaluate gross savings resulting from ECM installations. Unlike more engineering-centric approaches, NMEC relies directly on metered data and does not provide *deemed savings* estimates. Res-Intel uses a modeling approach that is consistent with the retrofit isolation approach outlined in the International Performance Measurement and Verification Protocol (IPMVP). Our analysis uses a variant of the time-of-week temperature (TOWT) linear regression model that is commonly used in NMEC evaluation. Res-Intel follows the recommendations of ASHRAE Guideline 14-2014 for evaluating baseline model quality. Additionally, TOWT model performance and results are compared to a more recently developed machine learning model described in Appendix A.

This study estimates the savings attributable to the MF HOPPs program in three steps. First, Res-Intel uses NMEC methods to calculate gross savings realized at each participating site. Gross savings approximate the savings that coincide with the installation of MF HOPPs energy conservation measures. Second, Res-Intel calculates adjusted gross savings by reducing the influence of outliers on baseline models and by adjusting for the effects of concurrent participation in other energy efficiency programs. Adjusted gross savings approximate the savings that are caused by the ECM after removing the effects of concurrent program enrollment and non-routine events (NRE). Third, net savings attributable to the MF HOPPs program are calculated by adjusting for free ridership.

Res-Intel's analysis of savings from the MF HOPPs program yields the following findings:

- NMEC-estimated gross electricity savings of 354 MWh among lighting replacements were, on average, approximately equal to deemed savings. Gross savings represented approximately 20% of baseline electricity use.

- NMEC-estimated gross gas savings of 69,225 therms from high-efficiency boilers which represented approximately 18% of baseline natural gas use.
    - Gross savings underperformed deemed savings projections by approximately 50 percent.
- Aggregate NMEC savings showed considerable variation in savings across sites. Electricity savings for individual customers often deviated significantly from deemed savings.
- Weather normalization had no statistical impact on electricity or gas savings for measures included in the MF HOPPs program.
- Evergreen Economics developed a phone interview guide to inform a net-to-gross ratio for net savings estimates. Only three of forty-one MF HOPPs participants completed the interview which yielded a net-to-gross ratio of 0.61, slightly greater than the ex-ante estimate of 0.55 for high-efficiency boilers and LEDs.
    - Because of the small number of responses, Res-Intel/Evergreen determined that the net-to-gross interview estimates were not statistically valid.
- The ex-ante net-to-gross savings ratio is 0.55 for boilers and LEDs was used instead. Using this ratio resulted in estimated net savings of 197 MWh and 38,073 therms from the MF HOPPs program.

Res-Intel's analysis of the MF HOPPs program contributes to the growing domain of knowledge on NMEC analysis with the following recommendations broken out into two categories for clarity:

Program Design and Data Collection

- Program overlap is difficult to measure empirically and can be better addressed with improved data collection during the retrofit process to isolate savings impacts:
    - Tracking service point numbers, affected meter numbers, and service addresses is essential: tracking account names and/or property names is less important.
- Concurrent participation in multiple energy efficiency programs does not necessarily make NMEC analysis infeasible. However, it does limit the scope of the analysis and could require omitting some customers from the final evaluation.
- Attempt to follow-up on Net to Gross interviews as soon as possible after program participation.
    - Update project tracking data to include detailed project contact information including name, phone number, and email address.
- It is fundamentally important that program implementers focus on accurately reporting (i) affected meter numbers and (ii) deemed savings values. Our evaluation produced some evidence that implementers may have over-reporting affected meters, which violates the retrofit isolation approach and attenuates NMEC savings estimates.

- Utilities can hedge against implementer misreporting by compiling a meter-to-service point mapping that allows evaluators to access all meters associated with a given property.
    - In multifamily and commercial evaluations, the mapping should also separate common-area from tenant meters whenever possible.
    - Res-Intel has performed this meter-to-property mapping for all residential and commercial meters for other projects for SCE, SDG&E, and PG&E and it has proven feasible and cost-effective.

NMEC Methods and Tools

- Evaluators should be careful not to place too much confidence on site-level evaluations due to data quality issues. Program or population-level estimates tend to align more closely with expectations, while site-level savings estimates can vary substantially.
- NMEC savings estimates should be accompanied by confidence intervals whenever possible to ensure that savings uncertainty is properly communicated.
- Researchers and evaluators must focus on developing better statistical methods for detecting whether baseline data is inadequate due to the presence of non-routine events.
    - Current statistical measures do not detect non-routine events and can convey false optimism about the quality of baseline data.
- Improvements in non-routine event detection may require using multi-year baseline periods to distinguish one-time events from seasonal patterns in energy consumption.
- Machine learning methods represent a viable alternative to traditional statistical models of energy consumption used for NMEC.
    - However, using machine learning methods appeared to have little effect on the final savings estimates for this sample.

1. Introduction

This document summarizes the methodology and results of Res-Intel's normalized metered energy consumption (NMEC) analysis of San Diego Gas and Electric's (SDG&E) Multifamily High Opportunity Project and Programs (MF HOPPs) energy efficiency program. Assembly Bill (AB) 802 requires California Investor Owned Utilities (IOUs) to report weather-normalized, meter-based program savings. This analysis is part of Res-Intel's work in developing methods and models to determine meter-based savings across SDG&E's portfolio.

The MF HOPPs program targets high energy use intensity (EUI) multifamily buildings built prior to 1980, regardless of income qualification or location. The program uses a direct install model to perform early replacement of common area energy efficiency measures. Res-Intel's NMEC evaluation of the MF HOPPs focuses on 41 multifamily properties that began participation between August 2017 and December 2018. During this time period each site had one or more of the following energy conservation measures (ECMs) installed: light-emitting diode (LED) fixtures, high-efficiency boilers, and variable speed pool pumps. This report uses NMEC methods to report the gross savings attributed to these measures.

NMEC fundamentally refers to the practice of relying heavily on statistical models to establish a property's energy-use baseline and to evaluate savings that result from energy conservation measure (ECM) installations. NMEC methods stand in contrast with more engineering-centric approaches that do not rely directly on metered data and provide *deemed savings* estimates. Guidelines for the proper implementation of NMEC methods are still evolving and uncertainty remains about reliability and best practices when using statistical models to calculate program savings. Res-Intel's analysis of the MF HOPPs program contributes to the growing domain of knowledge on this topic with the following findings:

- NMEC-estimated electricity savings among lighting replacements were, on average, approximately equal to deemed savings.
- Aggregate NMEC savings belied considerable variation in savings across sites. Electricity savings for individual customers often deviated significantly from deemed savings.
- NMEC-estimated gas savings from high-efficiency boilers underperformed deemed savings projections by approximately 50 percent.
- Weather normalization had no statistical impact on electricity or gas savings for measures included in the MF HOPPs program.
- The existence of concurrent program participation can severely bias NMEC-estimated savings and may justify exclusion of some customers from the evaluation.
- The detrimental effects that concurrent participation has on the quality of the evaluation can be mitigated by more detailed and accurate reporting from project implementers.

- Although energy-use recorded by common-area meters is often highly predictable, non-routine events (NREs) can severely hamper site-level evaluations.
- The impacts of NREs are often not revealed by traditional metrics of model quality, such as those recommended in ASHRAE Guideline 14.
- Evaluators may consider multi-year baseline periods to improve detection of NREs and shift the focus in model evaluation to year-over-year out-of-sample prediction accuracy.
- The impacts of NREs can be mitigated by down-weighting outlier meter readings.
- Evaluation of common-area measures is more difficult when the measure represents a small share of the common-area meter's total load.
- Using a retrofit isolation approach is often required when evaluating multifamily retrofits, and extra care must be taken to ensure implementors report *only* the affected meters.
- NMEC analysis can benefit from access to a meter or service point inventory for multifamily properties undergoing retrofit. Such an inventory could be used to supplement and cross-reference implementor-supplied data that may be incomplete, particularly among customers participating in multiple programs.
- Using machine learning methods for NMEC improves model accuracy, though the savings estimates they produce are roughly equal to those estimated by more traditional regression models.

The remainder of this report explains the methods Res-Intel employed in its evaluation and summarizes the results in greater detail. The following section provides a high-level outline of our approach to energy savings evaluation, Section 3 explains mathematical and statistical methods used to calculate gross savings, Section 4 summarizes the program and energy data made available by SDG&E, Section 5 summarizes the savings evaluation results, Section 6 presents findings on key measurement uncertainties discovered in the modeling process, and Section 7 finishes by offering recommendations for future NMEC projects.

2. Gross and Net Savings Evaluation

This study sets out to estimate the savings attributable to the MF HOPPs program. It begins using Normalized Metered Energy Consumption (NMEC) methods to evaluate gross savings realized at each participating site. We then calculate adjusted gross savings by (i) reducing the influence of outliers in hourly metered consumption on baseline models and (ii) adjusting for the effects of concurrent participation in other energy efficiency programs. A separate memo will report on net savings, which are derived by altering the adjusted gross savings estimates to account for survey-based measures of free ridership.

Table 2.1 enumerates each of these steps in the order of execution. The first measurement, gross savings, approximates savings that coincide with the installation of the MF HOPPs energy conservation measure (ECM). The adjustments approximate the savings that are directly caused by the ECM by removing the effects of concurrent program enrollment and NREs. Finally, net savings adjusts for free ridership to determine the savings attributable to the MF HOPPs program. The table also lists the methods employed in these measurements, each of which is detailed in Section 3.

Table 2.1: NMEC Savings Estimates

Measurements	Methods Employed	Primary Data Requirements	Interpretation
Gross Savings	Statistical evaluation.	Hourly metered consumption data.	Savings coinciding with ECMs.
Adjusted Gross Savings	Model adjustments; customer exclusion.	EE program data, participant survey.	Savings caused by ECMs.
Net Savings	Survey analysis.	Participant survey.	Savings attributable to MF HOPPs.

In addition to reporting observed savings attributable to the program, this report also extrapolates these findings to calculate savings expected during a typical weather year. We refer to this as the *normal savings*. The *normal savings* are defined as savings that one would expect to observe during a year in which average temperatures prevail.

3. Methods

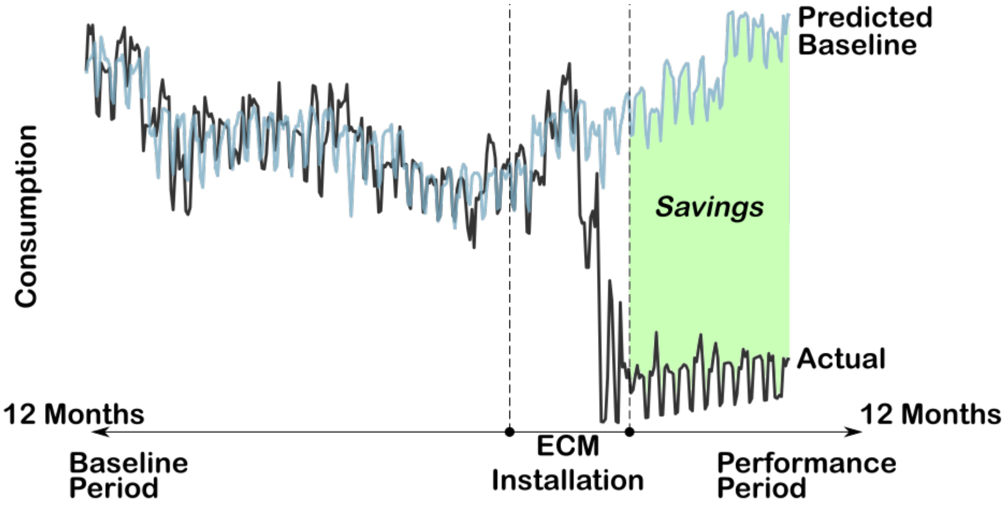
Our analysis uses statistical methods to assess the overall impact of the MF HOPPs program and the savings realized for individual sites and ECMs. It begins by estimating baseline consumption models for gas and electricity meters that were affected by ECMs. These models are used to predict consumption for each time interval after the ECM installation date. Model predictions after the ECM installation date represent the baseline consumption estimate ($\widehat{Baseline}$), which can be compared to actual consumption to determine changes caused by ECMs. The changes in consumption that coincide with the ECM installation dates are aggregated across gas and electricity meters to determine gross savings attributable to the program:

$$Gross\ Savings = \widehat{Baseline} - Actual.$$

Predicting baseline energy consumption requires first estimating a statistical model using 12 months of metered consumption data. The savings estimates that appear in this report are estimated using a commonly employed linear regression model. This linear regression model is a variant of the commonly used time-of-week temperature (TOWT) model, summarized in Mathieu et al. (2011), and is commonly used in NMEC analysis. Additionally, the Appendix compares the TOWT model performance and results with a more recently developed machine learning approach. For the sake of clarity and because the savings differ very little across modeling approaches, we present only TOWT model findings in the main portion of the report.

Gross savings are first calculated at the meter level by subtracting actual consumption from the predicted baseline consumption during the performance period (see Figure 3.1). Savings at the site-level are then determined by aggregating gross savings across affected meters at each site. Calculating gross savings for each ECM requires a more sophisticated approach. Regressing site-level savings on the quantity of ECMs installed at each site yields a set of coefficients that indicate the savings attributable to each type of ECM. Finally, program-level savings are calculated by simply aggregating site-level savings.

Figure 3.1: NMEC Gross Savings Calculation



In addition to calculating gross savings, this report also reports the *realization rate* for each site and ECM. The realization rate is defined as the ratio of meter-based gross savings to deemed savings. It can be interpreted as the percentage of the deemed savings that was realized by the NMEC evaluation:

$$Realization\ Rate = \frac{Gross\ Savings}{Deemed\ Savings}.$$

Statistical Modeling of Energy Use

Our proposed modeling approach is consistent with the retrofit isolation approach outlined in the International Performance Measurement and Verification Protocol (IPMVP) and satisfies many of the data requirements for whole-building analysis, often referred to as “Option C” (IPMVP 2002). This section summarizes the properties of the TOWT regression and the criteria for evaluating baseline model quality, following the recommendations of ASHRAE Guideline 14-2014.

Time-of-the-Week (TOWT) Model

The TOWT model takes the form of a simple linear regression with a flexible intercept to accommodate changes in consumption observed at different (i) times of week, (ii) months of the year and (iii) outside temperature ranges. Expressed mathematically, the TOWT regression predicts consumption c at time t with the following equation:

$$c_t = \sum_{\tau} \beta_\tau I(\tau(t) = \tau) + \sum_{m} \beta_m I(m(t) = m) + \sum_{k} \beta_k I(temp_t \in \mathrm{k}) + \varepsilon_t.$$

The terms τ, m and k index the time-of-week, month-of-year and temperature range. The temperature range takes on 10 different value ranges, reflecting the deciles of the outside temperature distribution associated with each meter. The m variable takes on 12 possible values, while the τ variable takes on 168 possible values in hourly data and 7 values in daily data. The indicator function $I(\cdot)$ equals one when its argument is true and zero otherwise. The β coefficients therefore represent the adjustments to the consumption prediction intercept when an hour falls in time-of-week τ, month m and temperature range k. Finally, the term ε represents random modeling error.

This model should be familiar to most NMEC practitioners. For instance, a version of TOWT model is described in the popular CalTrack NMEC methodology guidelines and TOWT models are used in numerous publications authored by researchers at the Lawrence Berkeley National Laboratories (Granderson et al. 2017).

Model Evaluation

To ensure the baseline models provide accurate predictions of consumption, we evaluate model performance using three different metrics: coefficient of determination (R^2), coefficient of variation of the root-mean-squared error $(CVRMSE)$ and normalized mean bias error $(NMBE)$. These metrics are calculated as follows:

$$R^2 = 1 - \sum_{T} \frac{(c_t - \hat{c}_t)^2}{(c_t - \bar{c})^2}$$

$$CVRMSE = 100 \times \frac{\left(\sum(c_i - \widehat{c_i})^2 / (n-p)\right)^{0.5}}{\bar{c}}$$

$$NMBE = 100 \times \frac{\sum_T (\hat{c}_t - c_t)}{(n-p) \times \bar{c}}$$

The terms $\hat{c}$, $\bar{c}$, n and p denote the predicted consumption values, average consumption, the number of observations in the baseline sample and the number of free parameters in the model, respectively. The R^2 measures the amount of variance in consumption explained by the model, ranging from 0 to 1. The ASHRAE Guideline 14-2014 recommends a minimum 0.7 for any model used in NMEC. The CVRMSE measures the model error rate, standardized by its mean consumption value, ranging from 0 to infinity. ASHRAE recommends a maximum CVRMSE of 35. Lastly, NMBE is the total bias of the model normalized by total consumption, where a value of zero indicates that no bias exists. ASHRAE recommends using only models whose NMBE does not exceed 0.005.

Gross Savings

Savings are expressed in either kWh or therms depending on whether the ECM affects electricity or gas consumption. Savings estimates do not account for any interaction effects between gas and electric ECMs. Given that all MF HOPPs ECMs included only common-area boilers and lighting retrofits, Res-Intel and SDG&E evaluation staff concluded that interaction effects would be either non-existent or too small to detect statistically.

Site

Site-level savings are simply the sum of savings calculated for each affected meter at the site. Let s be an index of sites and i be an index of meters. Total savings for site s is calculated by summing the model residuals over all time intervals in the 12-month performance period:

$$Savings(site) = \sum_{i \in s} \sum_{t \in i} (\hat{c}_{i,t} - c_{i,t}).$$

Program

Total program savings are the sum of savings over all sites:

$$Savings(program) = \sum_{s} \sum_{i \in s} \sum_{t \in i} (\hat{c}_{i,t} - c_{i,t}).$$

ECM

Calculating ECM-level savings is a two-step process. First, site-level savings must be calculated. Then site-level savings can be regressed by the quantity of each ECM installed at each site, $Q_{s,e}$, where e indexes the different ECMs installed. The regression equation is

$$Savings_s = \sum \beta_e Q_{s,e} + \varepsilon_s.$$

The coefficient β_e represents the marginal impact of one unit of ECM e on site-level savings:

$$Savings(ECM_e) = \beta_e.$$

Confidence Intervals

Confidence intervals communicate uncertainty over savings estimates and require estimating the variance in gross savings. At the meter level, savings variance is simply the variance of the sum of residuals:

$$\begin{aligned} Var(Savings) = \sigma^2 &= Var\left(\sum(\widehat{c_t} - c_t)\right) \\ &= n^2 \times Var\left(\bar{\hat{c}} - \bar{c}\right) \\ &= n \times \left(Var(\hat{c}) + Var(c) - 2 \times Cov(\hat{c}, c)\right) \\ &= \sum\left(\widehat{c_t} - \bar{\hat{c}}\right)^2 \times \alpha_{\hat{c}} + \sum(c_t - \bar{c})^2 \times \alpha_c - 2 \times \sum\left(\widehat{c_t} - \bar{\hat{c}}\right) \times (c_t - \bar{c}). \end{aligned}$$

The second line separates the variance of predicted and actual consumption while correcting for the covariance between the two terms. The last line introduces the α term which corrects for autocorrelation in each variable: $\alpha_x = (1 + \rho_x)/(1 - \rho_x)$. Site-level variance is the sum of meter-level variances at the site and program-level variance is the sum of meter-level variance across all affected meters.

The confidence intervals that appear in this report are calculated at the 95-percent confidence level. Each site's savings confidence interval is therefore

$$CI_s = (Savings_s - 1.96 \times \sigma_s, Savings_s + 1.96 \times \sigma_s).$$

Adjusted Gross Savings

Adjusted gross savings are derived from gross savings by applying adjustments that correct for the existence of non-routine events (NREs) and concurrent participation in other energy efficiency programs. We reduce the effect of non-routine events by down-weighting the effects of outlier meter readings, and we adjust for concurrent participation by deducting its average impact among all affected customers:

$$Adjusted\ Gross\ Savings\ =\ Gross\ Savings + (Outlier, Concurrent\ \ Adjustments)$$

Outlier Adjustment

Our method of outlier classification identifies outliers at the meter level, focusing on hourly data points. An outlier is defined as an hourly observation that deviates considerably from a meter's average consumption at a given hour-of-week. Let c_t represent a meter's consumption at hour-of-year $t\colon t \in [1, 8760]$. There are 168 hours in every week and 52 weeks in a year, so classification begins by calculating hourly average consumption for 168 groups of hours, labeled $\bar{c}_k$ for $k = 1 \dots 168$. An outlier score is then assigned to each hour using the following equation:

$$Outlier\ Score_t = \frac{1}{\sigma_k} \times \sqrt{\left(c_t - \bar{c}_{k(t)}\right)^2}$$

where σ_k is the standard deviation of consumption at hour-of-week k. The outlier score, therefore, represents the absolute value of the distance between an hour's consumption and its hour-of-week average, measured in standard deviations.

The characterization of outliers can be used to adjust and improve the data input into the baseline NMEC models. One approach to adjusting baseline data is to drop observations that exceed a certain outlier score threshold prior to estimating the model. Removing extreme outliers can reduce the confounding effects that erroneous data have on the resulting baseline model. Indeed, this approach is advocated by the SEM guidelines (SEM 2017), which recommend dropping all hourly observations that deviate more than three standard deviations from the overall hourly average.

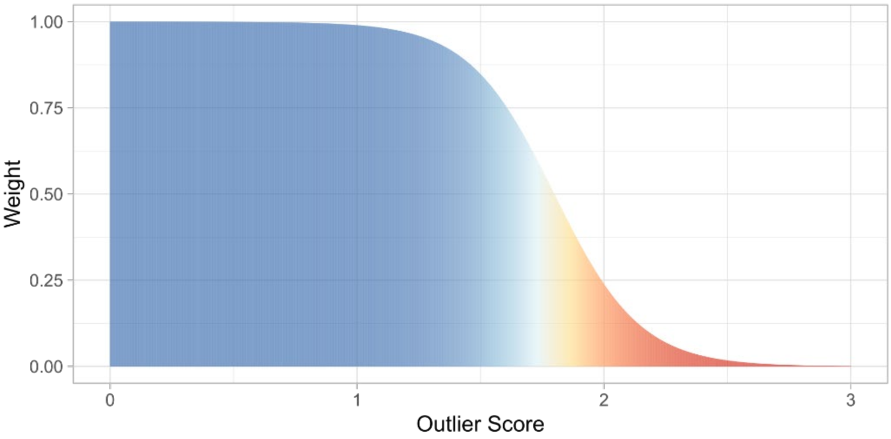
An alternative approach, which we use here, is to apply weights to baseline observations that reflect their outlier scores. These weights dictate how much confidence the baseline regression model places on each observation. Observations with high outlier scores are assigned small weights, while those with low outlier scores are assigned large weights.

Figure 3.2 illustrates how observation weights vary as a function of the outlier score. The sigmoid function used to translate outlier scores into observational weights is

$$Weight_t = \left(1 + e^{5.8 \times Outlier\ Score_t\ -\ 10.3}\right)^{-1}.$$

Applying this weighting function to the modeled data has the effect of placing more emphasis on the central tendencies of the consumption data. Observations that are close to the hour-of-week average affect the resulting model more than observations that deviate significantly from typical usage. It also replicates the SEM rule of dropping observations more than three standard deviations from the mean by applying near-zero weights to these observations.

Figure 3.2: Weighting Observations on Outlier Score



Concurrent Program Participation

Estimation of MF HOPPs program savings is complicated by the fact that 16 of the 41 participating customers were simultaneously receiving retrofits in connection SDG&E's Multifamily Energy Efficiency Retrofit (MFEER) program. Many of the MFEER ECM installation dates coincided on the same day as MF HOPPs installation or were installed only a few days apart. Additionally, many customers had MFEER installations taking place both before and after the MF HOPPs installation. The existence of simultaneous ECM installations presents serious difficulties to statistical modeling of energy use and savings.

Figure 3.3: Effect of MFEER Installations

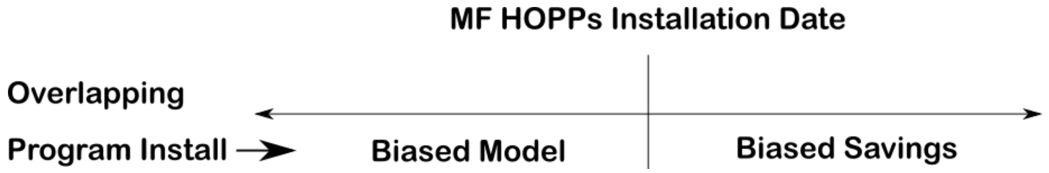


Figure 3.3 illustrates the consequences of ECM installations from other programs that occur during the MF HOPPs baseline or performance period. Installations that occur during the baseline period cause bias in the baseline model, while installations that occur in the performance period bias savings measurements.

The MFEER program included installations of lighting fixtures, water fixtures, and HVAC equipment. In some circumstances, it might be possible to statistically adjust for the impact of these measures using data that describes specific MFEER installations to modify the site- and meter-level models. However, the following deficiencies in data availability prevent site-level adjustments:

1. The SDG&E program database reports MFEER installations and deemed savings at the utility *account* level. Multi-family sites can have many utility accounts and each account is often linked to several different meters. The retrofit isolation approach used in our analysis, however, requires program information to be available at the meterlevel.
2. When reporting on installations at multifamily sites, implementers often report the site's primary account number, even though the affected meters are connected to different accounts at the same site.
3. SDG&E does not keep an inventory of accounts and meters associated with each site beyond what is reported by individual implementers. Access to such an inventory would create several possibilities for overcoming (1) and (2), such as conducting a whole-building evaluation.

Due to the above deficiencies in MFEER program data, we decided to instead *statistically* estimate the *average* impact of MFEER among the sites that participated in MF HOPPs. This statistical average can then be deducted from the total MF HOPPs savings among the 16 sites that participated in both programs. This approach is likely to cause some glaring inaccuracies among the site-level evaluations—because site-level MFEER savings are likely to deviate significantly from the average—but it nevertheless provides a consistent correction for the aggregate program-level evaluation.

Determining the average impact of MFEER participation on MF HOPPs savings requires estimating the following regression:

$$Savings_s = \beta_D Deemed_{s,e} + \beta_{mfeer} I_{mfeer} + \varepsilon_s.$$

where I_{mfeer} equals one if a site concurrently participated in MFEER and zero otherwise, and $Deemed$ is the reported MF HOPPs deemed savings. The coefficient β_D therefore represents the deemed savings realization rate and the coefficient β_{mfeer} represents the impact of MFEER participation on savings, conditional on the MF HOPPs deemed savings. The adjusted savings for each of the 16 sites concurrently participating in MFEER is equal to $Savings_s - \beta_{mfeer}$.

Weather Normal Savings

Our primary estimate of measured savings compares predicted energy use to actual energy use, where predictions adjust to the actual temperatures observed during the performance period. The primary estimate therefore reveals the *observed* level of avoided energy use. Using climate normal data, we can also estimate the savings that would occur during a typical weather year. This section describes the methods we use to calculate *normal* savings, defined as savings that one would expect to observe during a year in which average temperatures prevail.

Modeling Approach

Our modeling approach follows the ASHRAE 14-2014 Guideline's recommendation to begin by estimating two separate models for baseline and performance period consumption, and then to compare the predictions of those two models when applied to temperature averages. We calculate normal savings using the following equation:

$$Normal\ Savings\ =\ Baseline\ |\ AvgTemps - \widehat{Actual}\ |\ AvgTemps,$$

where the term $\widehat{Actual}$ denotes the predicted consumption using the performance period model.

The baseline period model is identical to the one that is used to calculate observed savings. The performance period model, however, is a new component to the savings calculation—it is needed to predict how much energy would be consumed if *typical* temperatures prevailed in the performance period, rather than the observed temperatures.

Data

The hourly weather normals are supplied by the CZ 2010 and CZ 2018 weather files, compiled by White Box Technologies at the request of the California Energy Commission. CZ 2010 calculates temperature averages for three different ISD weather stations in San Diego based on 2010 data, while CZ 2018 calculates averages for all weather stations based on 2018 data. We match each MF HOPPs site to the closest ISD weather station that is included in CZ 2010. Given that the set of ISD stations available for current-year observed temperatures is greater than the set of stations in the CZ 2010 dataset, many sites have CZ 2010 stations that are different from their observed and CZ 2018 temperature stations. Columns 1 and 2 of Table 3.2 list the pairings of ISD stations used to reference observed and normal temperatures.

Table 3.2: Stations and Average Temperatures (CZ10 = 2010, CZ18 =2018)

Station		Average Temperature (F)		
Observed & CZ18	CZ10	Observed	CZ10	CZ18
722904	722900	63.5	62	62.6
722906	722900	64.5	62	65.2
722903	722903	64.8	61.2	63.7
722907	722903	64.4	61.2	63.6

Columns 4 to 6 of the table list the average observed temperature during the MF HOPPs evaluation window, which spanned from August 2016 to May 2019. Average hourly temperatures ranged from 63.5 to 64.8 F during this time period. CZ 2010 average temperatures are significantly lower, ranging from 61.2 to 62 F; CZ 2018 averages are closer to the observed averages, ranging from 62.6 to 65.2 F.

Figure 3.4 illustrates the hourly difference by hour-of-day and shows that much of the CZ 2010 gap owes to higher observed temperatures during the mid-day and afternoon hours. Figure 3.5, which shows average hourly temperature by month, reveals that gaps between temperatures were most pronounced during the summer.

Figure 3.4: Hourly Temperatures- Observed vs. Normals

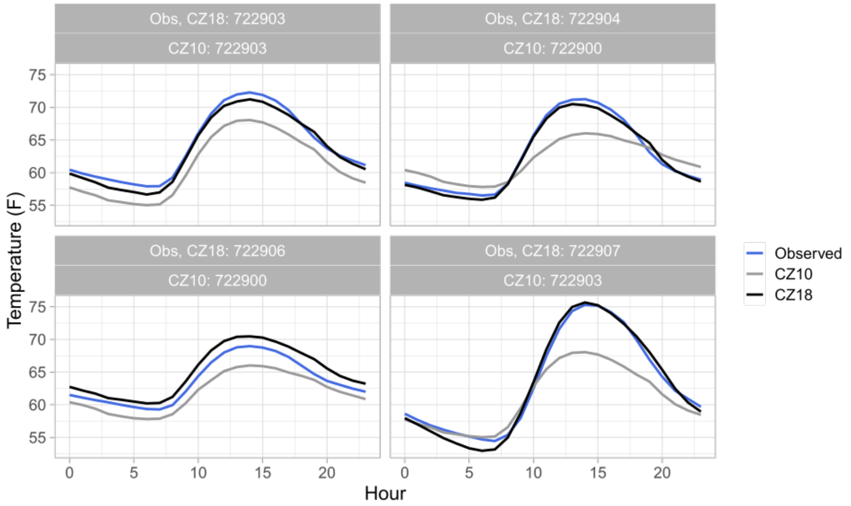
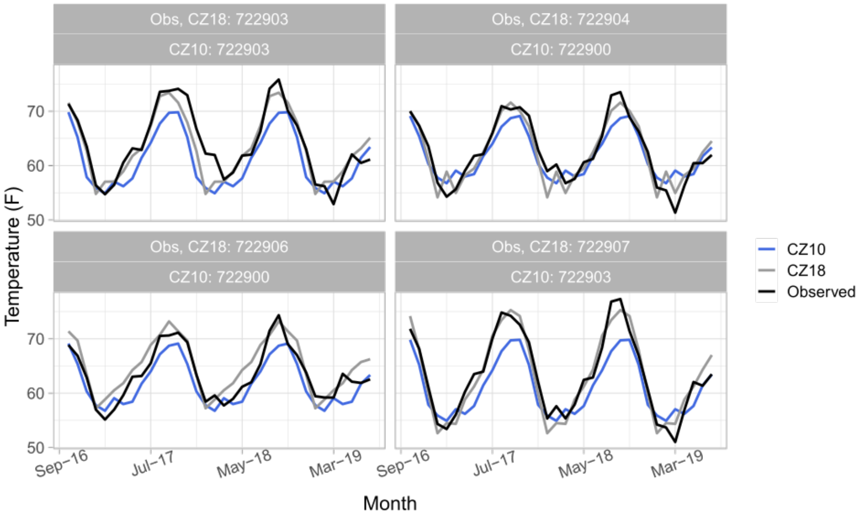


Figure 3.5: Monthly Temperatures- Observed vs. Normals



4. Data Summary

Res-Intel collaborated with SDG&E staff to obtain the data required to perform the analysis, which included:

- Implementor-supplied data:
  - ID numbers for each meter affected by an ECM.
  - ECM quantities and installation dates.
  - Maintenance/repair history for the affected equipment.
  - Operating hours for affected equipment.
  - Ex-Ante deemed savings for each ECM.
- Utility-supplied data:
  - Hourly (electricity) and daily (gas) consumption data for a minimum of one year prior to the first ECM installation date and one year after the installation.
  - Metadata associated with each affected meter, including latitude/longitude, service address, customer name and account number.
  - Program metadata used to identify MF-HOPPs sites participating in multiple programs.

The implementor-supplied dataset listed a total of 41 sites with 332 combined common area meters (gas and electric) impacted by the MF HOPPs program. The dataset tabulates ECMs, deemed savings and installation dates at the meter level. There are 1,760 total ECMs reported in the dataset. The installation dates ranged from 8/22/2017 to 12/4/2018. Table 4.1 reports total ECMs and deemed savings for each site. In total, the deemed estimates projected 257,924 kWh and 114,863 therms in program-related savings. Each site was linked to the nearest weather station with data available between 2016 and 2019. The last column of Table 4.1 shows the weather station assigned to each site.

Table 4.1: Site Summary

		Deemed Savings		Meters		
Site	ECMs	kWh	Therms	Electric	Gas	Weather Station
Site 1	82	15,948	3,104	4	3	722907-53143
Site 2	2	0	1,242	0	1	722900-23188
Site 3	19	2,964	0	3	0	722904-03178
Site 4	2	0	1,242	0	1	722900-23188
Site 5	2	0	1,242	0	1	722904-03178
Site 6	1	0	621	0	1	722900-23188
Site 7	2	0	1,242	0	1	722900-23188

Site 8	35	5,460	0	10	0	722903-03131
Site 9	26	3,276	3,104	14	2	722907-53143
Site 10	187	29,172	0	7	0	722903-03131
Site 11	10	0	6,209	0	5	722903-03131
Site 12	32	4,992	0	15	0	722907-53143
Site 13	74	11,284	6,830	9	10	722903-03131
Site 14	26	4,056	0	2	0	722907-53143
Site 15	5	0	3,104	0	1	722903-03131
Site 16	100	15,600	0	4	0	722904-03178
Site 17	129	18,856	4,346	9	3	722904-03178
Site 18	84	13,104	0	12	0	722907-53143
Site 19	2	0	1,242	0	1	722904-03178
Site 20	105	13,728	10,555	3	1	722904-03178
Site 21	21	0	13,038	0	21	722907-53143
Site 22	278	38,328	21,110	34	34	722904-03178
Site 23	5	0	3,104	0	1	722900-23188
Site 24	50	7,800	0	15	0	722904-03178
Site 25	97	13,104	8,071	10	13	722907-53143
Site 26	70	11,136	0	15	0	722907-53143
Site 27	33	5,858	0	6	0	722907-53143
Site 28	70	10,920	0	8	0	722903-03131
Site 29	13	1,716	1,242	3	1	722903-03131
Site 30	15	1,716	2,484	4	3	722904-03178
Site 31	8	0	4,967	0	1	722904-03178
Site 32	20	2,184	3,725	2	6	722903-03131
Site 33	59	9,076	621	5	1	722903-03131
Site 34	3	0	1,863	0	1	722903-03131
Site 35	16	2,028	1,863	4	1	722903-03131
Site 36	40	9,930	0	14	0	722906-93112
Site 37	0	2,100	0	1	0	722900-23188
Site 38	5	0	3,104	0	1	722903-03131
Site 39	4	0	2,484	0	1	722903-03131
Site 40	5	0	3,104	0	2	722903-03131
Site 41	23	3,588	0	1	0	722906-93112
	1,760	**257,924**	**114,863**	**214**	**118**	

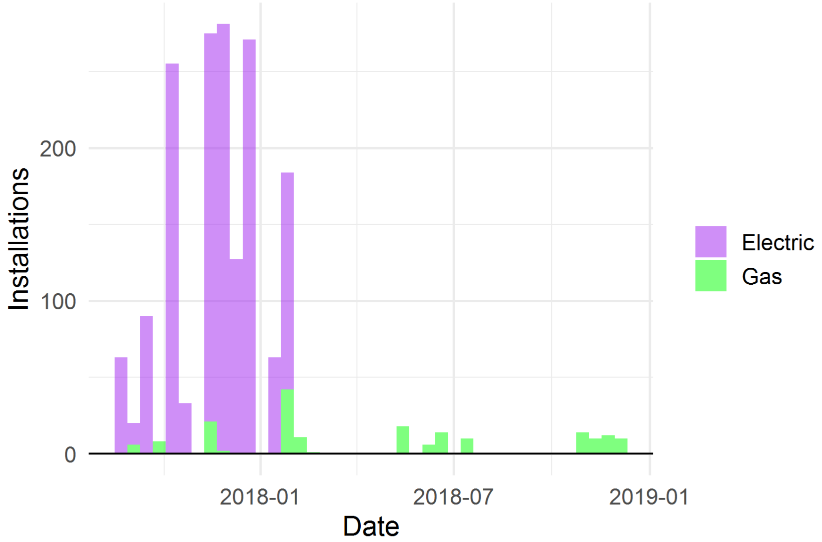
Improved lighting fixtures made up the majority of ECMs. Table 4.2 lists the quantity of each type of ECM installed and the number of meters and sites that were affected by these installations. By far the most frequent installations were 28-watt light-emitting diodes (LED28W). The 185 high-efficiency boilers make up the only ECM category affecting gas consumption.

Table 4.2: Energy Conservation Measures (ECMs)

Measure	Quantity	Affected Meters	Affected Sites
LED28W	1429	200	25
LED40W	2	2	1
LED43W	29	6	5
LED50W	64	21	3
LED70W	66	20	3
Boiler	185	118	27
Pool Pump	1	1	1

All the LED installations took place between 2017 and early 2018. Figure 4.1 plots the frequency of ECM installations throughout the duration of the program. Boiler installations are distributed more uniformly across the program period: the first were installed in early 2017 and the final group of installations occurred in late 2018.

Figure 4.1: ECM Installation Dates



Concurrent Program Participation

It is not uncommon for program implementers to cross-market offerings from other retrofit programs while conducting site visits. Indeed, we discovered that 16 of the 41 MF HOPPs sites participated in SDGE's Multifamily Energy Efficiency Retrofit (MFEER) program concurrently with their participation in MF HOPPs. Although concurrent participation in different programs can be beneficial to implementers and property managers, it adds considerable difficulty to program evaluation.

For reasons explained in the previous section, we cannot precisely estimate the impact that MFEER participation has on individual sites. However, MFEER program documentation does reveal that sites included in our sample received installations that were expected to greatly reduce electricity consumption and slightly increase gas consumption. Therefore, we should expect a positive bias on kWh savings among MFEER participants and a negative bias in therms savings.

Missing Data

While reviewing the implementer data, Res-Intel worked with SDG&E staff to correct inconsistent or missing entries. Several meters listed in the dataset were not assigned deemed savings even though they were assigned a non-zero number of ECMs. Additionally, several affected meters did not report ECM quantities. Given uncertainty about the accuracy of ECM quantity and deemed savings reporting at the meter-level, Res-Intel decided to only use these fields when aggregated to the site level. The possibility of misreporting in the data should be viewed as a source of uncertainty in our analysis.

Among the 214 electricity meters and 118 gas meters included in the NMEC analysis, there was a very small incidence of missing data. For each meter, we calculated the (%) rate of missing data across all time intervals. Table 4.3 reports the distribution of missing rates for electricity and gas meters. Most electricity meters had less than 0.01% missing values and the maximum missing rate was 0.34%. Gas meters did not report any missing data.

Table 4.3: Missing and Duplication Rates per Meter

	Missing %	
Percentile	Electric	Gas
0	0.01	0
25	0.01	0
50	0.01	0
75	0.07	0
100	0.34	0

All missing observations were replaced with hour-of-day average, where the average is evaluated by meter-month pairings.

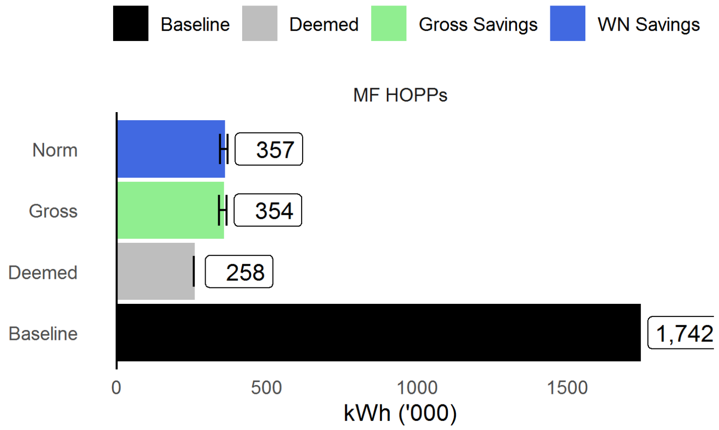
5. Results

In this section, we report several different calculations of energy savings. The first is the total gross savings estimated using the TOWT model, followed by weather normal savings and adjusted gross savings that correct for (1) outlier meter readings and (2) concurrent program participation. Since this report is exploratory in nature, no single savings estimate should be interpreted as more authoritative than any other—instead, they simply offer the opportunity to gauge the impact of the various NMEC savings adjustments.

Electricity Savings

The TOWT models produce an initial estimate of 254 MWh in gross savings, which implies a 1.37 realization rate when compared to deemed savings. In other words, unadjusted gross electricity savings exceeded deemed savings by 37 percent, as illustrated in Figure 5.1. Weather normal estimates show a nearly identical estimate of 257 MWh in gross savings. This could be explained by the fact that lighting, the predominant electrical measure, is not affected very much by changes in outside temperatures. Baseline consumption among all measured meters totaled 1.7 GWh, so these savings signify a roughly 20 percent reduction in baseline consumption. The deemed savings estimates in Figure 5.1 were provided by the MF HOPPs implementer at the property level based on their energy savings calculations.

Figure 5.1: Unadjusted Electricity Savings



Note: The whiskers indicate 95 percent confidence intervals for each estimate.

Columns 3 and 4 of Table 5.1 break down electricity savings based on whether sites were concurrently enrolled in MFEER (Concur/Non-Concur). The breakout in savings presents a clear indication that MFEER participation introduced a positive bias in estimated savings: concurrently enrolled participants achieved a realization rate of 1.66 while those not enrolled in MFEER achieved a rate of 1.05. Adjusting for concurrent enrollment reduces estimated gross savings to 272 MWh, bringing the aggregate realization rate down to 1.06.

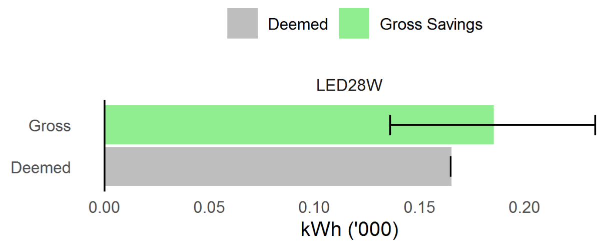
Table 5.1: Electricity Savings (kWh)

	Unadjusted				*Adjusted*		
	All	Norm	Concur.	Non-Concur.	Concur.	Outlier	Concur. & Outlier
Gross	354,079	357,133	227,497	126,583	272,211	306,480	222,315
Deemed	257,924	257,924	137,124	120,800	257,924	257,924	257,924
R-Rate	1.37	1.38	1.66	1.05	1.06	1.19	0.86
Baseline	1,741,905	1,741,905	1,063,852	678,052	1,741,905	1,741,905	1,741,905
Gross %	20	21	21	19	16	18	13
Deemed %	15	15	13	18	15	15	15

Our second adjustment involves reducing the impact of outlier meter readings using the methodology outlined in Section 3. Applying this adjustment has the effect of reducing gross savings estimates to 306 MWh with a realization rate of 1.19. Combining the outlier adjustment with the concurrent enrollment adjustment further reduces gross savings to 222 MWh. The combined adjustments bring the realization rate down to 0.86, signifying gross savings that are 14 percent *below* deemed estimates.

Overall, the range of estimates suggests that lighting measures performed roughly on-par with deemed savings projections. Though combined outlier and concurrent enrollment adjustments bring gross savings below deemed, it bears emphasizing that this estimate should not carry more authority than any other that this stage of our analysis.

Figure 5.2: Electricity ECM Savings



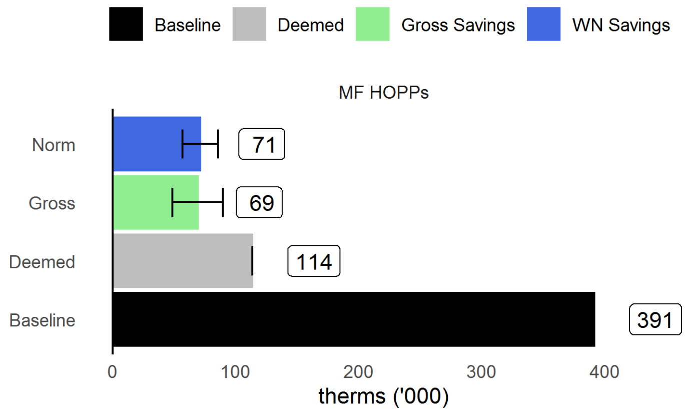
Note: The whiskers indicate 95 percent confidence intervals for each estimate.

Figure 5.2 plots our ECM-level estimate for the unadjusted gross savings attributed to 28-watt LEDs, the most common electricity measure. Consistent with our program-level findings, the LEDs produced savings well within the range of the roughly 160 kWh deemed savings projection. Indeed, the 95-percent confidence interval for gross savings, illustrated by the whisker in Figure 5.2, encompasses the deemed savings estimate.

Gas Savings

Gas savings, deriving solely from high-efficiency boiler installations, totaled to 69 thousand therms prior to any adjustments. Weather normal savings were slightly higher, totaling 71 thousand therms, though the confidence intervals illustrated in Figure 5.3 show that this difference is not statistically significant. Unadjusted and weather normalized gas savings underperformed the deemed savings projection provided by the implementer of 114 thousand therms by nearly 50 percent, achieving realization rates of 0.53 and 0.55. This underperformance is statistically significant.

Figure 5.3: Unadjusted Gas Savings



Note: The whiskers indicate 95 percent confidence intervals for each estimate.

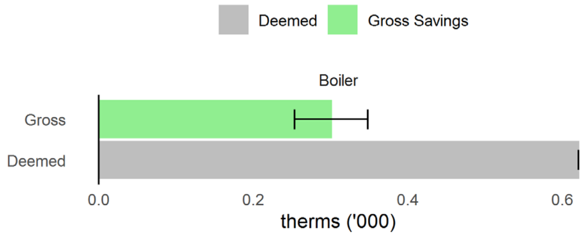
Table 5.2 shows that the changes in gas savings caused by outlier and concurrent participation adjustments are relatively small. The concurrent enrollment adjustment reduces savings to 66 thousand, the outlier adjustment reduces it to 63 thousand, and combined they reduce savings to 61 thousand. Although these adjustments imply lower realization rates (0.51, 0.49 and 0.47), their differences from the unadjusted estimate are not statistically significant, and therefore do not undermine the unadjusted realization rate of 0.53.

The savings estimates for individual boilers, illustrated by Figure 5.2, is 300 therms, equal to roughly 50 percent of their 620 therm deemed savings. The estimated confidence interval for boiler savings does not include 620 therms, proving that this underperformance is statistically significant.

Table 5.2: Gas Savings (therms)

	Unadjusted				*Adjusted*		
	All	Norm	Concur.	Non-Concur.	Concur.	Outlier	Concur. & Outlier
Gross	69,225	71,404	18,363	50,861	66,127	63,510	61,238
Deemed	129,528	129,528	47,956	81,572	129,528	129,528	129,528
R-Rate	0.53	0.55	0.38	0.62	0.51	0.49	0.47
Baseline	391,302	391,302	123,645	267,657	391,302	391,302	391,302
Gross %	18	18	15	19	17	16	16
Deemed %	33	33	39	30	33	33	33

Figure 5.4: Gas ECM Savings



Note: The whiskers indicate 95 percent confidence intervals for each estimate.

Site-Level Savings

An examination of site-level gross savings reveals considerable variation in outcomes across participating customers. Figures 5.3 and 5.4 illustrate unadjusted electricity and gas savings for each site. The Appendix reports these results in table format. Among sites receiving electrical ECMs, several achieved savings far above deemed savings, including Site 13, Site 36 and Site 10. However, there were also several sites where electricity savings did not reach levels statistically above zero, including Site 3, Site 26 and Site 29. Although all sites aggregated to a realization rate of approximately 1, individual site-level savings more commonly deviated significantly from deemed savings.

Gas savings, in contrast, consistently underperformed deemed values, with all sites achieving realization rates below 1. Gas savings estimates were also considerably more uncertain: confidence intervals often included zero and sometimes included deemed savings. The level of uncertainty implies that an accurate measure of gas savings cannot be determined at the site level for many individual customers.

The site-level figures also reveal a potential data quality concern, showing that deemed savings exceeded baseline consumption for several sites receiving gas and electricity ECMs. In particular, deemed kWh exceeded baseline consumption for Site 16 and deemed therms exceeded baseline consumption for Site 25 and Site 33. These data inconsistencies may signal deficiencies in the accuracy of implementer reports or miscalibrations in the application of deemed savings estimates.

Figure 5.5: Site-Level Electricity Savings

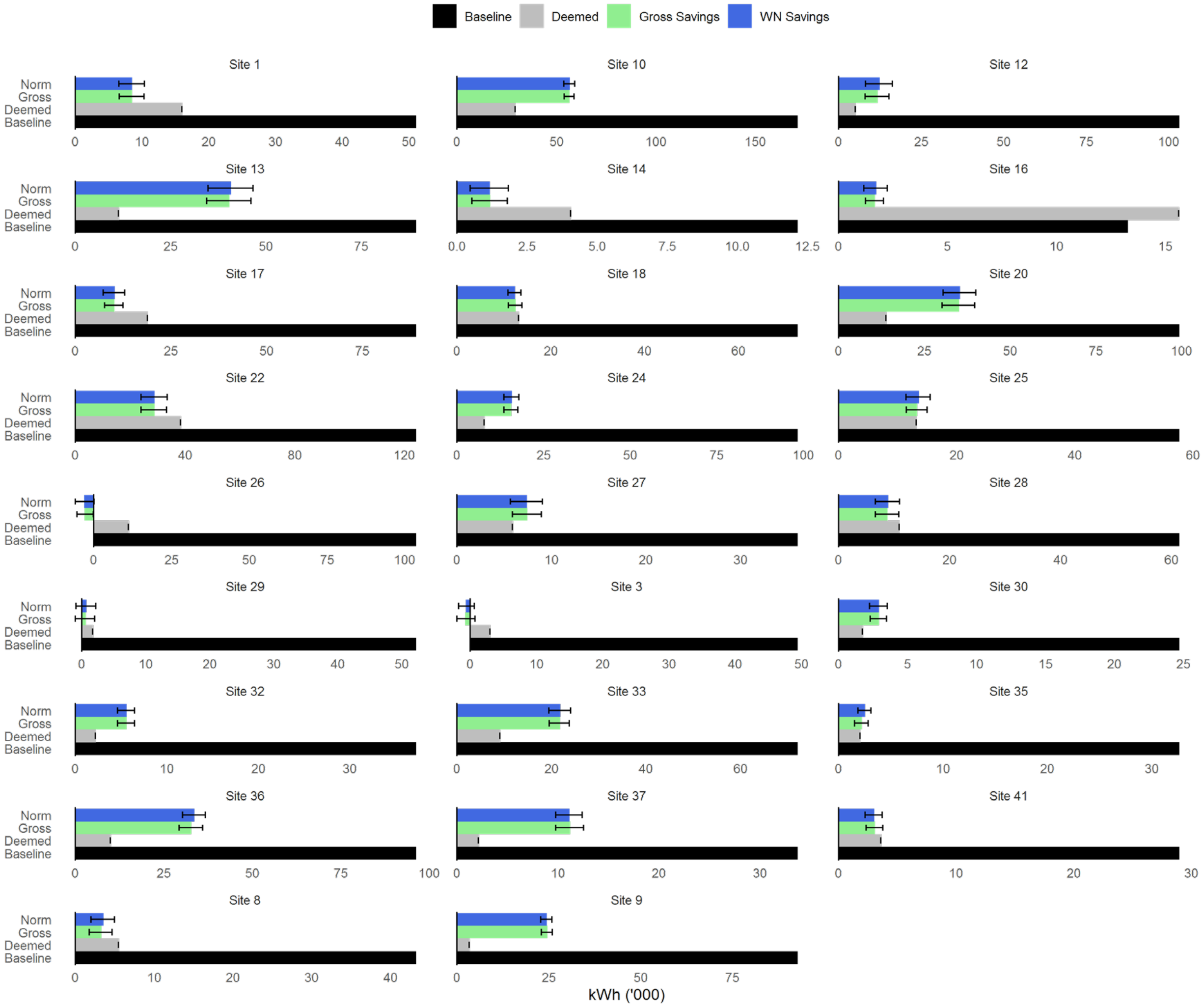
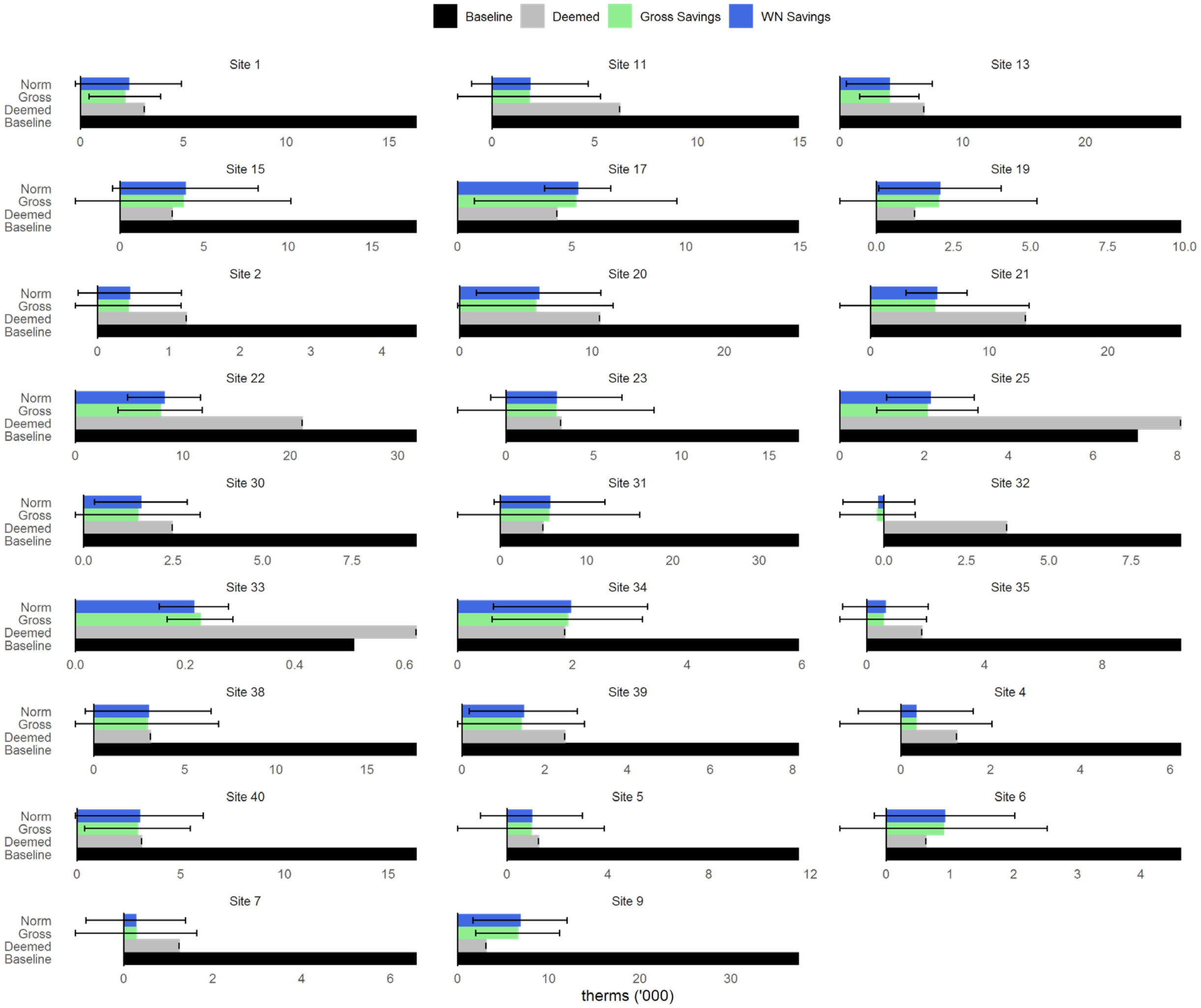


Figure 5.6: Site-Level Gas Savings



6. Net Savings Interview

Property managers at all MF HOPPs properties were solicited in 2020 to participate in a structured interview aimed at determining the likelihood that the managers would have installed retrofits in absence of the program incentives (Appendix B). These interview responses can be used to calculate a *net-to-gross ratio*, the ratio of net savings caused by the program to gross savings associated with the program retrofits. Multiplying the gross savings by the net-gross savings produces an estimate of the program's net savings.

Evergreen Economics developed a phone interview guide focused on non-routine events such as participation in other energy efficiency programs, installation of other energy efficiency measures, construction events, equipment and maintenance changes, and extended power outages. In addition, the interview guide included questions that would inform a net-to-gross calculation. Evergreen modeled the interview guide based on the 2018 Custom Industrial, Agricultural, and Commercial impact evaluation (See Appendix B for the interview guide). The contacts were first invited via email to participate in a phone interview about their experience with MF HOPPs and offered a check of $75 for completing the interview. The participants were subsequently called to request their participation in our interview; participants were contacted by phone at least three times and were left three voicemails. Depending on the reliability of the email addresses provided to us or given to us via our phone calls, we sent one more follow-up email to request their participation.

The interview was completed by only three of the forty-one program managers that received solicitation. The low response rate was due to COVID-19, the long duration between measure installation and the interview deployment, and a lack of good contact information for MF property managers/owners.

Table 6.1: Interview Respondents' Attributes

	Property 1	Property 2	Property 3
Were there external sources of funding for cost of program equipment?	None	None	Don't know
Number of apartments	108	60	193
Number of multifamily complexes company owns or manages	13	Don't know	"Quite a few"
Total number of apartments company owns or manages	Approximately 1,200	Don't know	Over 2,000
Have other properties participated in SDG&E programs?	No. No other properties located in SDG&E territory	Don't know	Don't know
Number of years participant worked at the property	13 years	2 years	5 years
Average occupancy rate	95%	85%	95%

Only one of the three respondents reported any non-routine events. They noted that they replaced their roofs, which likely influenced energy usage for four to five months in 2020. This site was observed through utility bills to have a significant increase in electricity usage from the summer of 2017 and on. When asked what was likely to have caused this increase, they attributed it to a both a heatwave in summer 2017 and increased temperatures overall.

Responses from two of the interviewees indicated that they were influenced by MF HOPPs, while one respondent indicated that they would have done the same thing with or without MF HOPPs. More specifically:

- One respondent reported that they would have installed standard efficiency equipment within one year.
- A second respondent reported that they would have installed equipment more efficient than code but less efficient than what they installed through MF HOPPs, but not for another two to three years.
- A third respondent reported that they would have done the exact same thing at the exact same time.

Due to the small number of completed interviews, the net-to-gross ratio for MF HOPPs is heavily affected by the way a single interviewee answers a question. The three responses indicated a net-to-gross ratio of 0.61, slightly greater than the ex-ante estimate of 0.55 for high-efficiency boilers and LEDs. Because of the small number of responses, Res-Intel/Evergreen determined that the net-to-gross interview estimates are not statistically valid.

Therefore, the ex-ante net-to-gross savings ratio of .55 for boilers and LEDs was utilized, which resulted in estimated net savings of 197 MWh and 38,073 therms from the MF HOPPs program.

7. Uncertainties

During our analysis of the MF HOPPs program, we encountered several obstacles and uncertainties that could undermine the reliability of the NMEC savings estimates. We've already noted some of these uncertainties, including the deficiencies in implementer reports, difficulties arising from concurrent participation in multiple energy-efficiency programs and model-based prediction error, which can lead to wider savings confidence intervals. In addition to the uncertainties already noted, there are two additional issues to highlight: (1) difficulty detecting non-routine events and (2) challenges in measuring savings that represent a modest share of baseline consumption.

Non-Routine Events

Non-routine events (NREs) affecting energy consumption are often easy to identify visually but difficult to detect using statistical methods. To aid with visual inspection of the data, Res-Intel developed a web-based dashboard that allows users to query and plot individual meters' loadpaths interactively. Using this dashboard, Res-Intel identified many NREs that are obvious upon close inspection but are not revealed by traditional model uncertainty metrics such as R^2 or CVRMSE. Conversely, energy baselines that do not contain NREs can sometimes perform poorly along these same statistical metrics. The inadequacy of traditional metrics in detecting NREs highlights a critical need for new and improved detection methods.

To illustrate this point, Figure 6.1 compares the baseline (white) and performance period (green) loadpaths for two meters: (a) a meter that has at least two visible NREs in the baseline period, occurring around Jan. and Oct. 2017, (b) a meter that does not have any visible NREs during the baseline period. Although the baseline model (a) contains visible NREs, it records a very high R^2 value of 0.89—much higher than the R^2 of 0.35 recorded by model (b). If we were to evaluate baseline model quality solely on the basis of ASHRAE recommended metrics such as R^2, we would conclude that model (a) is highly accurate while model (b) is so inaccurate that it is inadequate for evaluation. However, visual inspection suggests that model (a) is far less reliable due to the presence of NREs; and, moreover, model (b) does indeed render a reasonable baseline prediction for the performance period.

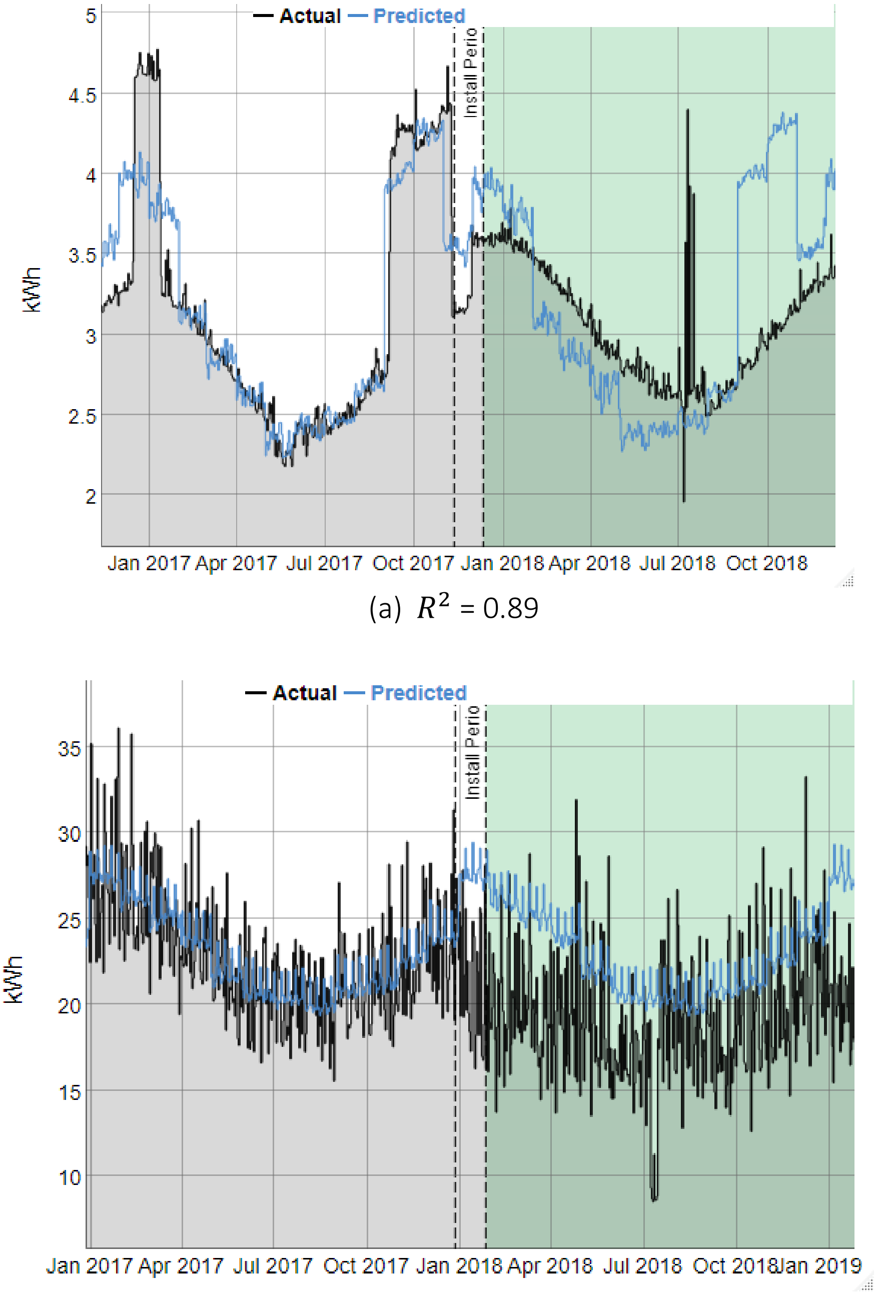
Future NMEC evaluations could benefit by using model-quality metrics that have a greater focus on detecting NREs. Res-Intel has experimented with several methods for outlier detection, using Dynamic Time Warping, Euclidian Distance, and Symbolic Aggregation Approximation. The results of this analysis are outlined in a previous memo delivered by Res-Intel.[1] Other researchers have also experimented with finding NREs using methods such as change-point detection (Touzani et al. 2019) and daily loadpath classification (Miller 2015). The literature on NRE detection is growing but there is yet to emerge a single production-ready method. We propose that a successful NRE detection method will have the following two attributes:

1. The ability to distinguish one-time events from repeated seasonal events.
2. A focus on out-of-sample predictive accuracy.

The current evaluation metrics described in ASHRAE and many of those proposed in the emerging literature fail to satisfy both requirements. We believe the current metrics place too much emphasis on within-sample prediction accuracy for a single baseline year. Improvements in methods will likely require a **multi-year baseline** in order to cross-validate models against out-of-sample seasonal fluctuations. Extending the model baseline period places a greater burden on data retrieval but may be necessary in order to distinguish one-time events from seasonal patterns.

[1] "Outlier Detection and Adjustment Proposal," June 14, 2019.

Figure 6.1: NRE Comparison



(a) $R^2 = 0.89$

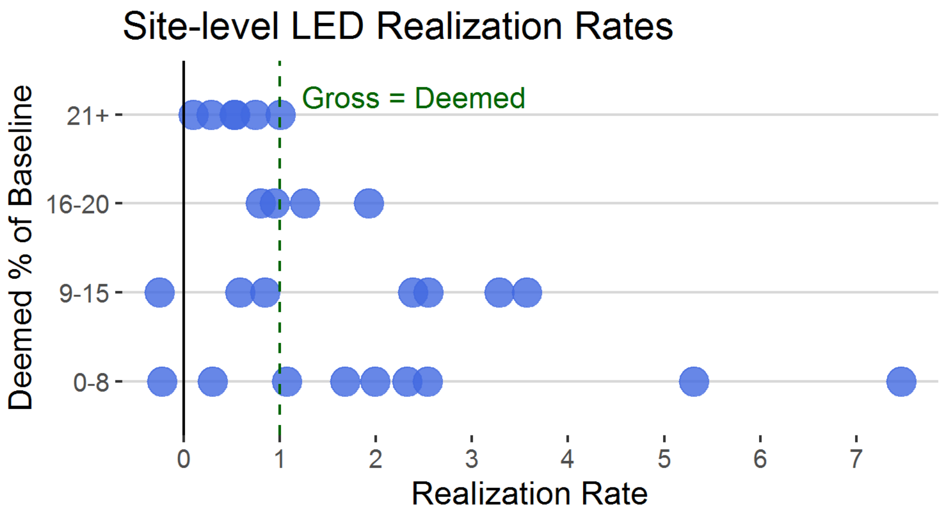
(b) $R^2 = 0.35$

Low Savings-to-Baseline Ratio

It is difficult to detect an ECM's savings if they are small in magnitude relative to the overall baseline consumption recorded at its associated meter. High variation in baseline consumption reduces the signal-to-noise ratio when measuring small changes in consumption. Large levels of baseline consumption can therefore make it difficult to distinguish ECM savings from routine variation that would typically be observed at the meter. This can have the effect of attenuating, or in some cases magnifying, the savings estimated for ECMs.

Figure 6.2 illustrates this point by plotting the ratio of deemed savings to baseline consumption against the distributions of realization rates. The bottom bin includes sites (represented by a blue dot) where deemed savings represented less than 8 percent of the site's baseline consumption. Realization rates among these sites vary widely, consistent with the hypothesis that a low savings-to-baseline ratio causes uncertainty in NMEC savings estimation. Among the sites in this group are Bonita Hills, SOFI Poway and the Dorchester, which all reported zero gross savings despite having substantial deemed savings predictions. Moreover, looking on to the second, third and fourth bins one can see that increasing the savings-to-baseline ratio reduces the spread in realization rates, consistent with the idea that these sites showed lower uncertainty in savings measurement.

Figure 6.2: Deemed Savings Ratio vs. Realization Rate



The measurement difficulties that arise from meters with low savings-to-baseline consumption highlight the importance of using a retrofit isolation approach that ensures accuracy in identifying affected meters. Whole building approaches to conducting NMEC analysis for multifamily retrofits are unlikely to be successful because baseline consumption at these sites can be much larger than expected ECM savings.

Additionally, **implementers must be very careful to select *only* the meters directly affected by ECMs** to ensure extra load is not included in the baseline, reducing the signal-to-noise ratio of the NMEC models. In our analysis, Res-Intel discovered that **only 159 of the 214 (74%) electricity meters reported by implementers showed savings** in the performance period, and savings were statistically significant for only 144 of those meters. These findings raise the possibility that many of the meters reported by implementers were not affected by the ECMs. To the extent that the addition of unaffected meters increases the baseline consumption, their inclusion in the analysis diminishes the quality of the baseline models and the savings estimates.

8. Recommendations for Future Evaluations

Res-Intel's NMEC evaluation of the MF HOPPs program has revealed several important lessons that could broadly inform the planning and execution of future NMEC projects, particularly those focused on multifamily residential or commercial building impact assessments. Based on the results summarized in this report, we submit the following recommendations broken out into two categories for clarity:

Program Design and Data Collection

- Program overlap is difficult to measure empirically and can be better addressed with improved data collection during the retrofit process to isolate savings impacts:
  - Tracking service point numbers, affected meter numbers, and service addresses is essential: tracking account names and/or property names is less important.
- Concurrent participation in multiple energy efficiency programs does not necessarily make NMEC analysis infeasible. However, it does limit the scope of the analysis and could require omitting some customers from the final evaluation.
- Attempt to follow-up on Net to Gross interviews as soon as possible after program participation.
  - Update project tracking data to include detailed project contact information including name, phone number, and email address.
- It is fundamentally important that program implementers focus on accurately reporting (i) affected meter numbers and (ii) deemed savings values. Our evaluation produced some evidence that implementers may have over-reporting affected meters, which violates the retrofit isolation approach and attenuates NMEC savings estimates.
- Utilities can hedge against implementer misreporting by compiling a meter-to-service point mapping that allows evaluators to access all meters associated with a given property.
  - In multifamily and commercial evaluations, the mapping should also separate common-area from tenant meters whenever possible.

    - Res-Intel has performed this meter-to-service point mapping for all residential and commercial meters for other projects for SCE, SDG&E, and PG&E and it has proven feasible and cost-effective.

NMEC Methods and Tools

- Evaluators should be careful not to place too much confidence on site-level evaluations due to data quality issues. Program or population-level estimates tend to align more closely with expectations, while site-level savings estimates can vary substantially.
- NMEC savings estimates should be accompanied by confidence intervals whenever possible to ensure that savings uncertainty is properly communicated.
- Researchers and evaluators must focus on developing better statistical methods for detecting whether baseline data is inadequate due to the presence of non-routine events.
    - Current statistical measures do not detect non-routine events and can convey false optimism about the quality of baseline data.
- Improvements in non-routine event detection may require using multi-year baseline periods to distinguish one-time events from seasonal patterns in energy consumption.
- Machine learning methods represent a viable alternative to traditional statistical models of energy consumption used for NMEC.
    - However, using machine learning methods appeared to have little effect on the final savings estimates for this sample.

References

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

Granderson, J., Touzani, S., Fernandes, S., & Taylor, C. (2017). Application of automated measurement and verification to utility energy efficiency program data. *Energy and Buildings*, 142, 191-199.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 27(2), 83-85.

International Performance Measurement and Verification Protocol (2002). In Concepts and Options for Determining Energy and Water Savings, DOE/GO-102002-1554, Volume 1; US Department of Energy, Office of Energy Efficiency and Renewable Energy: Washington, DC, USA.

Itron (2017). 2013-2015 Regional Energy Networks Multifamily Programs Impact Evaluation Final Report. Prepared for the Energy Division of the California Public Utility Commission. CALMAC Study ID: CPU0150.

Mathieu, J. L., Price, P. N., Kiliccote, S., & Piette, M. A. (2011). Quantifying changes in building electricity use, with application to demand response. *IEEE Transactions on Smart Grid*, 2(3), 507-518.

Miller, C., Nagy, Z., & Schlueter, A. (2015). Automated daily pattern filtering of measured building performance data. *Automation in Construction*, 49, 1-17.

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, 158, 1533-1543.

Touzani, S., Ravache, B., Crowe, E., & Granderson, J. (2019). Statistical change detection of building energy consumption: Applications to savings estimation. *Energy and Buildings*, 185, 123-136.

Appendix A

Baseline Model Evaluation

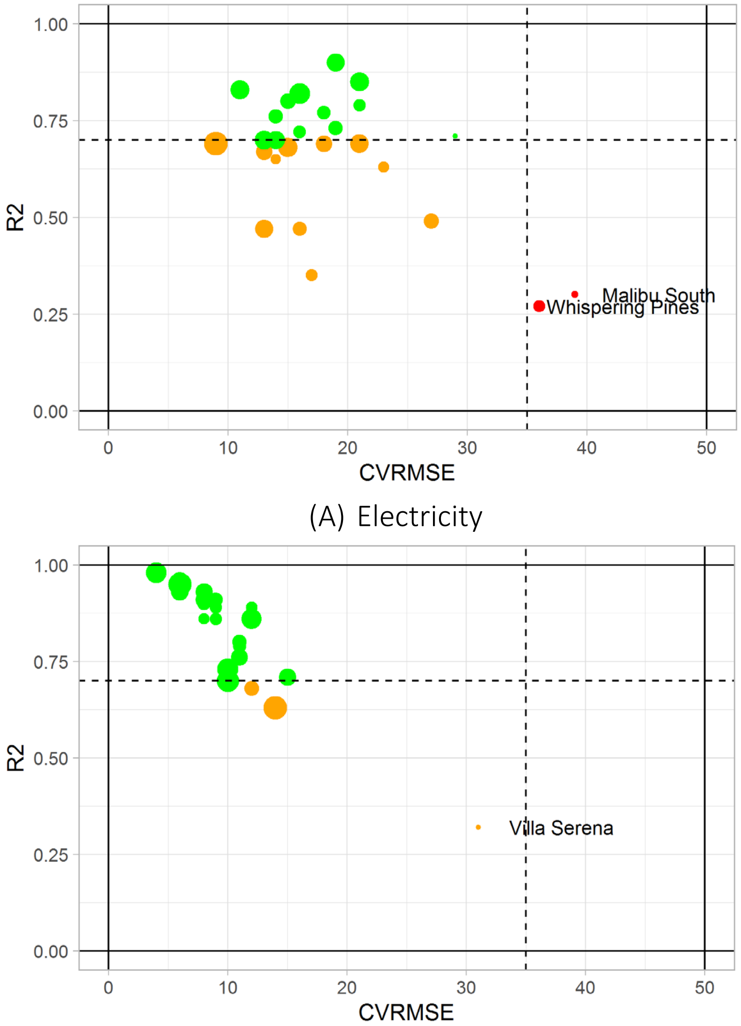
The quality of NMEC evaluated savings depends in large part on the level of uncertainty in the underlyingstatistical baseline energy models. ASHRAE guideline 14 recommends relying on baseline models that have a maximum CVRMSE of 35, a minimum R^2 of 0.7 and a maximum NMBE of 0.005. Table A.1 summarizes the distribution of each of these metrics among the TOWT models estimated for our sample of MF HOPPs sites. The median baseline model for gas and electric use satisfies all three ASHRAE requirements.

All models satisfy NMBE requirements by a significant margin, though a small minority of electricity and gas models fail to satisfy the requirements for R^2 and CVRMSE. Figure A.1 plots CVRMSE against R^2 for each site, and highlights the underperforming models in red. The size of the circles in the scatterplot indicates the total baseline consumption recorded at the site. Among electricity models, we find particularly poor model performance at Malibu South and Whispering Pines, and among gas models we document poor performance at Villa Serena. Fortunately, these sites comprise only a small minority and have relatively low baseline consumption, indicated by their small size on the scatter plot. Their site-level model uncertainties, therefore, signify only a small detriment to the overall program evaluation.

Table A.1: Baseline Model Metrics

	Electricity			Gas		
Percentile	R^2	**CVRMSE**	**NMBE**	R^2	**CVRMSE**	**NMBE**
0	0.27	9	9×10^{-15}	0.32	4	5×10^{-14}
25	0.64	14	5×10^{-14}	0.74	8	2×10^{-14}
50	0.7	16.5	9×10^{-14}	0.88	9	5×10^{-14}
75	0.77	21	1×10^{-13}	0.91	11	1×10^{-13}
100	0.9	39	2×10^{-13}	0.98	31	3×10^{-13}

Figure A.1: CVRMSE vs. $\boldsymbol{R^2}$



(A) Electricity

(B) Gas

GBM versus TOWT Comparison

This section compares the quality of the TOWT and GBM models using a subset of the MF HOPPs metered consumption data. The gross savings used in the report derive only from the TOWT model predictions. Nevertheless, future evaluations can benefit from insights gained by comparing the two models.

Tables A.2 and A.3 summarize property-level model fit (R^2), model error $(CVRMSE)$ and bias $(NMBE)$ for each type of model. The recommended $R^2$of greater than 0.7 was satisfied by the majority of TOWT and GBM models, having median R^2 values of 0.88 and 0.95. Additionally, about half of the TOWT models satisfied the recommended maximum value of 35 for CVRMSE, while most GBM models performed well below this threshold.

Table A.2: Electricity Model Evaluation

	R^2		CVRMSE		NMBE	
	TOWT	GBM	TOWT	GBM	TOWT	GBM
Mean	0.84	0.92	39	29	-5.64E-14	-0.0069
Median	0.88	0.95	36	27	-7.00E-14	-0.005
Max	0.97	0.98	78	69	1.00E-13	4.00E-04
Min	0.32	0.47	16	10	-2.00E-13	-0.03

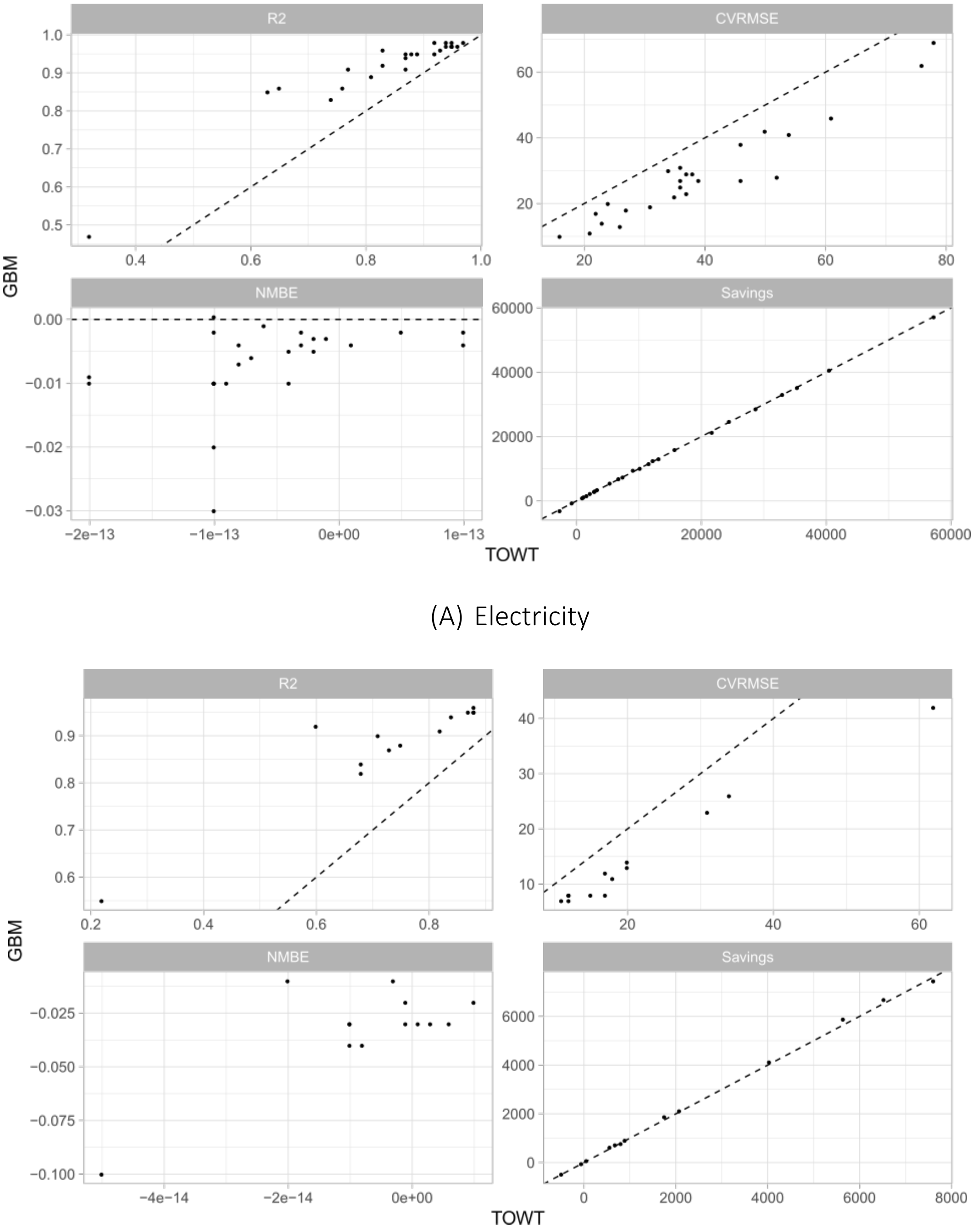
Table A.3: Gas Model Evaluation

	R^2		CVRMSE		NMBE	
	TOWT	GBM	TOWT	GBM	TOWT	GBM
Mean	0.73	0.88	22	14	-7.15E-15	-0.032
Median	0.75	0.91	17	11	-3.00E-15	-0.03
Max	0.88	0.96	62	42	1.00E-14	-0.01
Min	0.22	0.55	11	7	-5.00E-14	-0.1

The GBM performed significantly better on measures of model fit and model error, while performing worse on the measurement of bias. The introduction of bias is not an uncommon feature of machine learning models: machine learning methods are known for trading off increases in bias for reductions in prediction variance (Hastie et al. 2005). It is worth noting, however, that bias is still relatively low among GBM models and typically does not violate the recommended 0.005 threshold set by ASHRAE.

Figure A.2 shows pair-wise comparisons of the three performance criteria for both models. It confirms that GBM indeed dominates TOWT on model fit and model error criteria while performing worse on bias. The final panel plots TOWT measured savings against GBM measured savings. Interestingly, despite substantial differences in model performance, GBM measured savings do not appear to deviate significantly from TOWT measured savings.

Figure A.2: Pair-Wise Model Comparison



(A) Electricity

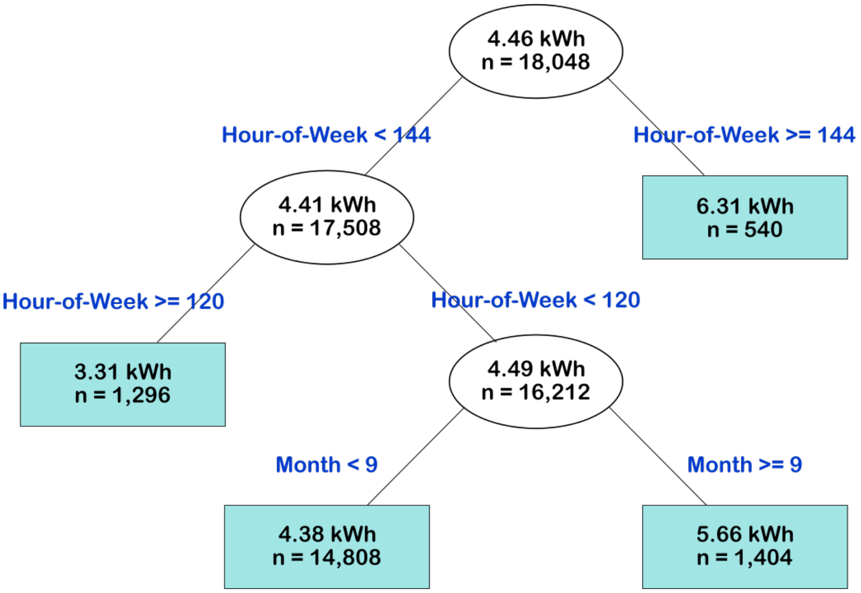
(B) Gas Models

GBM Model Summary

The GBM is a machine learning model that also predicts consumption based on time-of-week, month and temperature, but does so using a more sophisticated algorithm. The fundamental component of the GBM is the decision-tree model. The GBM combines predictions from multiple decision trees to generate a single prediction for consumption during a given time interval.

In simple terms, a tree model predicts consumption based on what bin an hour or day falls into, where bins are determined by a series of conditions on the explanatory variables (time-of-week, month, temperature). These conditions can be expressed as a sequence of splits, forming a decision tree, as illustrated in Figure 3.2.

Figure A.3: Tree Example



The tree begins at the top node, reporting average consumption (4.46 kWh) across all hours (18,048) recorded by the meter. The tree then determines the series of splits in the data that produce the greatest reduction in prediction error, and the terminal nodes in the tree contain the tree's final predictions. For example, this tree predicts 6.3 kWh hourly consumption for any hour in the last day of the week ("Hour of Week > 144"), and it predicts 5.66 kWh for the first five days of the week in months September to December.

The GBM relies on predictions from an *ensemble* of tree models $T_0, T_1 \ldots T_J$. These tree models are generated using the following sequential algorithm:

1. $Initialize\ T_0$

```
2. $For\ j = 1\ to\ J$:
    a. $Compute\ residuals$: $r_t = c_t - T_{j-1}\ \forall\ t.$
    b. $Fit\ regression\ tree\ to\ \{r_t\}$:
        $$T_j(\Theta) = argmin_{\Theta} \sum \left(r_t - T_j(\Theta)\right)$$
    a. $Update\ T_j = T_j + T_{j-1}$
1. $Output\ \hat{c} = T_J$
```

The GBM requires setting some additional *hyper*-parameters including the number of trees (J), the *scale* parameter and the maximum depth of each tree. We estimate these parameters using a cross-validation procedure. Setting the correct hyper-parameters is critical to ensuring that the model does not *overfit* the data and thereby underperform when predicting out-of-sample.

The preceding is a very brief explanation of the GBM and the reader may wish to refer to Touzani et al. (2018) for a more detailed summary of the GBM and its application to NMEC. They show that the GBM offers considerable improvements in accuracy and performance compared to linear regression models such as TOWT.

Site-Level Savings (Cont.)

Table A.4: Site-Level Electricity Savings

Site	Baseline	Deemed	Deemed/Baseline	Gross	RRate
Bella Vista	50,933	15,948	0.31	8,429	0.53
Bonita Hills	49,326	2,964	0.06	-664	-0.22
College Campenile	43,147	5,460	0.13	3,212	0.59
Colonnade at Fletcher Hills	92,535	3,276	0.04	24,453	7.46
Coral Bay	170,769	29,172	0.17	56,302	1.93
Forest Park	102,930	4,992	0.05	11,617	2.33
Gramercy	89,190	11,284	0.13	40,307	3.57
Grand Regency	12,120	4,056	0.33	1,164	0.29
Malibu South	13,237	15,600	1.18	1,640	0.11
Naples Court	88,873	18,856	0.21	10,043	0.53
Oak Valley	72,494	13,104	0.18	12,386	0.95
Palomar	99,115	13,728	0.14	34,949	2.55
Park Villas	123,896	38,328	0.31	28,596	0.75
Seawind	98,114	7,800	0.08	15,524	1.99
Shadow Hill	57,568	13,104	0.23	13,202	1.01
SOFI Poway	103,357	11,136	0.11	-2,768	-0.25

Spring Valley Apts	35,976	5,858	0.16	7,391	1.26
Stone Arbor	61,256	10,920	0.18	8,720	0.8
The Dorchester	52,100	1,716	0.03	523	0.3
Towne Centre	24,659	1,716	0.07	2,882	1.68
Villa Pacific	37,151	2,184	0.06	5,547	2.54
Villa Serena	71,965	9,076	0.13	21,644	2.38
Vista Capri North	32,578	2,028	0.06	2,180	1.08
Vista Del Sol	96,014	9,930	0.1	32,603	3.28
Whispering Pines	33,671	2,100	0.06	11,153	5.31
Woodland Hills	28,929	3,588	0.12	3,046	0.85

Table A.5: Site-Level Gas Savings

Site	Baseline	Deemed	Deemed/Baseline	Gross	RRate
Bella Vista	16,297	3,104	0.19	1,722	0.55
Colonnade at Fletcher Hills	37,418	3,104	0.08	4,641	1.49
Gramercy	27,730	6,830	0.25	3,344	0.49
Naples Court	14,920	4,346	0.29	4,972	1.14
Palomar	25,576	10,555	0.41	5,075	0.48
Park Villas	31,699	21,110	0.67	7,080	0.34
Shadow Hill	7,047	8,071	1.15	1,810	0.22
Towne Centre	9,286	2,484	0.27	1,265	0.51
Villa Pacific	8,992	3,725	0.41	-354	-0.09
Villa Serena	507	621	1.22	192	0.31
Vista Capri North	10,646	1,863	0.17	450	0.24
Bonita Glen	4,473	1,242	0.28	389	0.31
Bonita Mesa	6,225	1,242	0.2	320	0.26
Bonita Terrace	11,526	1,242	0.11	921	0.74
Bonita View	4,618	621	0.13	824	1.33
Bonita Woods	6,570	1,242	0.19	252	0.2
Foothill Courtyards	14,922	6,209	0.42	248	0.04
La Jolla Terrace	17,600	3,104	0.18	3,684	1.19
Pacific View	9,868	1,242	0.13	1,934	1.56
Park Grossmont	26,125	13,038	0.5	5,089	0.39
Rancho Bonita	16,670	3,104	0.19	2,761	0.89
Villa Bonita	34,515	4,967	0.14	5,619	1.13
Village View	5,932	1,863	0.31	1,854	1

Windsor Manor	17,679	3,104	0.18	2,892	0.93
Windsor Manor II	8,125	2,484	0.31	1,383	0.56
Windsor Manor III	16,334	3,104	0.19	2,871	0.92

Appendix B

Net-to-Gross Interview Report

EVERGREEN MF HOPPS