NATURAL GAS MODEL ACCEPTANCE CRITERIA RESEARCH AND DEVELOPMENT

FINAL REPORT

PREPARED FOR

SOUTHERN CALIFORNIA GAS COMPANY
PACIFIC GAS & ELECTRIC COMPANY
SAN DIEGO GAS & ELECTRIC COMPANY

PREPARED BY

kW ENGINEERING
287 17TH ST, SUITE 300
OAKLAND, CA, 94612
510.834.6420

SEPTEMBER 28, 2022

kW ENGINEERING

Executive Summary

California has adopted a new approach to capturing the savings potential in existing buildings. Known as Normalized Metered Energy Consumption (NMEC), this approach requires that an accurate regression-based or other data-driven model be developed based on a year of a customer's energy usage and independent variable data. The independent variable data typically includes the ambient temperature and often other influential parameters, such as time of use and building operation modes.

To participate in a site-level NMEC program, the customer's data-driven model's goodness of fit metrics must meet certain "criteria" as specified in the California Public Utilities Commission's (CPUC) NMEC Rulebook (CPUC 2020).[1] A key metric is the coefficient of variation of the root mean squared error CV(RMSE), which is one measure of the random error between a model and the energy data it is developed from. Good models have small values of CV(RMSE). According to the CPUC Rulebook acceptable models must have values lower than 25%.

By definition, CV(RMSE) is normalized by average energy use over the period. Because of this normalization, many commercial building's natural gas energy models often fail to meet the CV(RMSE) criterion due to low energy usage in warmer months. Figure 1 shows examples of low natural gas use during these months. For such buildings the annual average energy use is low and because it is in the denominator, the CV(RMSE) is high, often above the 25% threshold. Visually, the model may track very well with the usage data, be a good predictor of energy use and enable reliable savings estimations. However, the project does not qualify for a gas NMEC program because it does not meet the CV(RMSE) criterion. Because of this, many building types including commercial offices, government, educational, and similar buildings are unable to access natural gas NMEC programs.

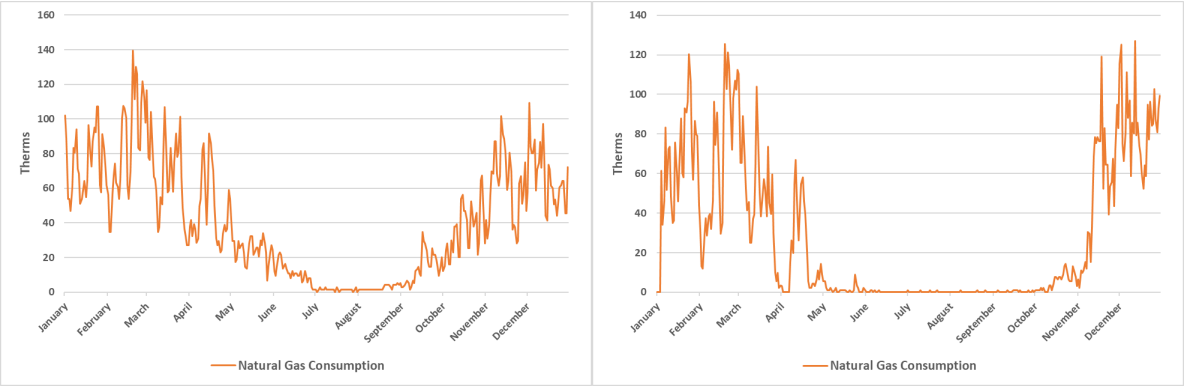


Figure 1. Examples of low use periods in commercial building natural gas consumption. The charts show the daily gas consumption over a typical year.

This research work was designed to address the problem of natural gas NMEC projects often failing the model acceptance criteria through a review of the literature for additional insight, employing and testing different modeling algorithms and modeling strategies, and developing and

[1] The Rulebook provides acceptable levels of goodness-of-fit metrics, but the industry takes these as criteria.

testing alternate model acceptance criteria. It was designed to answer the following research questions:

1. What are appropriate modeling algorithms or strategies that accurately model gas use in commercial buildings?
2. What alternate or more generalized acceptance criteria may be used to overcome participation barriers faced by natural gas commercial building NMEC projects due to failure to meet the goodness of fit criteria described in the CPUC's NMEC Rulebook 2.0? Should the acceptance criteria focus on model goodness of fit metric values, predictive accuracy, or on savings uncertainty?
3. Are there more generalized model acceptance procedures and criteria that should be followed for cases when natural gas usage is low in commercial buildings during some portions of the year?
4. What changes to the modeling acceptance criteria as described in the CPUC NMEC Rulebook 2.0 should be proposed as a result of this research?

A literature review was conducted to inform the selection of different modeling methods and strategies, as well as to identify different acceptance methods – whether they be based on alternate model goodness of fit metrics or savings uncertainty. A data set was requested of California natural gas utilities to assure an adequately large representation of commercial buildings throughout California's different climate zones and utility service areas.

The modeling algorithms, modeling strategies, and proposed acceptance criteria were tested using the data set. Distributions of results were tabulated and then analyzed to identify potential improvements in modeling and acceptance criteria that may help more buildings and customers participate in natural gas NMEC efficiency programs.

The findings from this work may also benefit cases in which any building energy use commodity (e.g. steam, chilled or hot water, electricity) has seasonally low usage. The work also provides more general insight on energy modeling useful to enable more buildings to participate in California's NMEC programs.

Recommendations and Conclusions

In this work, we examined alternate goodness of fit criteria to overcome the cause of many natural gas models failing the established goodness of fit criterion of CV(RMSE) < 25%. After assembling a dataset consisting of Affected and Unaffected buildings (buildings with gas use models not meeting and meeting the criterion, respectively), we ran two different modeling algorithms, calculated several alternate goodness of fit criteria, and compiled summary tables for additional evaluation of the alternate metrics.

Before proceeding, it is important to note that we did not examine individual building gas datasets to determine the presence of poor data or unidentified non-routine events as would be done on a case-by-case basis in preparation of participation in a site-level NMEC program. This would require additional information from each building to be collected, regimes of operations to be identified, and additional analysis to be performed. This work instead proceeded by modeling the data as is without additional insight into each building that otherwise may have improved

individual building gas models. For every site-level NMEC project, we recommend this preliminary analysis of the data be pursued.

Gas usage data typically have more data quality issues than electricity use data. Gas usage is usually measured in units of volume, which requires flow measurements. Flow measurement is typically less accurate than electric measurement. Gas data is typically more 'noisy' in that it has more day-to-day random fluctuations, often has large periods of missing data, and may reflect the manual operation of gas-consuming equipment in commercial buildings. It is acknowledged that there are more data quality issues to address with natural gas data.

The goodness of fit metric CV(RMSE) quantifies the amount of random error between a model and the data the model is developed from. For commercial buildings that generally use natural gas for space and water heating only, it is a common reason natural gas models fail to become NMEC projects. A literature search suggested several alternate goodness of fit metrics: a weighted mean absolute percent error (wMAPE), a range-normalized root mean squared error (nRMSE), and a root mean squared error normalized by total energy use (tRMSE). The fractional savings uncertainty (FSU) developed by Claridge and Reddy (2000) and used in ASHRAE Guideline 14-2014 combined the CV(RMSE) and savings in a new metric that directly addresses the important question of how accurate will the resulting savings estimate be given the proposed model? ASHRAE's FSU was tested as an alternate qualifying metric. Another metric was to separate the low use and high use periods and model them separately, weighting the CV(RMSE) by that period's total gas consumption.

We tested two different modeling algorithms, the time-of week and temperature (TOWT) model (Mathieu, et.al. 2011) and ASHRAE's three-parameter change-point model (3PH) (Kissock et. al, 2004), and performed a manual modeling strategy by separating the low use from high use period data and modeling separately with the TOWT model.

After testing 635 building data sets with the TOWT algorithm, we found 50% of the buildings did not pass the current CV(RMSE) criterion. When using the FSU as a metric on these failed buildings, we found that 13% of them passed when assuming the project would yield 10% savings or more. When the same buildings were modeled with the 3PH model, 45% initially failed the CV(RMSE) criterion, but 10% of the failed buildings passed when using the FSU.

The binary classification analysis was used to evaluate the CV(RMSE), FSU, wMAPE, nRMSE, tRMSE, and 3PH model wCV(RMSE). This analysis confirmed that the CV(RMSE), FSU, and wCV(RMSE) metrics were superior to the wMAPE, nRMSE, and tRMSE metrics, as they showed high true positive rates along with reasonably low false positive rates. False positive rates were too high for wMAPE and nRMSE. The results were consistent whether using the TOWT or 3PH modeling algorithms (the wCV(RMSE) with was tested only for the 3PH model).

Manual separation of the low gas use and high gas periods, development of separate models for each period, then calculating the energy-weighted CV(RMSE) of the two models showed that when the CV(RMSE) of a model built on the full dataset did not excessively exceed the 25% criterion, this strategy could be used to qualify more gas NMEC projects.

Based on this work, it follows that alternate metrics and modeling strategies may be used to qualify natural gas site-level NMEC projects. Our recommendations are summarized below:

1. Examine the gas use data for data quality issues. Assure a full dataset is obtained for each building. Identify and resolve data quality issues such as outliers and extensive gaps in gas use throughout the baseline year. Determine whether a different modeling approach or whether different regimes of operation or non-routine events are present. Obtain information from building operators to substantiate modeling assumptions and strategies.

2. Use the current CV(RMSE) criterion of 25% to determine whether a natural gas NMEC project is acceptable (models based on daily or monthly time interval data only). Should the model fail the CV(RMSE) test, calculate the FSU assuming 10% savings. If there is a savings estimate available, use it in the FSU equation instead. If the FSU is < 50% at a 90% confidence level, accept the building as an NMEC project. FSU may be used as a criterion as long as the model is an ordinary lease squares regression-based algorithm.

3. If the current gas model fails the CV(RMSE) criterion by a small amount such as 5%, consider separating the low gas use from the high gas use period and modeling each period separately. Calculate the energy-weighted average CV(RMSE) from the two models individual CV(RMSE). If the weighted average CV(RMSE) passes the criterion, accept the building as an NMEC project.

Table of Contents

Abbreviations

ASHRAE – American Society of Heating Refrigeration and Air-conditioning Engineers

ARIMA – Autoregressive Integrated Moving Average

CDD – Cooling Degree Day

CPUC – California Public Utilities Commission

FSU – Fractional Savings Uncertainty

FNR – False Negative Rate

FPR – False Positive Rate

GLM – Generalized Linear Model

GOF – Goodness of Fit (refers to regression model fitness or accuracy metric)

HDD – Heating Degree Day

HVAC – Heating Ventilating and Air-Conditioning

IPMVP - International Performance Measurement and Verification Protocol

ISD – Integrated Surface Database

LBNL – Lawrence Berkeley National Laboratory

MAPE – Mean Absolute Percent Error

MSE – Mean Squared Error

NAICS – North American Industry Classification System

NMBE – Normalized Mean Bias Error

NMEC – Normalized Metered Energy Consumption

NOAA – National Oceanic and Atmospheric Administration

OAT – Outside Air Temperature

OLS – Ordinary Least Squares

RMSE – Root Mean Squared Error

- **CV(RMSE)** – Coefficient of Variation of the Root Mean Squared Error
- **wCV(RMSE**) – weighted Coefficient of Variation of the Root Mean Squared Error
- **n(RMSE)** – range normalized Root Mean Squared Error
- **t(RMSE)** – total energy use normalized Root Mean Squared Error

TOWT – Time-of-Week and Temperature (refers to a modeling algorithm)

TPR – True Positive Rate

TNR – True Negative Rate

wMAPE – weighted Mean Absolute Percent Error

3PH – Three Parameter Heating (refers to a modeling algorithm)

Introduction

California has adopted a new approach to capturing the savings potential in existing buildings. Leveraging the short time interval data made available from the widespread installation of advanced meters throughout the state, utilities and third parties are offering meter-based program approaches to customers. In these programs the savings are quantified based on the difference between baseline and post-installation period energy use, each normalized to a common set of conditions. Known as Normalized Metered Energy Consumption (NMEC), this approach requires that an accurate regression-based or other data-driven model be developed based on a year of a customer's energy usage and independent variable data. The independent variable data typically includes the ambient temperature and often other influential parameters, such as time of use and building operation modes.

To participate in a site-level NMEC program, the customer's data-driven model's goodness of fit metrics must meet certain "criteria" as specified in the California Public Utilities Commission's (CPUC) NMEC Rulebook (CPUC 2020).[2] A key metric is the coefficient of variation of the root mean squared error CV(RMSE), which is one measure of the random error between a model and the energy data it is developed from. Good models have small values of CV(RMSE). According to the CPUC Rulebook acceptable models must have values lower than 25%.

By definition, CV(RMSE) is normalized by average energy use over the period. Because of this normalization, many commercial building's natural gas energy models often fail to meet the CV(RMSE) criterion due to low energy usage in warmer months. Figure 2 shows examples of low natural gas use during these months. For such buildings the annual average energy use is low and because it is in the denominator, the CV(RMSE) is high, often above the 25% threshold. Visually, the model may track very well with the usage data, be a good predictor of energy use and enable reliable savings estimations. However, the project does not qualify for a gas NMEC program because it does not meet the CV(RMSE) criterion. Because of this, many building types including commercial offices, government, educational, and similar buildings are unable to access natural gas NMEC programs.

While not unique to natural gas use in buildings, this problem is far more prevalent than electric or other energy uses in buildings.

CV(RMSE) is only one metric used to evaluate the appropriateness of a model. Other goodness of fit metrics, such as those discussed in this report and others, can be used and may be more appropriate, depending on the researchers needs.

This research project evaluated alternate modeling methods and model acceptance criteria to determine how more natural gas projects could participate in site-level NMEC energy efficiency programs.

[2] The Rulebook provides acceptable levels of goodness-of-fit metrics, but the industry takes these as criteria.

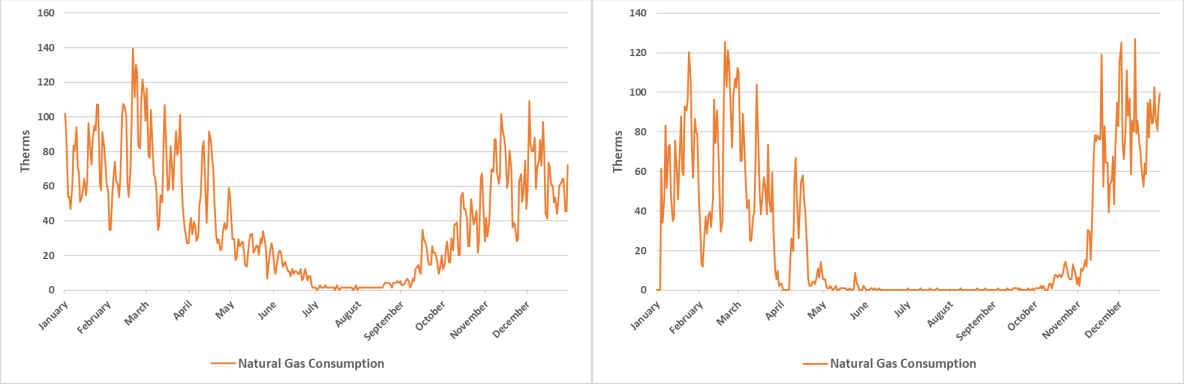


Figure 2. Examples of low use periods in commercial building natural gas consumption. The charts show the daily gas consumption over a typical year.

Background

Quantifying savings based on the reduction in a building's normalized metered energy consumption is the same approach as the International Performance Measurement and Verification Protocol's Option C: Whole Facility approach (IPMVP, 2016). Using data monitored over time savings are determined from empirical energy models that accurately describe baseline and reporting period energy use behavior. The energy models are typically based on regression methods developed from available energy use and independent variable data, which generally include ambient temperature, but may also include time of use, building occupancy, or other parameters.

For NMEC savings analysis we are concerned with how accurately a model can predict energy use under the expected conditions. Model accuracy may be achieved through use of different modeling algorithms or modeling strategies. How we quantify modeling accuracy through use of modeling metrics plays a key role as well.

Modeling algorithms have improved from ordinary least squares regression methods, which were sufficient for use with monthly billing data, but inadequate with shorter time interval data. An example includes ASHRAE's piecewise linear change-point modeling algorithms (Kissock, et. al. 2004) that capture energy use behavior with ambient temperature. Another example is Lawrence Berkeley National Laboratory's time-of-week and temperature (TOWT) model (Mathieu, 2011), which captures energy use dependance on both temperature and weekly building operation schedules. Use of more advanced machine learning algorithms is emerging in some projects. The algorithms may be augmented with additional independent variables, which may be indicator variables that identify different building operation modes, including low-gas use periods. More information on ASHRAE's change-point and LBNL TOWT modeling algorithms are provided below.

Energy engineers have used modeling strategies such as separating data from different periods of unique building operations and creating separate models for them. The different building operation periods may be identified from occupied and unoccupied periods, operation periods of equipment, as well as high and low use periods. After models have been developed for the different operation periods, they may be combined using binary variables representing each operation period so that predictions are made correctly.

In California, practitioners implementing NMEC projects assess their models based on model goodness of fit (GOF) guidance provided by the CPUC (described in the Literature Review section below). Three GOF metrics are used with suggested thresholds for acceptance (LBNL, 2019): the coefficient of variation of the root mean squared error, CV(RMSE); the normalized mean bias error, NMBE; and the coefficient of determination, R^2. Each of these GOF metrics are quantified based on energy data and model predictions for the baseline period.

The CV(RMSE) is a measure of random error or how closely any individual data point may be predicted by the model. While smaller values of CV(RMSE) are desired, it is not possible to eliminate the CV(RMSE) so the goal is to minimize it. CPUC's guidance requires CV(RMSE) to be less than 25% for acceptable models.

The NMBE is a measure of bias error, or how accurately the total energy consumption predicted by the model matches that of the measured data. Regression methods are based on minimizing bias error. CPUC's guidance requires the NMBE to be between 0.5% and -0.5%.

R^2 describes how well the independent variables explain the energy use behavior. R^2 values above 0.7 indicate a strong relationship between the independent and dependent variables. Unfortunately, when there is small or no variation in the dependent variable, R^2 values will be low despite how accurate the model may be. R^2 is not an accuracy metric, rather it is useful when comparing one model to another.

Meeting these GOF requirements represent one way for practitioners to determine a project's appropriateness for an NMEC approach. However, they may present artificial barriers for participation of more natural gas projects due to the failure to meet the CV(RMSE) criterion.

Objectives and Work Plan

This research work was designed to address the problem of natural gas NMEC projects failing the model acceptance criteria through a review of the literature for additional insight, employing and testing different modeling algorithms and modeling strategies, and developing and testing alternate model acceptance criteria. It was designed to answer the following research questions:

1. What are appropriate modeling algorithms or strategies that accurately model gas use in commercial buildings?
2. What alternate or more generalized acceptance criteria may be used to overcome participation barriers faced by natural gas commercial building NMEC projects due to a failure of meeting the goodness of fit criteria described in the CPUC's NMEC Rulebook 2.0? Should the acceptance criteria focus on model goodness of fit metric values, predictive accuracy, or on savings uncertainty?
3. Are there more generalized model acceptance procedures and criteria that should be followed for cases when natural gas usage is low in commercial buildings during some portions of the year?
4. What changes to the modeling acceptance criteria as described in the CPUC NMEC Rulebook 2.0 should be proposed as a result of this research?

A literature review was conducted to inform the selection of different modeling methods and strategies, as well as to identify different acceptance methods – whether they be based on alternate model goodness of fit metrics or savings uncertainty. A data set was requested of

California natural gas utilities to assure an adequately large representation of commercial buildings throughout California's different climate zones and utility service areas.

The modeling algorithms, modeling strategies, and proposed acceptance criteria were tested using the data set. Distributions of results were tabulated and then analyzed to identify potential improvements in modeling and acceptance criteria that may help more buildings and customers participate in natural gas NMEC efficiency programs.

The findings from this work may also benefit cases in which any building energy use commodity (e.g. steam, chilled or hot water, electricity) has seasonally low usage. The work also provides more general insight on energy modeling useful to enable more buildings to participate in California's NMEC programs.

Literature Review

A literature review was conducted to identify alternate model goodness of fit and accuracy metrics, modeling algorithms, and modeling strategies. Sources of the literature included statistical texts, industry literature, discussions with experienced modelers and statisticians, and a review of available natural gas NMEC data and modeling approaches.

Goodness-of-Fit Metrics

ASHRAE Guideline 14 and the California NMEC Rulebook prescribe the following three metrics to evaluate the predictive performance of energy models: CV(RMSE), NMBE, and R^2.

As described above, R^2 is an indicator of how well the independent variable explains the variation in the energy use.

Eqn. 1: $\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{n}(E_i - \hat{E}_i)}{\sum_{i=1}^{n}(E_i - \bar{E})}}$, *where* $\boldsymbol{E_i}$ *and* $\boldsymbol{\hat{E}_i}$ *are the model's measured and predicted energy use at each time interval i respectively,* $\boldsymbol{\bar{E}}$ *is the mean energy use over the baseline period, and n is the number of baseline model points.*

The NMBE describes the bias error of a model's predictions versus the data.

Eqn. 2: $\boldsymbol{NMBE = \frac{\sum_{i=1}^{n}(E_i - \hat{E}_i)}{E_{tot}}}$, *where* $\boldsymbol{E_{tot}}$ *is the total baseline energy use.*

CV(RMSE) is calculated as the random error of the model (RMSE) normalized by the average energy use over the period. Note that the random error indicates how well a model follows the load profile.

Eqn. 3: $\boldsymbol{RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{E}_i - E_i)^2}{n-p}}}$, *where* $\boldsymbol{p}$ *are the number of model parameters.*

Eqn. 4: $\boldsymbol{CV(RMSE) = RMSE/\bar{E}}$, *where* $\boldsymbol{\bar{E}}$ *is the mean energy use.*

Reddy and Claridge (2000) describe that there are fundamental differences between the R^2 and CV(RMSE) metrics. They each are normalized indices, but their normalizations are different,

therefore their interpretations are different. While their numerators are similar, their denominators differ. R^2 represents the variation of the dependent variable explained by the model compared to the variation in data about the mean value. CV(RMSE) is the mean variation in the data not explained by the model normalized by the mean energy use. If the accuracy of the savings estimate is the quantity to use when selecting baseline models, the CV(RMSE) is the criteria to use over R^2.

However, as described above for natural gas models, the mean energy use as a normalizing term often unduly influences the CV(RMSE) of the model.

An alternative normalizing term is the range of the energy use data, which is the difference between the maximum and minimum values of energy use in the model training period. This is the range normalized RMSE, or nRMSE:

Eqn. 5: $$\boldsymbol{nRMSE = RMSE/(E_{max} - E_{min})}$$

The range normalized RMSE scales the random variation of the model predictions about the data to the actual range of data without overly compensating for repeated high or low values. The normalized RMSE has been tested in energy models and found to be a meaningful and accurate representation (Chakraborty and Elzarka, 2017).

Similarly, the total energy use may be used as an alternate normalizing term. The total normalized RMSE is:

Eqn. 6: $$\boldsymbol{tRMSE = RMSE/\sum_{i=1}^{n} E_i}$$

Similarly to the range normalized RMSE, the total normalized RMSE scales the random variation of the model predictions about the data to the total amount of energy use of the model training period. Various sources report on the differences in normalizing the RMSE.[3]

An alternate way to calculate a measure of the random error is to weight the contribution to model variation by the amount of energy represented. The weighted CV(RMSE) may be used:

Eqn. 7: $$\boldsymbol{wCV(RMSE)} = \frac{\boldsymbol{E_{low-use} \cdot CV(RMSE)_{low-use} + E_{high-use} \cdot CV(RMSE)_{high\_use}}}{\boldsymbol{E_{low\_use} + E_{high\_use}}}$$

In this case the model is broken up into two distinct use regimes, where E_{low_use} and E_{high_use} are the totals used during those periods, with $E_{low\_use} + E_{high\_use} = \sum_{i=1}^{n} E_i$. The wCV(RMSE) reduces the contribution of the model variation that occurs from low energy use periods of time during the training period. Reddy and Claridge (2000) discuss the use of weighted CV(RMSE) with change-point models.

Other random error metrics do not use the square root of squared residuals $\left(E_i - \hat{E}_i\right)$, which tend to over-emphasize the contribution of large errors between predictions and measured values. Instead, an absolute value of the error is used since we are interested in the magnitude of the random error and not in its direction (positive or negative). One such metric is the mean absolute percentage error, MAPE.

[3] One such source is: https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/#

Eqn. 8: $$\boldsymbol{MAPE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left|\frac{\boldsymbol{E_i}-\widehat{\boldsymbol{E}}_{\boldsymbol{i}}}{\boldsymbol{E_i}}\right| \times \mathbf{100}$$

However, this definition suffers when actual values of energy use are zero or near zero. To account for low values, a modified version is used, called the weighted MAPE, wMAPE.

Eqn. 9: $$\boldsymbol{wMAPE} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}|\boldsymbol{E_i}-\widehat{\boldsymbol{E}}_{\boldsymbol{i}}|}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}|\boldsymbol{E_i}|} \times \mathbf{100}$$

Uncertainty

Savings can never be directly measured, it can only be estimated from the difference between the baseline model predictions and measured energy values. We determine how well we know the savings by estimating the uncertainty of our calculations. The uncertainty is a probabilistic statement of the confidence we have that the actual amount of savings lies within a specified interval. Proper statements of uncertainty require a precision and confidence level.

ASHRAE Guideline 14-2014 provides a relatively simple formula for estimating savings uncertainty that relates the CV(RMSE) and the amount of savings expected for the project, referred to as the Fractional Savings Uncertainty (FSU). For a selected confidence level, this formula states how accurately the savings can be calculated for a given model based on the expected amount of savings. CPUC uses a 90% confidence level by convention, while ASHRAE Guideline 14-2014 requires a 68% confidence level for compliance. The uncertainty must be less than 50%, as savings is stated as the value 'plus or minus' the uncertainty (since plus or minus 50% leads to a 100% band). ASHRAE provides different versions of the FSU formula based on whether the model residuals have autocorrelation. Models based on shorter time periods such as daily or hourly have higher degrees of autocorrelation. For cases with autocorrelation, the ASHRAE formula is provided below.

Eqn. 10: $$\boldsymbol{FSU} = \frac{\boldsymbol{\Delta E_{save}}}{\boldsymbol{E_{save}}} = \frac{(\boldsymbol{aM^2+bM+C})\boldsymbol{t}}{\boldsymbol{m\bar{E}_{base,n}F}}\left[\boldsymbol{MSE'}\left(\mathbf{1}+\frac{\mathbf{2}}{\boldsymbol{n'}}\right)\boldsymbol{m}\right]^{\mathbf{0.5}}$$ *, where* $\boldsymbol{\bar{E}_{base,n}}$ *is the mean baseline use,* $\boldsymbol{F}$ *is the fraction of savings from baseline use, M is the number of months in the post period, and a = -0.00024, b = 0.03535, and c = 1.00286 for daily models.*

Eqn. 11: $$\boldsymbol{MSE'} = \frac{\mathbf{1}}{\boldsymbol{n'-p}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(\boldsymbol{E_i}-\widehat{\boldsymbol{E}}_{\boldsymbol{i}}\right)^{\mathbf{2}}$$

Note that CV(RMSE) in this equation has been modified to use the apparent number of independent data points n', which is less than the total number of baseline period points n.

Eqn. 12: $\boldsymbol{n'} = \boldsymbol{n}\frac{(\mathbf{1}-\boldsymbol{\rho})}{(\mathbf{1}+\boldsymbol{\rho})}$*, where* $\boldsymbol{\rho}$ *is the correlation coefficient between residuals offset by one time step (referred to as lag 1 autocorrelation).*

The inclusion of the polynomial $(aM^2 + bM + c)$ was an improvement by Sun and Balthazar (2013) to the original ASHRAE Guideline 14-2014 version provided by Claridge and Reddy (2000).

Accounting for autocorrelation in energy use data is a complicated issue. For example, the data could be related to other data points besides its immediate successor or predecessor in time. In daily models, energy use on Mondays may be related to energy use on previous or following

Mondays in buildings, as Monday natural gas use may be higher than other weekdays due to building warm up after a weekend shutdown. Autocorrelation complicates methods to estimate savings uncertainty. Not accounting for autocorrelation can greatly underestimate the savings uncertainty.

A study was performed by LBNL to test the reliability of different uncertainty methods to estimate prediction uncertainties (Touzani, et. al. 2019). While not 100% reliable in all cases, ASHRAE's FSU for daily linear models was found to provide reliable uncertainty estimations in over 75% of cases, using TOWT models the number dropped to approximately 60%.

Koran (2017) examined four methods for estimating uncertainty using four different data sets. The methods included ASHRAE's FSU, the improvement to the FSU equation described above, an algebraic solution for aggregated uncertainty from OLS methods, and bootstrap resampling methods. Three of the data sets consisted of synthetic data with increasing levels of autocorrelation; the fourth data set was data from a real building. He tested the synthetic data sets with linear relationships and the real data set with a four-parameter change point model (Kissock, 2004). Several interesting findings were made:

a) The improved FSU, aggregated OLS uncertainty, and bootstrapping methods produced almost identical results for datasets without autocorrelation,
b) The improved FSU method significantly overstated its impact on uncertainty in comparison with bootstrap methods,
c) The aggregated OLS and improved FSU methods overestimated the uncertainty, and
d) All of the approaches provided reasonable results, with no approaches differing by orders of magnitude or even a factor of two.

Shonder and Im (Shonder and Im, 2012) find that assessment of savings accuracy are infrequently considered, in part because classical statistical methods are difficult to apply. Bayesian inference provides an alternate method to quantify savings and savings uncertainty in efficiency projects, by applying probability distributions to parameters used in the analysis and estimating the results with numerical techniques. A natural gas boiler replacement project is used to compare results using classical statistical methods with the Bayesian inference approach and shown to be the same. A second example is used to show the power of the Bayesian inference method when the data exhibit nonlinearity and serial autocorrelation, a situation in which there are no analytical solutions.

It would be interesting to carry out such a comparison of different uncertainty methods on a larger number of data sets, however this was beyond the scope of the current effort.

Modeling Algorithms

Energy-use prediction is traditionally carried out using a variety of linear models, including change point (or piecewise linear), time-of-week and temperature, and degree-day models. These models are based on ordinary least squares (OLS) regression, allowing the use of the ASHRAE uncertainty calculations described above.

In modeling building energy consumption, the most common independent variable is the outdoor air temperature. Ambient weather conditions affect a building's HVAC systems, which are approximately 40% of a typical non-residential building's energy use. Change-point models are a

class of models that capture the trends of energy consumption over the range of ambient temperatures. These models range from two parameter heating or cooling to five-parameter heating and cooling models. The number of parameters refer to the number of coefficients that represent the model. Two parameter models capture a single linear relationship between outside air temperature (OAT) and whole-building energy consumption and are used for either heating or cooling scenarios. Three parameter models capture the linear relationship between energy use and OAT above or below a change point, and a constant relationship between OAT and energy use (i.e., energy-use is assumed constant) at other times. The change-point is said to have physical significance for the building, as it delineates the temperature below which heating is required and above which the required heating, if any, is insensitive to temperature.

Four parameter models are similar to three parameter models, but with two different linear relationships (slopes) on either side of the change-point. Five parameter models are for electric heated and cooled buildings (Kissock, 2004).

Time-of-week and temperature (TOWT) models (Mathieu, et. al. 2011) use time-of-week indicator variables and piecewise linear temperature dependence to capture energy dependence on both weekly operations and ambient temperature. As with change-point models, the temperature dependences are piecewise linear, however multiple segments may be used and their change-points may not always be physically significant. There are seven time of week indicator variables for models developed from daily data, as is common for natural gas.

Other strategies to consider are autoregressive integrated moving average (ARIMA) models and generalized linear models (GLM). While ARIMA models assume an underlying stationary process (constant mean and variance), processes with a cyclical component (such as seasonal variations) can be decomposed into multiple ARIMA process. ARIMA models have shown promise in predicting natural gas demand (Erdogdu, 2010) based on previous demand and prices. GLM are similar to classical linear regression models (Dunn & Smyth, 2018), but do not assume a linear relationship between the predictors and the response variable. Leading to greater flexibility in the processes modeled. Furthermore, GLMs do not require the errors of the response variable to be normally distributed. While the scope of this project did not allow investigation into these strategies, further research may be done on the ability of these types of models to compensate for seasonally low natural gas usage.

Modeling Strategies

Modeling strategies involve breaking up a building dataset and creating separate models for each portion of the data. Modeling strategies are often used when distinctly different operation modes are present, such as a school's in-session and vacation and summer periods. Low gas use periods of non-residential buildings may be separated and modeled separately, often using simple averages of the low use period. An appropriate modeling algorithm may be applied to the remaining high-use data. A key to making predictions using a baseline model that is made up of two or more sub-models is knowing when to apply each sub-model. An indicator variable that takes on a value of one or zero depending on the time of year or the temperature exceeding a certain value may be used to apply a sub-model in a grand equation that includes each sub-model. Such modeling strategies may be complicated to program and are often implemented manually in spreadsheets.

The overall model pieced together with sub-models must pass the goodness of fit criteria to participate in a site-level NMEC program, but how should the metrics be calculated? Two options may be used: 1) determine the goodness of fit metrics based on the final pieced-together model estimates and training period data, and 2) determine the goodness of fit metrics for each sub-model and combine them using an energy-weighted average (Eqn. 7). As described above, weighting the contribution to the overall model goodness of fit by the energy use it represents is the logical approach.

Predictive Accuracy

Evaluating a model over only the training period provides an incomplete picture about its predictive ability. A common approach to evaluating a model is to test its predictions on a similar but new dataset. In scenarios where at least 18 months of data is available, this approach might be useful in evaluating good models for sites with gas-use NMEC projects. Eighteen months of data generally includes a full range of gas use and temperature conditions. Due to the issue with the normalizing term in CV(RMSE) described above, the models may present with summary statistics that show poor model accuracy. However, if the predictive accuracy of these models can be evaluated and quantified, this approach might provide a pathway for allowing gas-use projects to participate in NMEC programs. The prediction intervals on the out-of-sample predictions may be used to quantify the uncertainty in savings calculated over the post-period data.[4]

Methodology

The literature review provided several alternate goodness of fit metrics, modeling methods, and modeling strategies worthy of investigating. To investigate them, a large number of individual building natural gas data sets were required. To assure the results were fairly representative of all potential natural gas site-level NMEC projects, several considerations were made in selecting the data. The main considerations were:

- The dataset should have a high percentage of buildings known to fail the current goodness of fit requirement, CV(RMSE) < 25%. The study seeks alternate methods to improve the rate of acceptance for these buildings.
- The dataset should also have a large number of buildings that pass the current goodness of fit requirement. Any recommendations for alternate methods must not reduce the rate of acceptance of projects that meet the current criteria.
- The dataset should include buildings from both warm and cool climate zones. Weather conditions heavily influence natural gas consumption in commercial buildings. The alternate methods should be generally applicable regardless of the climate zone where the building is located.
- The dataset should include many types of buildings in the commercial and public sectors from all over the state.
- Buildings should have a minimum 10,000 therms of annual natural gas use to assure the dataset includes buildings with a significant potential for savings in an NMEC program.
- Individual premise natural gas data was collected at daily time intervals over two years prior to the shutdown order in March 2020 due to the COVID-19 pandemic.

[4] See discussion in https://online.stat.psu.edu/stat508/lesson/2/2.

- Accuracy of NMEC savings analysis is improved through use of advanced modeling algorithms with shorter (than monthly) time interval data.
- Two years of data enable quantification of model prediction errors which are useful in the evaluation.
- The dataset should not be complicated by the known low gas use impacts that occurred due to the COVID-19 pandemic.

A deeper discussion is provided here to address the first few items in the above list.

Building types that typically fail the existing model goodness of fit criteria are those that have low and no gas usage over significant portions of the year. This is the case in buildings where there are few natural gas end uses, such as only for space and domestic water heating. These cases include most commercial building types such as commercial office, educational (K-12, college and university classroom or administrative), and public sector buildings (state and local government office, city halls, and community centers). These are referred to as 'Affected' building types. Other building types including laboratories, hospitals, and medical offices have additional end-uses for natural gas, and do not exhibit long periods of low use during the year. We refer to these as 'Unaffected' building types. Residential buildings were excluded from the study since these buildings typically consume gas for other end-uses, such as cooking and gas dryers, which do not exhibit seasonal dependance, and are not good candidates for site-level NMEC projects.

Because of the above considerations, it was assumed that Affected buildings could be represented by a few distinct building types. The IOUs record North American Industry Classification System (NAICS) code[5] in their customer databases. Several NAICS codes were identified for the Affected building types, including:

- 551114: Corporate Offices
- 921190: Personnel Offices, Government
- 92XXXX: Public Administration
- 61111X: Elementary and Secondary Schools (elementary, charter, high school, etc.)
- 6112XX: Junior Colleges
- 6113XX: Colleges, Universities, Professional Schools
- 6114XX: Business Schools
- 624120: Community Centers

NAICS codes for Unaffected building types included:

- 622110: Hospitals
- 621512: Medical radiological laboratories
- 621511: Laboratories, medical
- 621511: Forensic laboratories
- 62151X: Medical and diagnostic laboratories

Thirty Affected and Unaffected natural gas building data sets were requested from each utilities' climate zone. The request was made based on the Affected and Unaffected categories, not based on particular NAICS codes. To maintain confidentiality, customer ID numbers were randomized

[5] NAICS codes may be found at https://www.census.gov/naics/.

but unique so that each building data set could be identified but not have any relationship with the actual customer. In some cases, the utilities reported back that thirty buildings were not available for every climate zone, particularly in the less populated inland areas of California.

A data collection plan was developed, and individual data requests were provided to each participating utility as guidance for collecting and providing data. All data was provided under non-disclosure agreements.

Once data was received from each utility, it was reviewed and prepared for analysis. A majority of the data sets were of high quality, included the requested two years, and were in daily time intervals. Some data quality issues were encountered. The issues and their resolutions included:

- Issue: Data with irregularly spaced timestamps, not uniform daily intervals. Irregular time intervals that spanned multiple days. These were from meters with low communication rates over multiple days. Resolution: This data was not used.
- Issue: Data sets with two or more identifying numbers. Resolution: These were identified and resolved with the source utility.
- Issue: Dropped data resulting from meter reading issues. Resolution: If the dropped data was a significant portion of the data, the entire dataset was discarded. This resulted in less than 2% of the data sets being discarded.

After data preparation, a total of 635 building data sets were available for analysis, where a building data set consists of two years of data from a single building

Using the zip codes, the climate zone was identified for each building. Ambient temperature data in hourly time intervals was downloaded from NOAA Integrated Surface Database (ISD) for each climate zone for the two year duration of the study. kW Engineering's open-source R package, nmecr (kW Engineering, 2019), was used to develop models based on the first year (training period) of each building.

The nmecr R code was developed for site-level NMEC projects. Among other features, it allows the user to select a modeling algorithm and an analysis time interval (hourly or daily) when developing energy models. Multiple modeling algorithms are included in the nmecr package: a version of LBNL's time-of-week and temperature (TOWT) model, the family of ASHRAE's piecewise linear change-point models, and heating- and cooling-degree day models. All modeling algorithms are based on ordinary least squares regression. The nmecr code calculates each of the required goodness of fit metrics CV(RMSE), NMBE, and R^2. It also calculates ASHRAEs FSU based on a user-input amount of savings, or a default 10% value.

An analysis platform was set up to intake individual natural gas and ambient temperature data sets, run the selected modeling algorithm, and output the goodness of fit metrics. TOWT was chosen to model each data set, because it accounts for both ambient temperature and weekly operation effects on energy use. Using the training period energy use and the model estimates, the alternate goodness of fit metrics were calculated in the platform. To test how accurately each model predicted energy use, second year temperatures were used in the training period models to predict second year energy use. We refer to the second year as the test period. The same metrics evaluated for the training period were also calculated for the test period. In addition, we calculated the percent difference between test period predictions with test period actual values.

The training and test period goodness of fit metrics, alternate metrics, FSU, and test period percent accuracy for each building data set were output to a summary spreadsheet. Included were each building's meta data, including building identifier, source utility, zip code, climate zone, and Affected or Unaffected status based on NAICS code and based on meeting the CV(RMSE) < 25% criterion, so that results could be filtered and analyzed under different conditions.

To test the relevance of different modeling algorithms, we re-ran the analysis on the same data using a three parameter heating (3PH) model. This created a new summary spreadsheet of results. To evaluate the relevance of a weighted CV(RMSE), the 3PH modeling platform was modified to determine the CV(RMSE) for the low and high use periods separately, then determine their weighted average.

Exploration of alternative modeling strategies were performed manually. For ten sites that had not passed the CV(RMSE) criterion, the low use portion of the year was separated from the high use portion of the training period and separate models were created for each portion. The sites were selected from the data sets to include only those with explicit low use periods in the training period.

The summary spreadsheets were used to quantify the percentage of buildings that passed the current goodness of fit CV(RMSE) criterion as well as the FSU criterion from ASHRAE (< 50% at a 90% confidence level). Currently no references provide suggested acceptance levels for the other metrics calculated for each model: nRMSE, tRMSE, and wMAPE. In addition, examination of each metric's summary statistics and distributions provides no real insight for an appropriate acceptance criterion.

To develop further insight, a binary classification method was used.[6] Binary classification may be used to understand the outcome of a particular test in the context of a whether a given condition holds true or not. For example, a new diagnostic test for whether a patient has a disease or not may be evaluated using binary classification. The condition is whether a patient has the disease, and the new diagnostic test outcome can be compared against it. Four possibilities exist:

- True Positive (TP): the patient is diseased and the test predicts as diseased
- False Positive (FP): the patient is healthy but the test predicts as diseased
- True Negative (TN): the patient is healthy and the test predicts as healthy
- False Negative (FN): the patient is diseased but the test predicts as healthy

where TP, TN, FP, and FN are the number of true positive, true negative, false positive, and false negative cases for a set of results, respectively. Binary classification is useful for focusing on individual rates, such as true positive rate (TPR), which is a measure of how good a test is at identifying positive cases, and true negative rate (TNR), which describes how good the test is at identifying negative cases. The complements of these rates are false negative rate (FNR) and false positive rate (FPR), respectively. Because they are complements, one only needs to look at two of these four metrics to get a complete picture of how good a test is at classification.

[6] A good description of binary classification may be found here: https://en.wikipedia.org/wiki/Binary_classification

Eqn. 13: $\boldsymbol{TPR = \frac{TP}{TP+FN}}$ *Eqn. 14:* $\boldsymbol{FNR = \frac{TN}{TP+FN} = 1 - TPR}$

Eqn. 15: $\boldsymbol{FPR = \frac{FP}{TN+FP} = 1 - TNR}$ *Eqn. 16:* $\boldsymbol{TNR = \frac{TN}{TN+FP}}$

The alternate goodness of fit metrics were evaluated using this binary classification methodology. For our purposes we focused on classification performance with respect to TPR, and FPR. The given condition was whether the building data set was 'modelable.' We defined 'modelable' as cases when the model's prediction error was low – that is, the model accurately predicted the total energy use of the test period. This condition was set when the absolute prediction error was lower than a selected value, such as 2%. Modelable cases are those cases that we believe are a good fit for an NMEC project regardless of their CV(RMSE)

To investigate the alternate goodness of fit metrics, we varied their acceptance criteria over a wide range of values. Using plots that show how the true positive rates and false positive rates change as the metric's acceptance criterion was varied provided insight on the worthiness of each metric as well as an indication of an appropriate value of its acceptance criterion. This also enabled insightful comparisons of the binary classification outcomes between different modeling algorithms. The results of a binary classification experiment as applied to this use-case can be represented with the contingency table shown in Table 1.

Table 1. Binary Classification Contingency Table

		Classification when using a particular criterion	
		Test Outcome *Positive*	**Test Outcome *Negative***
Ground truth classification, based on NMBE	**Condition *Positive* (NMBE ≤ 0.02)**	True *Positive*	False *Negative*
	Condition *Negative* (NMBE > 0.02)	False *Positive*	True *Negative*

Results and Discussion

NAICs Code as Indicator of Affected Buildings

The NAICs codes used to select buildings for this study turned out to be a fairly good indicator of buildings that have low natural gas use during the warm summer months. Categorizing by NAICs codes, about 55% of the building data set collected were considered Affected buildings (had low energy use in warm months). These NAICs codes were:

- 551114: Corporate Offices
- 921190: Personnel Offices, Government
- 92XXXX: Public Administration
- 61111X: Elementary and Secondary Schools (elementary, charter, high school, etc.)
- 6112XX: Junior Colleges
- 6113XX: Colleges, Universities, Professional Schools

- 6114XX: Business Schools
- 624120: Community Centers

After modeling the first year of natural gas use and checking the CV(RMSE), we found that 50% of the data sets did not pass the CV(RMSE) criterion. The percentage of Affected NAICs code cases that actually did not pass the CV(RMSE) criterion was 91%. Further, 85% of the total data set had agreement between the building status as determined by the NAICs code and the status as determined by modeling the building (with the TOWT algorithm). This showed that the NAICs code provided a good indication of a building that may not pass the Rulebook goodness of fit criterion. Additional considerations as outlined here should be considered in order to accept the building as a natural gas NMEC project.

Data Quality

The observed quality of the natural gas data was poor in comparison with electricity use data sets. Gas data tends to show more erratic usage patterns than electricity. Figure 3 shows two examples of issues with gas use in commercial buildings. As demonstrated by the left chart in Figure 3, there can be excessive periods of zero values in the data set. This may indicate data recording or collection errors. As the chart on the right of Figure 3 shows, step-wise shifts in gas use are often present, which may be caused by non-routine events in the building, or reflect manual operations of gas equipment. The chart also shows periods where repeated values are present, which may be the result of poor resolution of the gas meter or problems with gas data telemetry. Utility natural gas meters generally measure gas flow, and the meters used are not as accurate as their electric meter counterparts. This work did not address treatment of individual building non-routine events, however it should be recognized that non-routine events can often be accounted for with different modeling approaches and strategies and qualify as a site-level NMEC project in most programs. This report does not address methods for treating individual project non-routine events.

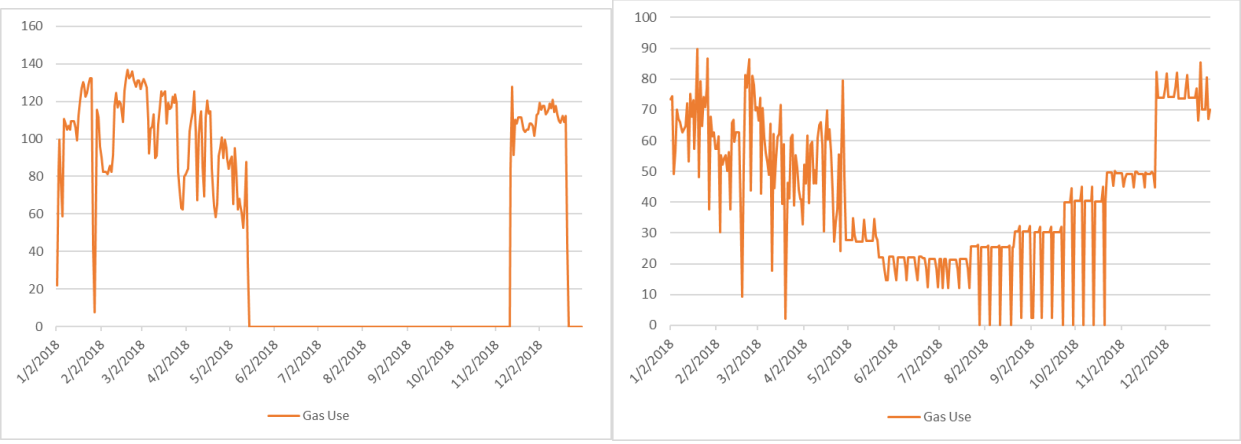


Figure 3. Examples of Poor Quality Gas Data Sets.

Binary Classification

To use the binary classification strategy outlined above, buildings were considered part of the positive class if they were a good candidate for nmec (modelable), and negative class if they were

not a good candidate (unmodelable). Ground truth of the class was determined by the test period absolute value of NMBE (Eqn. 2), with a 2% limit between the model's prediction and actual total year two gas use. Thus, buildings that had a test period absolute value of NMBE less than or equal to 0.02 were considered part of the positive class. Of the 635 buildings considered, 95 (15%) qualified as modelable using this criterion. A true positive case was when the selection criterion agreed with a modelable building case. A false postive case was when the selection criterion indicated an acceptable model however the building was not considered modelable. Figure 4 provides an example of a true positive case and Figure 5 provides an example of a false positive case.

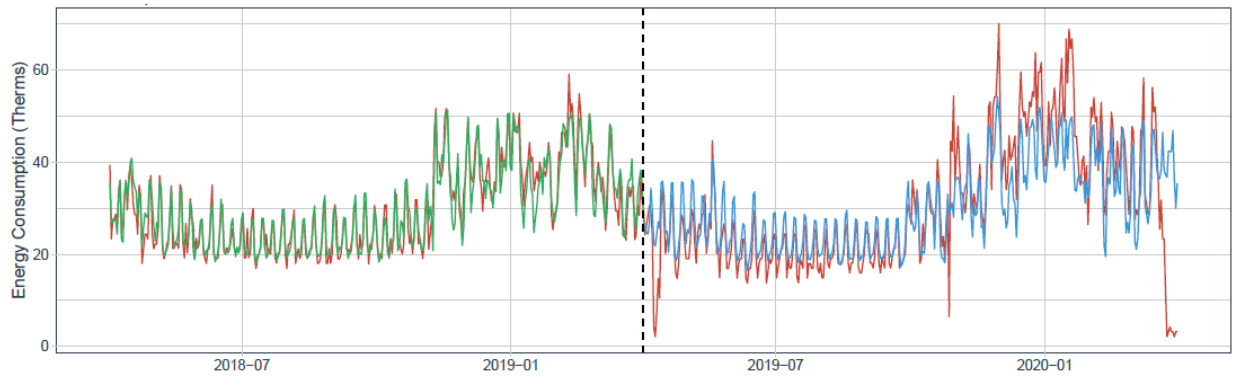


Figure 4. True positive case: |Test NMBE| = 0.4%, CV(RMSE) = 11%.

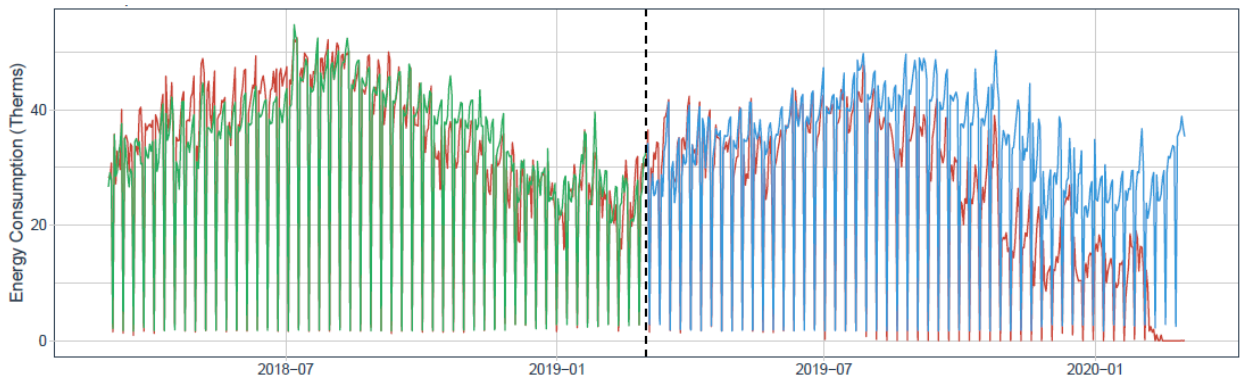


Figure 5. False positive case: |Test NMBE| = 33%, CV(RMSE) = 12.5%.

Following are the results of the evaluation of the existing and alternate goodness of fit metrics for two modeling algorithm cases: TOWT and a three-parameter change-point model (3PH).

Modeling with TOWT

We developed TOWT models for the first year of gas data for 635 building data sets, then compiled the various metrics into a summary spreadsheet. As described above, 50% of those buildings did not pass the CV(RMSE) criterion (must be lower than 25%). Of these Affected buildings, we calculated the FSU assuming 10% savings and found that 13% of these Affected buildings passed the FSU < 50% criterion. This result supports the FSU metric as superior to CV(RMSE) for selecting models. FSU considers model variation in context with the amount of savings. Because FSU varies inversely with the amount of savings, higher estimated savings will yield lower values of FSU.

Considering that natural gas has relatively fewer end uses in buildings as compared to electricity, it is mainly used for space and water heating, efficiency applications in natural gas end uses can often yield savings over 10% of annual use.

When using the binary classification method to evaluate all datasets (Affected and Unaffected), Using CV(RMSE) as the selection criteria for NMEC led to a TPR of 74%, a FPR of 47% at the current cut-off threshold of 25% (shown in Figure 6, with the 25% cut-off threshold marked by a vertical line). This chart shows how the TPR and FPR change as the cut-off threshold (shown on the x-axis) is increased. Note that the TPR increases faster than the FPR initially, then maintains a relatively constant 28% separation between the 15% and 25% cut-off threshold, then shows how increasing the cut-off threshold beyond 25% leads to relatively smaller increases in the TPR, while the FPR increases faster. This supports the selection of the 25% CV(RMSE) criterion currently required for site-level NMEC projects.

The number of false positive cases (unmodelable buildings that pass the goodness of fit criterion) seems high at 47%. A metric that significantly reduces the number of false positive cases while maintaining high numbers of true positive cases would be preferred.

Figure 6. True positive and false positive rates for a TOWT model when using CV(RMSE) as a criterion to qualify a data set for use in an NMEC project.

Using FSU instead of CV(RMSE) led to a TPR of 70%, a FPR of 43% at the ASHRAE Guideline 14 50% criterion. (See Figure 7, with the 50% cut-off threshold marked). As with the trends with the cut-off threshold in the CV(RMSE) chart above, this chart shows that the FSU TPR increases faster than the FPR initially as the cut-off threshold increases, but slows considerably at higher thresholds. At the 50% threshold, the TPR and FPR rates are similar but slightly lower than those for CV(RMSE).

Figure 7: True and false positive rates on a TOWT model when using FSU as a criterion.

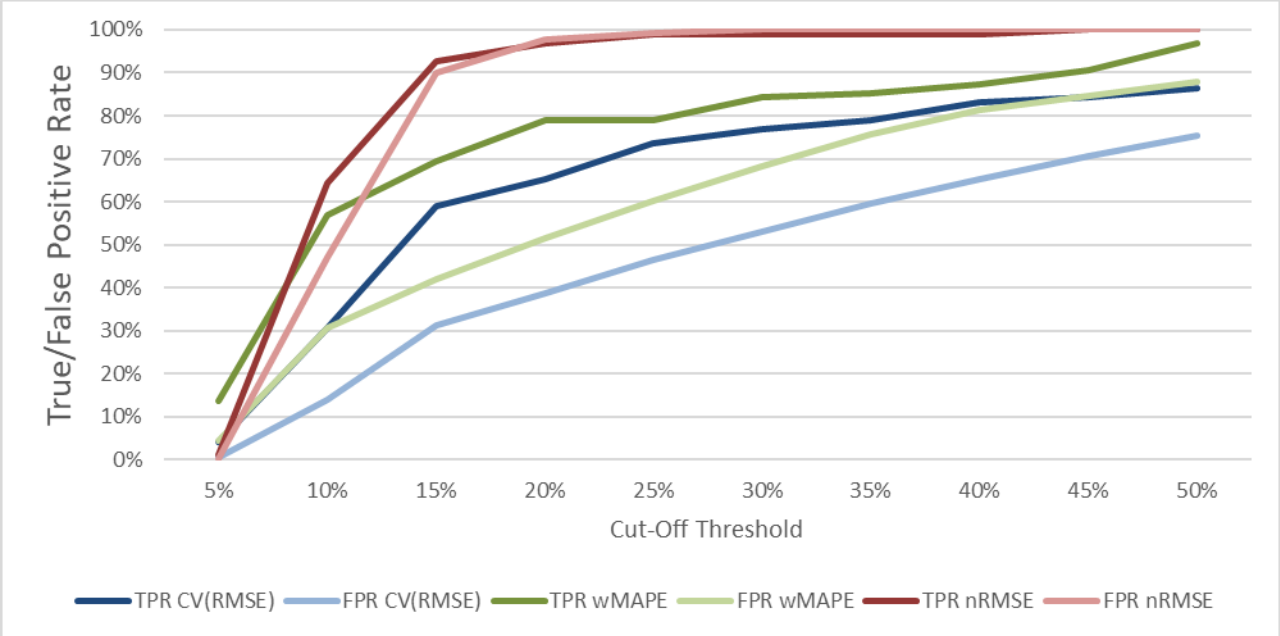


Figure 8: TPR and FPR for alternate metrics wMAPE and nRMSE as compared to CV(RMSE).

The binary classification results for the alternate metrics are shown in Figure 8. The nRMSE TPR and FPR have similar behavior over the entire range of cut-off thresholds, with little separation between them, showing little effectiveness in this metric's ability to identify modelable and unmodelable buildings. The results for wMAPE show higher TPR than that for CV(RMSE), but also significantly higher FPR from the 20% threshold and above, which is too high. The binary classification analysis for tRMSE yielded threshold values two orders of magnitude smaller than shown for CV(RMSE). Further analysis was discontinued for tRMSE as it was not considered to be informative.

Modeling with 3PH

The TOWT model is generally more accurate than simpler model forms for commercial building gas use because the model uses the time of week as an additional regressor along with ambient temperature. To understand whether the model algorithm choice had any significant differences from those described above, we re-ran the analyses using a three-parameter change-point (3PH) model (Kissock, et. al., 2004). This is a piecewise linear model with only ambient temperature (or other weather variable) as the independent variable. This model form was selected as it was considered appropriate for buildings with low gas use in warm months. As shown on the left side of Figure 9, energy use decreases as temperatures increase from low values until heating is no longer necessary, at which time natural gas use flattens. The 3PH model has a change-point between the sloped and flat portions of the linear segments. The charts also show how a TOWT model predicts this building's gas use. It shows that it can predict negative values of gas use during these low use periods.

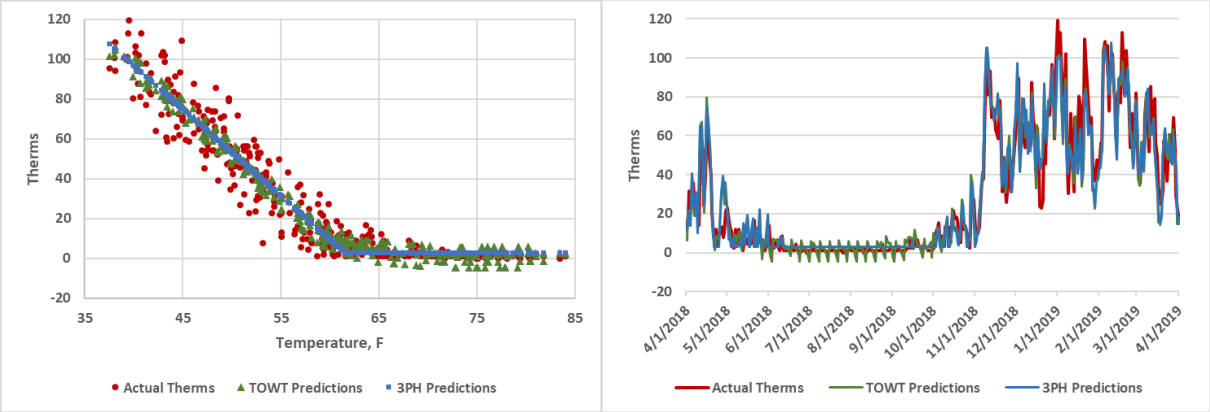


Figure 9: Comparison of 3PH and TOWT models, left: scatter plot with temperature, right: time series plot.

The 3PH models were analyzed in the same way as the TOWT models and the binary classification scheme was used to compare the results. In this case, CV(RMSE) led to a TPR of 65% and a FPR of 41% at the 25% threshold, which was a poorer outcome for the TPR and a somewhat better outcome for the FPR in comparison with the TOWT results.

Figure 10: True and false positive rates for a 3PH model when using CV(RMSE) as a criterion.

Similarly to the TOWT results, the FSU analysis led to comparable though slightly poorer outcomes when compared to the CV(RMSE) results, with a TPR of 62%, and an FPR of 40% at the 50% threshold.

In a similar analysis on Affected buildings (not using the binary classification method), 45% of the buildings failed the CV(RMSE) criterion when the buildings were modeled with the 3PH model. Using FSU on the 3PH models that failed the CV(RMSE) criterion, 10% of the Affected buildings then passed.

Figure 11: True and false positive rates for a 3PH model when using FSU as a criterion.

Weighted CV(RMSE)

Using the 3PH model we developed a weighted CV(RMSE) by developing CV(RMSE) for the high and low gas use period separately, then combining them using each periods gas usage as a weighting factor. This strategy reduced the contribution of the CV(RMSE) from the low use period in the final weighted CV(RMSE). Using the binary classification analysis, the results are shown in Figure 12. In comparison with the unweighted CV(RMSE), the TPR and FPR were practically the same at 65% and 41% respectively. A summary of the binary classification results is shown in Table 2.

Figure 12. True and false positive rates for a 3PH model when using wCV(RMSE) as a criterion.

Table 2: Summary of Binary Analysis Results with Two Models

	TOWT Model		3PH Model		
	CV(RMSE) at 25% Threshold	FSU at 50% Threshold	CV(RMSE) at 25% Threshold	FSU at 50% Threshold	wCV(RMSE) at 25% Threshold
True Positive Rate (TPR)	74%	70%	65%	62%	65%
False Positive Rate (FPR)	46%	42%	41%	40%	41%

Climate Zone Effects

We analyzed the data to understand whether there were any weather-related effects. Natural gas use in commercial buildings is mainly for heating, and California has sixteen defined climate zones throughout the state. Colder climate zones will have longer periods of gas use throughout the year

while milder climate zones will have longer periods on little or no gas use. We analyzed the data by cold and mild climate to determine if there were any weather effects.

To sort the data into cold and mild climate zones, we collected heating and cooling degree day (HDD and CDD) data for each climate zone from a published source (PG&E, 2006). This report provided the HDD and CDD data for four locations in each of California's 16 climate zones. The mean HDD and CDD were recorded for each zone and plotted in a bar chart, Figure 13. Based on the chart, cold climate zones were identified as those having over 2,000 HDD annually, while mild climate zones were those with under 2,000 HDD. Table 3 identified the cold and mild climate zones and Figure 14 shows their general location in California.

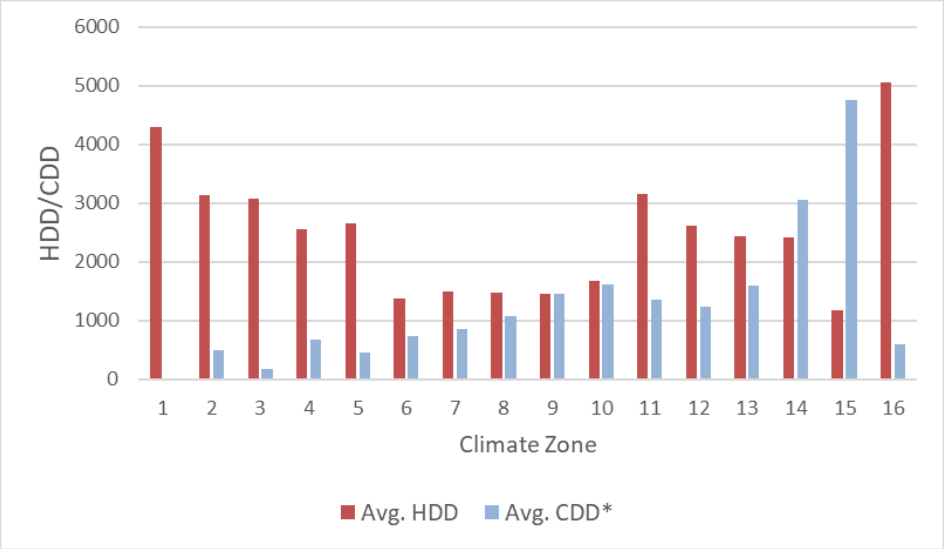
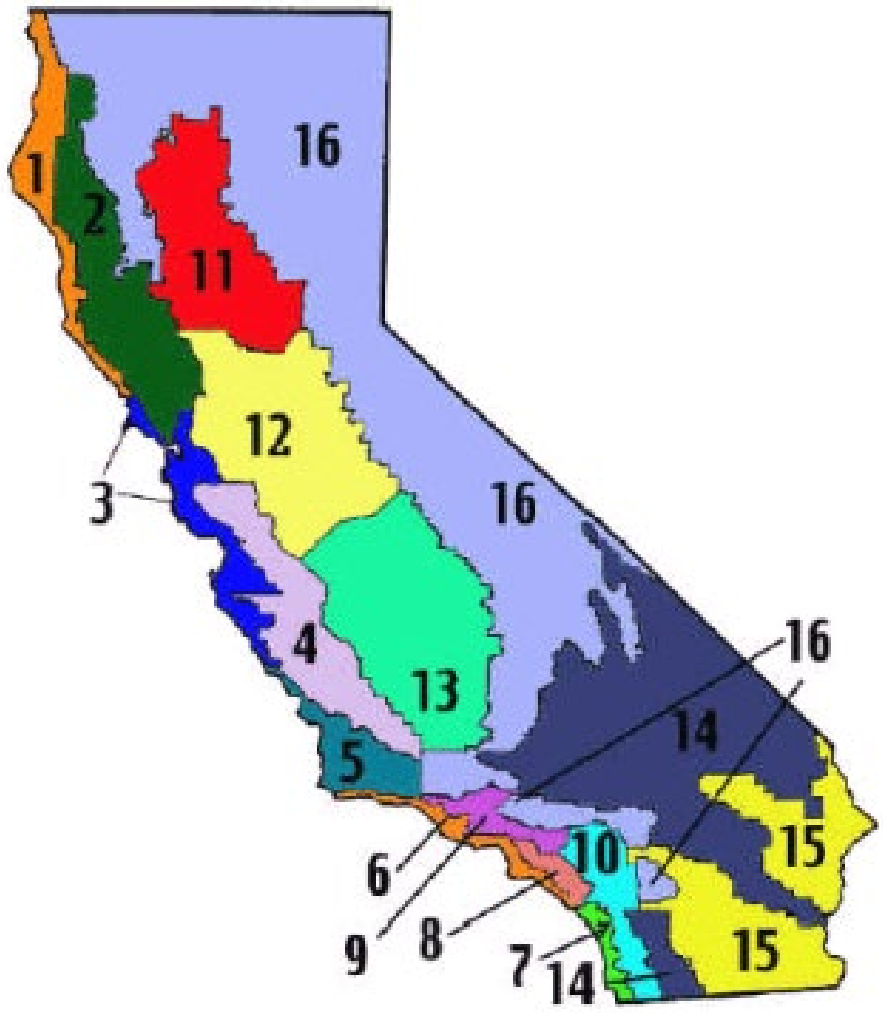


Figure 13. Average heating and cooling degree days for California's 16 climate zones.

Table 3. Climate zones classification.

Classification	Climate Zones
Cold (> 2,000 HDD)	6, 7, 8, 9, 10, 15
Mild (< 2,000 HDD)	1, 2, 3, 4, 5, 11, 12, 13, 14, 16

Figure 14. California climate zone map.[7]



Of the 635 data sets, 70% of them were in cold climates. We separated the cold and mild building data sets into separate spreadsheet tabs to complete similar analysis comparing CV(RMSE) and FSU as described above. The results are compiled in Table 4.

Table 4. Results by climate zone type.

Climate Zone Type	# Buildings	% Affected Buildings	% Additional Buildings
Cold	445	52.4%	13.4%
Mild	189	45.2%	12.7%

[7] Image taken from CEC Title 24 Residential Compliance Manual, https://www.title24express.com/what-is-title-24/title-24-california-climate-zones/.

Table 4 shows that there are more buildings that do not pass the CV(RMSE) criterion in colder California climates, but not by a dramatic amount over that in milder climates. When FSU is calculated for these Affected buildings, the rate of improvement is consistent between the two climate classifications at approximately 13%. This shows that there isn't a bias introduced by using FSU for cases when the building initially fails the CV(RMSE) criterion.

Modeling Strategy

A modeling strategy of manually separating the low use period gas data and developing separate models on each was performed on then randomly selected Affected sites, each that failed to pass the CV(RMSE) criterion. Separate TOWT models were developed for the high use and the low use periods, the CV(RMSE) for each model was determined and an energy-weighted CV(RMSE) was developed for the entire year. Results are shown in Table 5.

Table 5. Weighted CV(RMSE) Results for Separating Low Use Period Modeling Strategy

No.	Original Model CV(RMSE)	Overall Use (Therms)	High Use Period (Therms)	Low Use Period (Therms)	High Use CV(RMSE)	Low Use CV(RMSE)	wCV(RMSE)
1	31%	22,002	21,619	383	21%	83%	22%
2	26%	27,561	24,642	2,919	18%	49%	21%
3	48%	31,062	30,895	167	28%	177%	29%
4	30%	16,338	16,047	291	21%	67%	21%
5	79%	34,574	34,093	481	50%	91%	51%
6	27%	26,445	25,984	461	20%	61%	21%
7	49%	51,288	49,374	1,914	30%	42%	30%
8	48%	21,912	21,331	581	33%	59%	34%
9	26%	31,980	29,227	2,753	18%	18%	18%
10	45%	14,244	13,361	883	32%	63%	34%

The weighting strategy did reduce the CV(RMSE) in five cases (highlighted in green). Note that in these cases, the original model CV(RMSE) were not excessively higher than the threshold criteria of 25%. Also, none of the cases in which the high use period CV(RMSE) exceeded the threshold could be reduced low enough to pass. Of those that did pass, the low use period CV(RMSE)'s were often very high. Based on this limited sample, the weighted CV(RMSE) is an appropriate alternate method to use when low gas use periods are present.

Recommendations and Conclusions

In this work, we examined alternate goodness of fit criteria to overcome the cause of many natural gas models failing the established goodness of fit criterion of CV(RMSE) < 25%. After assembling a

dataset consisting of Affected and Unaffected buildings (buildings with gas use models not meeting and meeting the criterion, respectively), we ran two different modeling algorithms, calculated several alternate goodness of fit criteria, and compiled summary tables for additional evaluation of the alternate metrics.

Before proceeding, it is important to note that we did not examine individual building gas datasets to determine the presence of poor data or unidentified non-routine events as would be done on a case-by-case basis in preparation of participation in a site-level NMEC program. This would require additional information from each building to be collected, regimes of operations to be identified, and additional analysis to be performed. This work instead proceeded by modeling the data as is without additional insight into each building that otherwise may have improved individual building gas models. For every site-level NMEC project, we recommend this preliminary analysis of the data be pursued.

Gas usage data typically have more data quality issues than electricity use data. Gas usage is usually measured in units of volume, which requires flow measurements. Flow measurement is typically less accurate than electric measurement. Communicating the gas consumption data to a central repository may also have problems not evident with electric meters. Gas data is typically more 'noisy' in that it has more day-to-day random fluctuations, often has large periods of missing data, and may reflect the manual operation of gas-consuming equipment in commercial buildings. It is acknowledged that there are more data quality issues with natural gas to address. However in preparation for an NMEC project, these issues may yet be overcome to qualify for a natural gas NMEC program.

The goodness of fit metric CV(RMSE) quantifies the amount of random error between a model and the data the model is developed from. For commercial buildings that generally use natural gas for space and water heating only, it is a common reason natural gas models fail to become NMEC projects. A literature search suggested several alternate goodness of fit metrics, a weighted mean absolute percent error (wMAPE), a range-normalized root mean squared error (nRMSE), and a root mean squared error normalized by total energy use (tRMSE). The fractional savings uncertainty (FSU) developed by Claridge and Reddy (2000) and used in ASHRAE Guideline 14-2014 combined the CV(RMSE) and savings in a new metric that directly addresses the important question of how accurate will the resulting savings estimate be given the proposed model? ASHRAE's FSU was tested as an alternate qualifying metric. Another metric was to separate the low use and high use periods and model them separately, weighting the CV(RMSE) by that period's total gas consumption.

We tested two different modeling algorithms, the time-of week and temperature (TOWT) model (Mathieu, et.al. 2011) and ASHRAE's three-parameter change-point model (3PH) (Kissock et. al, 2004), and performed a manual modeling strategy by separating the low use from high use period data and modeling separately with the TOWT model.

After testing 635 building data sets with the TOWT algorithm, we found 50% of the buildings did not pass the current CV(RMSE) criterion. When using the FSU as a metric on these failed buildings, we found that 13% of them passed when assuming the project would yield 10% savings or more. When the same buildings were modeled with the 3PH model, 45% initially failed the CV(RMSE) criterion, but 10% of the failed buildings passed when using the FSU.

The binary classification analysis was used to evaluate the CV(RMSE), FSU, wMAPE, nRMSE, tRMSE, and 3PH model wCV(RMSE). This analysis confirmed that the CV(RMSE), FSU, and wCV(RMSE) metrics were superior to the wMAPE, nRMSE, and tRMSE metrics, as they showed high true positive rates along with reasonably low false positive rates. False positive rates were too high for wMAPE and nRMSE. The results were consistent whether using the TOWT or 3PH modeling algorithms (the wCV(RMSE) with was tested only for the 3PH model).

Manual separation of the low gas use and high gas periods, development of separate models for each period, then calculating the energy-weighted CV(RMSE) of the two models showed that when the CV(RMSE) of a model built on the full dataset did not excessively exceed the 25% criterion, this strategy could be used to qualify more gas NMEC projects.

Based on this work, it follows that alternate metrics and modeling strategies may be used to qualify natural gas site-level NME projects. Our recommendations are summarized below:

1. Examine the gas use data for data quality issues. Assure a full dataset is obtained for each building. Identify and resolve data quality issues such as outliers and extensive gaps in gas use throughout the baseline year. Determine whether a different modeling approach or whether different regimes of operation or non-routine events are present. Obtain information from building operators to substantiate modeling assumptions and strategies.

2. Use the current CV(RMSE) criterion of 25% to determine whether a natural gas NMEC project is acceptable. Should the model fail the CV(RMSE) test, calculate the FSU assuming 10% savings. If there is a savings estimate available, use it in the FSU equation instead. If the FSU is < 50% at a 90% confidence level, accept the building as an NMEC project. FSU may be used as a criterion as long as the model is an ordinary lease squares regression-based algorithm.

3. If the current gas model fails the CV(RMSE) criterion by a small amount such as 5%, consider separating the low gas use from the high gas use period and modeling each period separately. Calculate the energy-weighted average CV(RMSE) from the two models individual CV(RMSE). If the weighted average CV(RMSE) passes the criterion, accept the building as an NMEC project.

References

American Society of Heating Refrigeration and Air-conditioning Engineers (ASHRAE) Guideline 14-2014 Measurement of Energy and Demand Savings. Available from www.ashrae.org.

California Public Utilities Commission (CPUC), Rulebook for Programs and Projects Based on Normalized Metered Energy Consumption, version 2.0, January 7, 2020, available at https://www.cpuc.ca.gov/-/media/cpuc-website/files/legacyfiles/n/6442463694-nmec-rulebook2-0.pdf

Dunn, P.K. and G.K. Smyth, 2018, Generalized Linear Models with Examples in R, Springer, New York.

Erdogdu, E., 2010, Natural Gas Demand in Turkey, Applied Energy, 87 (1), p. 211-219.

International Performance Measurement and Verification Protocol (IPMVP), 2016, Efficiency Valuation Organization, www.evo-world.org.

Kissock, K. K., J. S. Haberl, D. E. Claridge, 2004, "ASHRAE Research Project 1050: Development of a Toolkit for Calculating Linear, Change-Point Linear, and Multiple-Linear Inverse Building Energy Analysis Models," available from: www.ashrae.org.

Koran, B., E. Boyer, Khawaja, M.S., Rushton, J, and J. Stewart, 2017, A Comparison of Approaches to Estimating the Time-Aggregated Uncertainty of Savings Estimated from Meter Data, proceedings of the 2017 International Energy Program Evaluation Conference, Baltimore, MD.

kW Engineering, 2019, nmecr - An implementation of peer-reviewed energy data analysis algorithms for site-specific M&V, https://github.com/kW-Labs/nmecr.

Lawrence Berkeley National Laboratory (LBNL), 2019, "Site-Level NMEC Technical Guidance: Program M&V Plans Utilizing Normalized Metered Energy Consumption Savings Estimation, Version 2.0, December 15, 2019," available at: https://www.cpuc.ca.gov/-/media/cpuc-website/files/legacyfiles/l/6442463695-lbnl-nmec-techguidance-01072020.pdf.

Mathieu, J. L., P. N. Price, S. Kiliccote, and M. A. Piette, 2011, "Quantifying Changes in Building Electricity Use, with Applications to Demand Response," IEEE Transactions on Smart Grid, 1949-3053.

Pacific Gas & Electric (PG&E) 2006, Pacific Energy Center's Guide to: California Climate Zones and Bio Climatic Design, October 2006. Available at: https://www.pge.com/includes/docs/pdfs/about/edusafety/training/pec/toolbox/arch/climate/california_climate_zones_01-16.pdf

Reddy, T. A. and Claridge, D. E., 2000. "Uncertainty of 'Measured' Energy Savings from Statistical Baseline Models." HVAC&R Research, January 2000.

Shonder, J. A. and P. Im, 2012, Bayesian Analysis of Savings from Retrofit Projects, ASHRAE Transactions, Volume 118, Part 2.

Sun, Y. and Baltazar, J.C. 2013. "(DE-13-C033) Analysis and Improvement on the Estimation of Building Energy Savings Uncertainty," Proceedings of the 2013 ASHRAE Annual Conference, Denver, CO.

Touzani, S., J. Granderson, D. Jump, D. Rebello, June 2019, “Evaluation of Methods to Assess the Uncertainty in Estimated Energy Savings,” Energy and Buildings, 193(1):216-225

Appendix: Comments and Responses from Concerned Parties

Following are comments and responses received from CPUC Energy Division

Comment 1: Am I reading the study correctly, that the alternate method proposed only results in a 13% pass rate among projects that fail the CV(RMSE) criteria?

Response 1: Yes, 13% of the projects that failed the CV(RMSE) criterion passed when estimating the uncertainty (ASHRAE's FSU) and assuming 10% savings. This seems very incremental, however it reinforces that projects with higher savings are more likely to pass if FSU (or other uncertainty method) are used instead of considering only model goodness-of-fit metrics such as CV(RMSE). Higher savings = better NMEC projects.

Comment 2: Do you have suggestions for other possible approaches that may necessitate further research/development but could be used for even more projects without increasing risk?

Response 2: Yes, some suggestions are in the recommendations section: prepare the data – gas use data is messier so eliminate gaps and outliers without exceeding a 25% data removal limit, try different modeling algorithms (see comment received from PDA below), use FSU for daily models when CV(RMSE) fails, use a modeling strategy of separating the data into low and high use periods and modeling each period separately, then use a weighted average CV(RMSE) from each period.

On the research front, the LBNL study was cited that showed that FSU determinations of uncertainty for hourly models was unacceptable, and better but not fully acceptable for daily models. This is because the energy use data is serially correlated or has autocorrelation. FSU makes an first-order attempt to account for autocorrelation (considers lag 1 autocorrelation only), however the uncertainty methods in general should be improved to better account for it. A diverse group of experts and practitioners on an EVO/IPMVP technical subcommittee (Statistics and Uncertainty Application Guide) are providing more general methodologies for estimating uncertainty in Option C Whole building M&V calculations. These include: improved FSU methods, exact matrix solutions (FSU is a subset of these), and a bayesian approach. These are higher-order statistical solutions that are not a common practice in our industry, but there are tools and software that make their use easier.

Comment 3: Would the results meaningfully change/improve if the % savings was increased beyond the 10% threshold? What about possibilities for fuel substitution (when possible)?

Response 3: Yes. Higher savings always reduces uncertainty as a general rule. Natural gas NMEC projects should not necessarily be required to achieve more than 10% savings, but it should be stressed that higher savings projects are preferred for NMEC. This applies to electric savings as well.

Fuel substitution NMEC projects, if allowed, would include the 'apparent' gas savings as an adder to any other gas EE savings and help qualify gas NMEC projects using the FSU criterion. The downside would be quantifying the amount of gas substituted separately from other gas EE measures if incentives are allowed for EE savings only.

Following is the comment received during the public review period from the Public Documents Area on the CPUC's Evaluation Studies Public Documents website: (https://pda.energydataweb.com/#!/documents/2657/view).

Comment: This analysis is handicapped by sticking to the existing "published" methods. Current methods are inadequate and more modeling research is needed. Granderson et al note that most building models only work for "well behaved' buildings. Many buildings are not well-behaved, so existing models don't work well. Hence the high CV(RMSE) values. HEA has analyzed over 25,000 homes over the course of 10 years and have seen significant improvements in gas and electric regression results through algorithmic advances in many different areas, including: * Balance point temperature selection, for degree day calculations, through slope analysis of average energy use vs outdoor temp, utilizing a LOESS interpolation. * Improved handling of building thermal inertia using rolling 5 day averages. * Use of degree hours to calculate degree days (see description here - https://github.com/energy-market-methods/caltrack/issues/120) * Removal of vacation/unoccupied days using K-means cluster analysis. * Identification of different heating modes (e.g. pool heating vs home heating). * Interpolation of integral therms, esp for improved accuracy during periods of low gas use. Each of these improvements have resulted in more accurate models for actual homes. We are happy to share any of these methods.

Response: This comment makes good points about improved modeling methods (which we call algorithms). In this report we did compare two modeling algorithms: TOWT and a 3PH algorithm. We acknowledge that other modeling algorithms are worthwhile pursuits, however it's unclear whether there is an algorithm that addresses the low gas use periods barrier that this report addresses. Many of the examples cited are for houses or homes, not commercial buildings where the thermal characteristics are different. The comment mentions HDD models with balance point temperatures, which are analogous to the 3PH model we used. We agree with the point that more rigorous modeling algorithms would be useful, the TOWT and 3PH algorithms used in this report have significant differences in their accuracy as the TOWT algorithm better captures the weekly energy use whereas the 3PH only considers temperature effects. The report was updated to suggest that alternate modeling algorithms should be attempted should initial modeling attempts fail.