Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

PACIFIC GAS AND ELECTRIC COMPANY

POPULATION NMEC CONTROL GROUP ACCURACY ASSESSMENT

Prepared for:
Pacific Gas and Electric Company

By:
Demand Side Analytics

FINAL REPORT

ACKNOWLEDGEMENTS

Research Team

- Adriana Ciccone, M.S.
- Josh Bode, M.P.P.
- Stephanie Bieler, M.S.
- Andrea Hylant

Pacific Gas & Electric Team

- Brian Arthur Smith
- Lauren Garcia

ABSTRACT

This study sought to determine whether alternative Normalized Metered Energy Consumption (NMEC) models that included comparison groups comprised of non-participants were able to accurately account for changes in energy consumption associated with the COVID-19 pandemic. This research is relevant for energy efficiency program impacts as current NMEC methods in California do not require a comparison group for estimating energy efficiency impacts and can misattribute the effects of the pandemic to the effects of energy efficiency programs. The accuracy assessment was conducted as a tournament, whereby methods were identified and judged using pseudo-participants both prior to and during the pandemic. Results varied by segment, customer sector, and fuel type, but generally found that the inclusion of comparison groups improved performance. The recommendations from this analysis include producing a transparent repository for qualified models for each future program savings estimate rather than requiring a specific model be used.

COMPLETION DATE: FEBRUARY 15, 2022

CONTENTS

1 EXECUTIVE SUMMARY

Accurate and unbiased estimates of energy efficiency (EE) impacts are critical for utility program staff, third-party program implementers, and regulators. Pacific Gas and Electric Company (PG&E) currently uses the CalTRACK Version 2.0 method (CalTRACK) to estimate avoided energy use for its energy efficiency (EE) programs based on the Population-Level NMEC (normalized metered energy consumption) methodology, including the Residential Pay-for-Performance (Res P4P) and On-Bill Financing Alternative Pathway (OBF-AP) programs. The Population NMEC method relies on whole-building granular electric and/or gas consumption data to estimate the savings associated with the installation of an individual or multiple energy efficiency measures (EEMs) at the site that is paired with typical and historical weather data to derive normalized energy savings.

A notable feature of the current population NMEC method, CalTRACK v2.0[1], is the lack of comparison groups to adjust the energy savings baseline and normalize the savings estimate for factors beyond weather. This method relies almost exclusively on weather normalization and effectively assumes that the only difference between the pre- and post-intervention periods is weather and the installation of EEMs. The COVID-19 pandemic laid bare the limitations of the adopted method. The pandemic led to changes in our commutes, business operations, and home use patterns. Not surprisingly, it has also changed how, when, and how much electricity and gas we use. Moreover, the impact on energy use differs for residential customers and various types of businesses.

Given the changes in energy consumption that have occurred over the course of the COVID-19 pandemic, the need for alternative approaches to CalTRACK and similar, simple pre-post regression methods for estimating EE impacts is paramount. While adding comparison groups typically improves the accuracy of these energy saving estimates, there are three, main logistical challenges:

- **Privacy of non-participant customer data.** Current California laws and regulation exist to protect the privacy of AMI/smart meter data for individual customers.[2]
- **Transparency Challenges.** Many evaluation methods that rely on a comparison group require extensive calculation in order to construct the group. This complexity can hinder independent review and/or replication of the findings.
- **Complexity and frequency.** PG&E and third-party EE program implementers target a wide range of customer segments and geographic areas, each of which require regular and specifically targeted non-participant data for evaluation. This is a proposition that adds complexity to existing program administration processes.

To determine if there are viable alternative models that can accommodate the effects of the COVID-19 pandemic or other wide-scale non-routine events, Demand Side Analytics was retained to conduct an accuracy assessment of the existing Population NMEC methods as well as a variety of other methods with and without comparison groups.

1 See http://docs.caltrack.org/en/latest/methods.html for more detail.

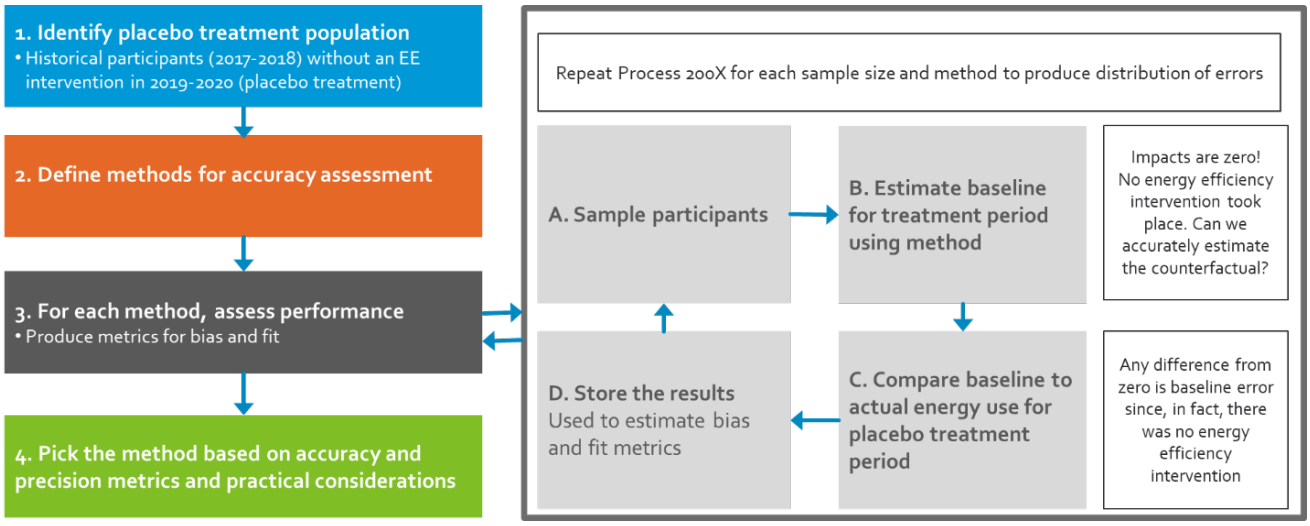
2 Including CPUC Decision D.97-10-031 (1997) Direct Access Proceeding (15/15 rule), Senate Bill No. 1476 (2010) Chapter 5 Privacy Protections for Energy Consumption Data, and CPUC Decision D.14-05-016 (2014) Decision Adopting Rules to Provide Access to Energy Usage and Usage-related Data While Protecting Privacy of Personal Data (2014)

Figure 1 summarizes the approach used to assess the accuracy and precision of alternative methods. The assessment is effectively a competition, where different methods are tested repeatedly, to identify the methods that are unbiased and accurate.

Accuracy is assessed by applying placebo treatment on customers that did not participate in EE programs during the period analyzed. The impact of a program (or in this case, a pseudo-program) is calculated by estimating a counterfactual and comparing it to the observed consumption during the post-treatment period.

A counterfactual is an estimate of what that participant would have done in the absence of the installation of EEMs. Because no EEMs were installed in this simulation, any deviation between the counterfactual and actual loads is due to error or other changes in consumption unrelated to the installation of EEMs. The process is repeated hundreds of times – a procedure known as bootstrapping – to construct the distribution of errors.

Figure 1: General Approach for Accuracy Assessment



Detailed findings for each customer segment and fuel type can be found in Section 4. The key findings, however, are as follows:

This analysis supports several conclusions and recommendation for future evaluations of energy efficiency impacts measured at the meter. While a wide variety of models tested performed well in this simulation, prescriptive requirements – requiring evaluators to use a single specific model for all evaluations – can lead to inaccurate estimates as not every model has been tested for every customer segment. Instead, a certification process should be employed. The benefit of such an approach is that it recognizes that, as broad as this accuracy study was, it was never possible to capture the specific characteristics of participants, regions, and measures of all varied population NMEC-measured programs that are currently operating or will operate in California. Results from this study, specifically around regression methods, control group matching methods and segmentation strategies, can inform future model selection under this suggested framework, but the goal is to encourage additional innovation and transparency rather than select a single approach. Instead, this approach allows for flexibility in methods while requiring transparency in how the methods are implemented and in documenting how well the methods can isolate the effect of treatment from background noise. Once a method is tested and certified, it can be applied for estimating savings without the need for extensive explanation.

Population NMEC Certification Principles

1. Certification needs to be implemented by an independent party.
2. NMEC methods need to be tested for reproducibility.
3. NMEC methods must meet pre-defined input analysis dataset structures and pre-defined output structures.
4. Population NMEC metrics of accuracy (bias) and precision should be calculated out-of-sample at a portfolio level.
5. The measurement of accuracy (bias) and precision metrics should be calculated by the independent party certifying the method using a blind test.
6. To be certified, an NMEC method must meet specific criteria for accuracy and precision.
7. NMEC methods must be separately certified for residential, small and medium businesses, and large businesses and for sites with and without solar.
8. The out-of-sample metrics for accuracy and precision of NMEC method tested for certification should be posted on a public repository such as CALMAC.
9. The code for estimating savings needs to be publicly available and include examples of how it is applied.
10. The method used must be selected and certified in advance of the program implementation.

2 INTRODUCTION

Accurate and unbiased estimates of energy efficiency (EE) impacts are critical for utility program staff, third-party program implementers, and regulators. Pacific Gas and Electric Company (PG&E) currently uses the CalTRACK Version 2.0 method (CalTRACK) to estimate avoided energy use for its energy efficiency (EE) programs based on the Population-Level NMEC (normalized metered energy consumption) methodology, including the Residential Pay-for-Performance (Res P4P) and On-Bill Financing Alternative Pathway (OBF-AP) programs. The Population NMEC method relies on whole-building site-specific electric and/or gas consumption interval data to estimate the savings associated with the installation of an individual or multiple energy efficiency measures (EEMs) at the site that is paired with typical and historical weather data to derive normalized energy savings. The advantage of this approach is that in a jurisdiction with near 100% penetration of advanced metering infrastructure (AMI), all program participants have revenue-grade meters measuring electricity consumption at hourly, or sub-hourly, intervals. Using the AMI data to estimate energy savings:

- Eliminates the need for sampling.
- Reduces by measurement and verification (M&V) burden on participants. If M&V can be conducted remotely using the revenue meter instead of sending a technician on-site, it creates less of an "ask" on program participants.
- Leads to faster feedback on energy savings to program implementers and utility staff.
- Opens up new opportunities for program design and delivery, such as pay-for-performance programs.
- Enables a better understanding of how energy savings vary for different customers and, thus, encourages better customer targeting and enables performance monitoring during the post-installation (performance) period to monitor for non-routine events and to track whether the expected savings are materializing.
- Provides the ability to time-differentiate energy savings to calculate savings load shapes.

A notable feature of the CalTRACK method[3] is the lack of comparison groups to adjust the energy savings baseline and normalize the savings estimate for factors beyond weather. This method relies almost exclusively on weather normalization and effectively assumes that the only difference between the pre- and post-intervention periods is weather and the installation of EEMs. Such an assumption has always been strong, given the frequency of other changes in energy consumption that can occur at a site, such as occupancy or installations of other end uses. Essentially, the challenge of such an evaluation framework is that it cannot effectively accommodate changes in energy consumption that are uncorrelated with weather. The COVID-19 pandemic demonstrated the degree to which methods without comparison groups are susceptible to estimation error when systemic non-routine events occur.

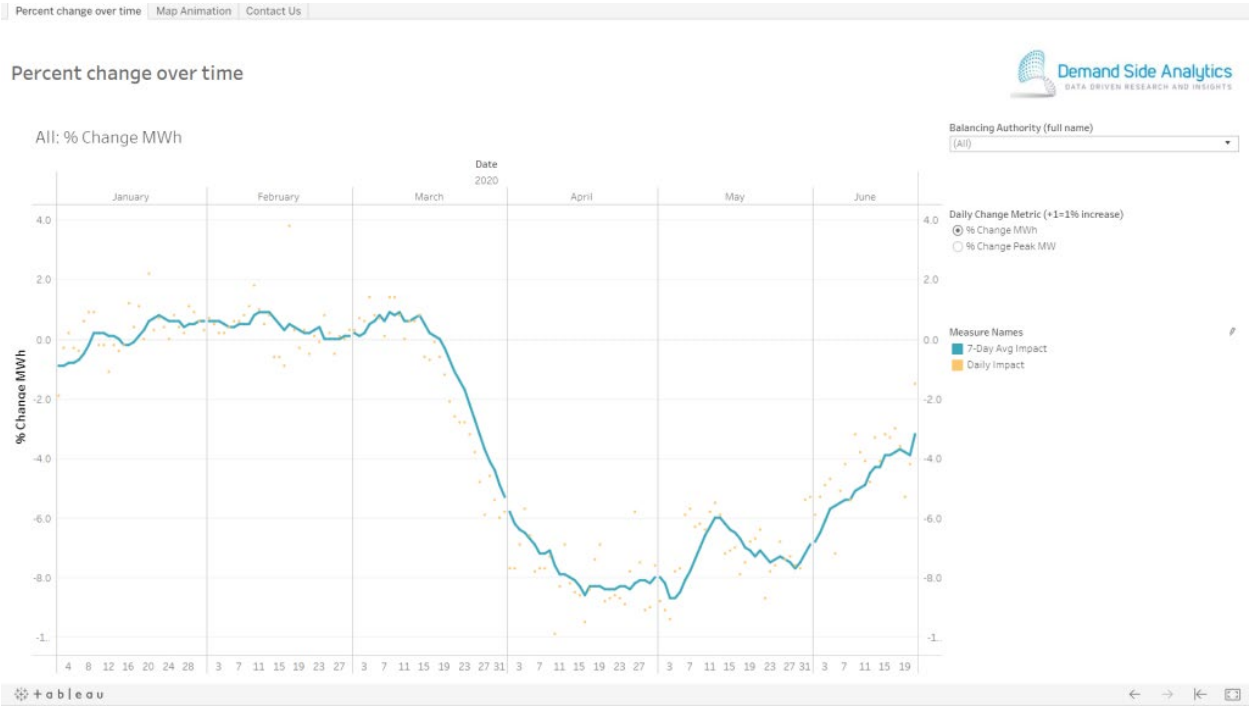
The COVID-19 pandemic laid bare the limitations of the adopted CalTRACK method. The pandemic led to changes in our commutes, business operations, and home use patterns. Not surprisingly, it has also changed

[3] See http://docs.caltrack.org/en/latest/methods.html for more detail.

how, when, and how much electricity and gas we use. Moreover, the impact on energy use differs for residential customers and various types of businesses.

The impact of the pandemic on energy use is dynamic and evolves based on the level of community spread, stay-at-home or reopening guidance, and social distancing practices. Figure 2 shows DSA's estimate of the pandemic's impact on U.S. electricity use and summarizes the daily percent change. The drop in electricity use peaked in April 2020 and rebounded over time. The effect of COVID on energy use varied by region, sector, and over time due in part to different social practices around social distancing and restrictions on travel, crowds, and indoor gatherings. The effect of COVID on energy use will persist over time, in part because it has modified commute and work-from-home patterns and because of different variants can arise over time.

Figure 2: Electricity Consumption Changes Associated with COVID-19 Pandemic



As a result of the pandemic, there has been renewed attention to the importance of incorporating the use of comparison groups to calibrate estimates of energy savings for pay-for-performance programs based on the CalTRACK method.[4]

2.1 EFFECTS OF THE COVID-19 PANDEMIC

The COVID-19 pandemic and associated stay-at-home orders in California and other parts of the country contributed to a substantial change in how energy was consumed by households and businesses. These changes fluctuated over time as different sectors of the economy alternately shut down and reopened, which in conjunction with concerns for human health, led to commensurate changes in household energy consumption. While the effects differed by sector, fuel type, and geography, the primary change in energy consumption

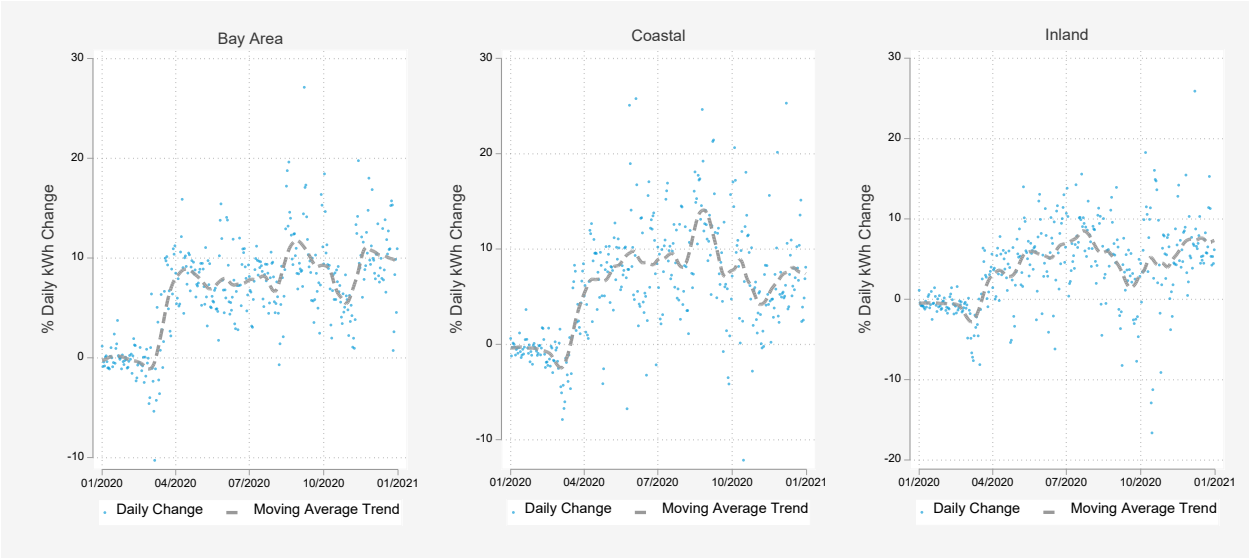
[4] For example, see the report resulting from a series of public meetings on this issue that is available here: https://grid.recurve.com/comparisons-groups-and-covid1.html.

compared to the prior year included an increase in residential energy consumption and a decrease in commercial and industrial consumption.

Residential Reference Loads

Figure 3 shows the percent change in residential electricity consumption by date as a result of the pandemic for several key segments of interest: by climate zone for customers without rooftop solar or low-income rates. These estimates of change in energy use were developed by constructing weather-normalized models of energy use from March of 2018 through February of 2020 and then predicting what consumption would have been in each segment for the post-COVID period starting in March of 2020. The scatterplot below shows daily estimates for the percent change in usage associated with the pandemic, while the grey line shows the trend over time. All residential segments showed a moderate increase in electricity consumption that persisted through the rest of 2020, but only three are shown below for clarity. The magnitude of these changes throughout the summer, fall and winter of 2020 is fairly consistent, suggesting that the behavior changes associated with stay-at-home orders and other restrictions were the primary drivers of this change. Said another way, changes in energy use in the residential sector responded in a relatively uniform way to the pandemic: shorter-term, region-specific responses to particular surges of COVID-19 infections do not seem to have been the main cause of the observed changes in usage.

Figure 3: Residential Changes in Electricity Consumption Associated with the Pandemic5



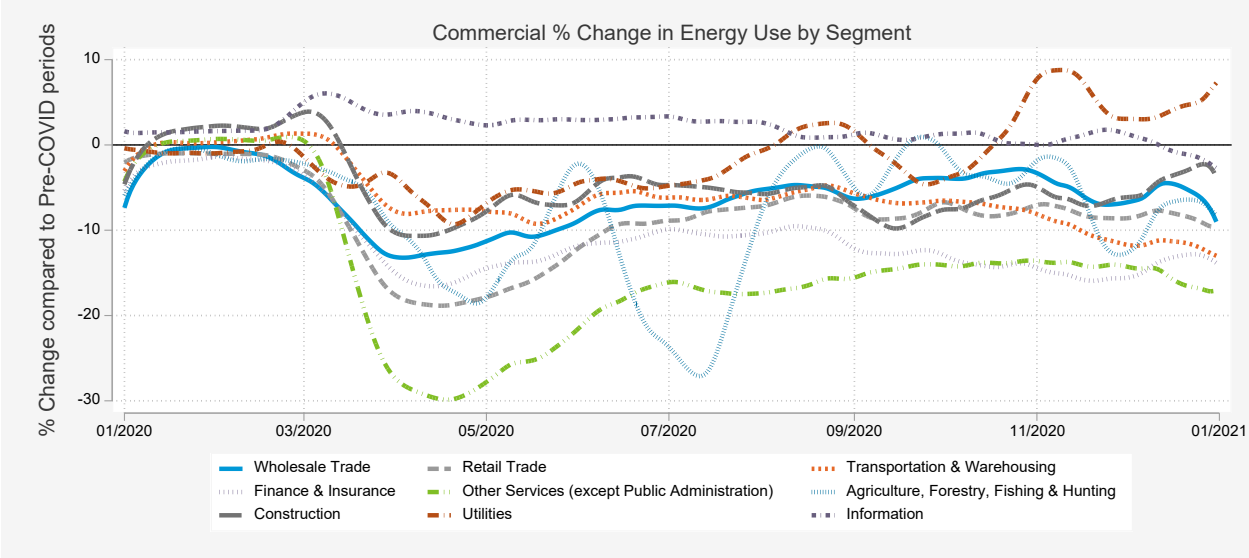
Non-Residential Reference Loads

The same analysis was conducted for commercial segments and shown in a slightly different format for Figure 4. Because customer response to the pandemic and associated designations of essential vs non-essential businesses were varied across industries, all segments with sufficient data are reported here as moving-average trend lines only. As expected, Retail, Finance and Insurance, and Other Services showed the highest decreases in consumption, normalized for weather. These segments represent businesses that would not be classified as essential, or that could relatively easily convert their operations to work-from-home configurations.

[5] For this analysis, the following climate zones were aggregated to the reported geographic locations: Coastal Bay Area: Climate Zones 1, 3, 5. Coastal Non-Bay Area: 2, 4. Inland: 11, 12, 13

Information segments actually saw an increase in consumption relative to the prior years. On average, the commercial sector saw approximately an 8-10% decline in consumption associated with the pandemic. Compared to the residential sector, the changes in consumption were slightly more variable, due to local and statewide changes in guidance about which business types were allowed to operate and in what capacity.

Figure 4: Commercial Changes in Energy Consumption Associated with the Pandemic



2.2 OPPORTUNITIES AND CHALLENGES OF USING NON-PARTICIPANT DATA FOR EVALUATION

Given the changes in energy consumption that have occurred over the course of the COVID-19 pandemic, the need for alternative approaches to CalTRACK and similar, simple pre-post regression methods for estimating EE impacts is paramount. While adding comparison groups typically improves the accuracy of these energy saving estimates, there are three, main logistical challenges:

- **Privacy of non-participant customer data.** Current California laws and regulation exist to protect the privacy of AMI/smart meter data for individual customers.6 While EE program participants consent to sharing their consumption data for the purposes of evaluation and settlement, no such agreement exists for non-participants. The regulations limit the ability to use non-participant data for settlement of pay-for-performance programs and create tradeoffs between accuracy, transparency, and access to data, particularly for third-party program implementers. Relatedly, disclosure of participant and non-participant consumption data in any public setting is limited to aggregations of customers that do not violate the 15/15 rule. This rule requires that any public disclosure of consumption patterns be aggregated such that at least 15 customers are included in the aggregation and that no one customer makes up more than 15% of the total consumption.
- **Transparency Challenges.** Many evaluation methods that rely on a comparison group require extensive calculation in order to construct the comparison group from a large pool of eligible non-

[6] Including CPUC Decision D.97-10-031 (1997) Direct Access Proceeding (15/15 rule), Senate Bill No. 1476 (2010) Chapter 5 Privacy Protections for Energy Consumption Data, and CPUC Decision D.14-05-016 (2014) Decision Adopting Rules to Provide Access to Energy Usage and Usage-related Data While Protecting Privacy of Personal Data (2014)

participants. Additionally, they require familiarity with statistical matching methods and access to high-powered servers with statistical applications to produce the matching variables and construct the comparison groups. This complex system, when operated by an experienced statistician, can result in precisely-matched comparison groups. It can hinder independent review and/or replication of the findings, however. Transparency is limited further if reviewers or implementers do not have access to the non-participant data due to privacy concerns.

- **Complexity and frequency.** PG&E and third-party EE program implementers target a wide range of customer segments and geographic areas. Program enrollment occurs throughout the year and different customer sites have unique pre-intervention periods used for baseline periods. Because developing comparison groups typically relies on finding non-participants with similar consumption profiles and demographic indicators during pre-intervention periods, it can require constructing distinct comparison groups for each implementer, participating customer segment, and treatment period; a proposition that adds complexity to existing program administration processes.

2.3 KEY RESEARCH QUESTIONS

As part of the study, we plan to address five main research questions:

- How does the current Population NMEC method (CalTRACK V2.0, without comparison groups) perform given COVID conditions, as compared to methods that adjust for changes in energy use observed in similar non-participant cohorts?
- What is the relative accuracy and precision of alternative baseline methods that use comparison groups during COVID periods?
  - ✓ To what extent do the accuracy and precision change if the installation period begins prior to the onset of COVID?
  - ✓ To what extent to the accuracy and precision change if the installation period begins after the onset of COVID?
- How does the accuracy and precision vary depending on the number of customers aggregated?
- How do baseline methods that rely on non-participant aggregated profiles perform as compared to methods that rely on customer level interval data?
- How do the results of this research inform the development of a set of recommendations for integrating comparison groups as part of Population NMEC in light of the concerns of a) measurement accuracy, b) transparency of data to all parties, c) reproducibility of results, and d) providing feedback to implementers on an ongoing basis to inform program performance?

3 METHODOLOGY

This section outlines the methodology our team used to develop the framework for the accuracy assessment as well as the different methods that were tested as a part of the assessment.

3.1 OVERVIEW OF STUDY APPROACH

The primary challenge of estimating energy savings is the need to accurately detect changes in energy consumption due to the energy efficiency intervention, while systematically eliminating plausible alternative explanations for those changes, including random chance and non-routine events. Did the introduction of energy efficiency measures cause a change in energy use? Or can the differences be explained by other factors (such as the effects of the COVID-19 pandemic)? To evaluate energy savings, it is necessary to estimate what energy consumption would have been in the absence of program intervention—the counterfactual or baseline.

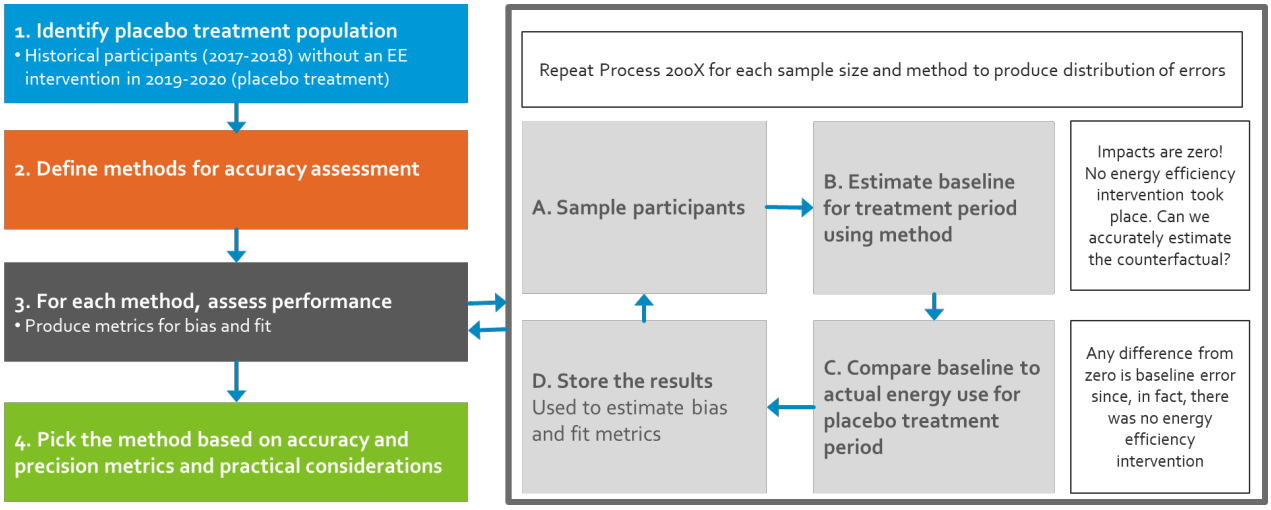
At a fundamental level, the ability to measure energy savings accurately depends on four key components:

- **The effect or signal size** – The effect size is most easily understood as the percent change in energy use following the intervention. It is easier to detect large changes than it is to identify small ones.
- **Inherent data volatility or background noise** – The more volatile the load, the more difficult it is to detect small changes. Non-routine events effectively add noise to the data.
- **The ability to filter out noise or control for volatility** – At a fundamental level, statistical models, baseline techniques, and comparison groups – no matter how simple or complex – are tools to reduce noise (or unexplained variation) and allow the effect or impact to be more easily detected.
- **Sample/population size** – For most of the programs in question, sample sizes are irrelevant since the full participant population is analyzed. However, it is still easier to precisely estimate average impacts for a large population than for a small population because individual customer behavior patterns "smooth out" and offset individual customer volatility across large populations.

To assess accuracy, one needs to know the correct values. When the correct answers are known, it is possible to determine whether each alternative method correctly measures energy use and, if not, by how much it deviates from the known values. Figure 5 summarizes the approach used to assess the accuracy and precision of alternative methods. The approach is effectively a competition, where different methods are tested repeatedly, to identify the methods that are unbiased and accurate.

Accuracy is assessed by applying placebo treatment periods on customers that did not participate in EE programs (non-participants) during the period analyzed. These customers were selected from the population of former participants in EE programs in PG&E's territory. Selecting former participants can ensure that the results in this study apply to the population of customers who typically participate in energy efficiency programs. However, because EEMs have not been installed in non-participant premises during the analysis period, any deviation between the counterfactual and actual loads is attributed to error. The statistical process of applying placebo treatments on non-participants is repeated hundreds of times – a procedure known as bootstrapping – to construct the distribution of errors from zero, which represents the constructed counterfactual of no pre/post changes in energy use due to no EEMs having been installed.

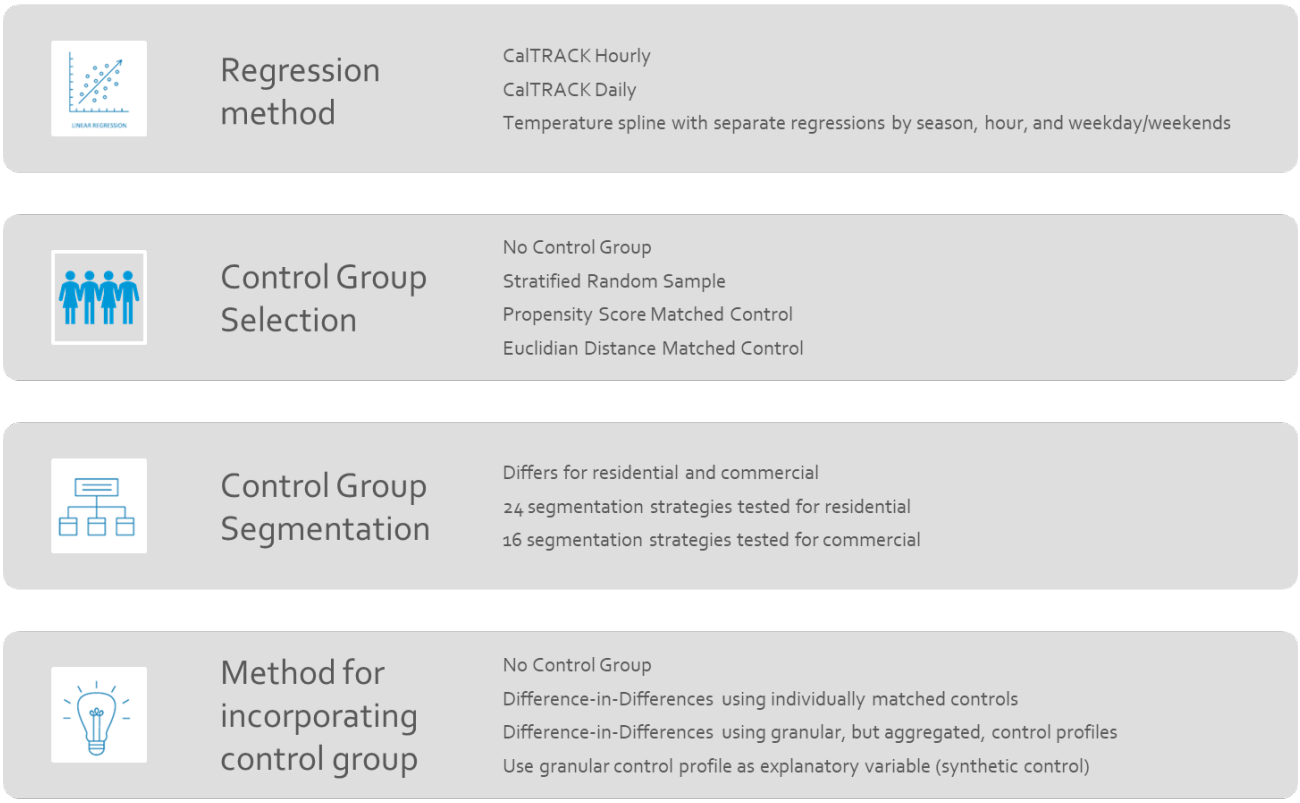
Figure 5: General Approach for Accuracy Assessment



Before the accuracy assessment can be implemented, it is necessary to identify the pseudo participant (non-participant) population, agree upon the baseline alternatives to include, and define the metrics for assessing performance. To conduct this evaluation, we used "pseudo participants" drawn randomly from the pool of customers that participated in EE programs in the 2017-2018 time period but did not participate in EE programs from 2019 onwards. These customers are most similar to recent EE program participants because they were past participants in EE programs. Critically, however, past participants do not have a current EEM treatment (or a recent treatment) in place, which allows us to conduct the accuracy assessment.

Once program participants and non-participant comparisons are identified, we can proceed with selecting a list of models to evaluate. DSA, in conjunction with PG&E, Recurve, and other stakeholders, developed a list of models to test. To ensure that feedback from all stakeholders was included, stakeholders were invited to share their input via an online survey. These models are constructed by defining a regression method, identifying the type of comparison group (if any) from which to control for exogenous factors, the method of selecting the comparison group customers, and the method by which the comparison group is incorporated into the regression model. Figure 6 below is a summary of the regression methods, comparison group types, segmentation strategies, and methods of incorporating the comparison group that were tested; each is discussed in further detail in the following section.

Figure 6: Model Options Tested



3.2 REGRESSION MODELS TESTED

Three main regression models were tested in the development of the accuracy assessment framework. Two of the three models are existing CalTRACK methods, described in detail on the CalTRACK methods website.[7] In this section, we review the model approach for each method.

CalTRACK V2 Hourly Method

The CalTRACK hourly model is a time-of-week/temperature (TOWT) pre-post model[8] that combines information about customer energy consumption during each hour of the week and outside temperature to create a model of electricity consumption.[9] Detailed documentation is available at the CalTRACK website. However, the main steps of the method include:

[7] http://docs.caltrack.org/en/latest/methods.html
[8] This model is a variant of the Temperature and Time-of-Week model developed by Lawrence Berkeley National Laboratory. More detail on this method can be found here: Matthieu, J.L., P.N. Price, S. Kiliccote, and M.A. Piette, "Quantifying Changes in Building Electricity Use, With Application to Demand Response," IEEE Transactions on Smart Grid, 2:507-518, 2011
[9] The CalTRACK documentation suggests that this model can be used for any hourly energy consumption data. Nevertheless, practically speaking gas data is only recorded on a daily basis. As a result, we only use this method for analyzing electricity consumption at the hourly level.

1. **Filtering the data to the appropriate time period.** The hourly model is run at the monthly level. To predict usage for July of 2020, the pre-treatment period is June through August of 2019, with both June and August given less weight than July in the regression specification.
2. **Construction of the occupancy flag.** Occupancy is determined at the day-of-week and hour level, for 24 x 7 = 168 unique states. The variable is constructed as a binary 0 or 1 value for each hour of the week.
3. **Construction of the temperature splines.** The model includes up to 7 temperature bins.[10] Fewer bins may be used if there are insufficient data in individual bins.
4. **Running the regression.** The regression on the pre-treatment period includes all 168 time-of-week dummy variables, the temperature spline variables, the occupancy binary variable interacted with the time-of-week dummy variables and the occupancy binary variable interacted with the temperature splines. The results are predicted using this model for the post period.

CalTRACK V2 Daily Method

The CalTRACK daily model can be run on either electricity or gas data with daily granularity or with billing data that has been normalized to the average daily consumption value for that billing period. It relies on the same data that the CalTRACK hourly method does, albeit at a lower level of granularity. Detailed documentation is available at the CalTRACK website. However, the main steps of the method include:

1. **Selecting the appropriate balance points for the model.** Heating and cooling balance points range from 30⁰F to 90⁰F, such that the heating balance point is lower than the cooling balance point and sufficient data is available for both the heating and cooling balance points
2. **Identifying the appropriate regression model.** Four models are tested, including both heating degree day and cooling degree day variables, either heating- or cooling-only degree day models, and a model with an intercept only. The winning candidate model must meet requirements for a high R-squared value as well as the correct (positive) sign on all coefficients.
3. **Running the regression.** The regression is run for the full pre-treatment year and then is predicted for the post-treatment period.

Alternative Regression Models

The remaining models tested were simple variations on temperature spline[11] models. The motivation for including these models was to test the performance of relatively simple alternative specifications that included temperature, time, season and weekday components, similar to the CalTRACK TTOW models. In these models, all data is used in a single model for the pre-treatment year. Four different spline models were tested:

1. Hourly temperature

[10] The initial number of bins proposed by the model are: <30F, 30-45F, 45-55F, 55-65F, 65-75F, 75-90F, >90F.
[11] A spline model in this context is a linear regression where the temperature variable of interest is sorted from lowest to highest, grouped into bins of similar temperatures, and the bins are then used as variables in the regression, rather than the temperature variable itself. This is done to account for varying participant responses to temperature, similar to a cooling degree hour or cooling degree day. Said another way, the relationship between temperature and energy consumption from 40-45F will be different to the relationship between temperature and energy consumption from 80-85F.

2. Average daily temperature
3. 3-hour moving average hourly temperature
4. 12-hour moving average hourly temperature

For daily models with this structure, only the average daily temperature spline model is constructed. All splines were constructed as 4-part splines, with temperature bins of: <50⁰F, 50⁰-60⁰F, 60⁰-70⁰F, >70⁰F. All of these spline models were run independently for each hour of the day, weekdays and weekends, and for summer and non-summer months (where summer was defined as May to October).

3.3 MATCHING AND SEGMENTATION

The regression methods listed above were tested with a variety of comparison group options incorporated in different ways into the regression model. The benefit of a comparison group is to ensure that exogenous factors that influence energy use can be fully accounted for and are not misattributed to the effect of the treatment. For a comparison group to be a good proxy for what the participants would have done in the absence of treatment, evaluators should ensure that they are statistically similar to the participants on observable characteristics such as size, location, or solar status. There are a variety of ways to construct this statistically-similar group of non-participants; in this assessment we reviewed three common matching methods:

1. **Stratified random sample**: for each participant, randomly select a non-participant to act as a comparator within a grouping of segmentation characteristics. For example, non-participants must be matched to participants within the same climate zone and rooftop solar system installation status (yes/no).
2. **Propensity score matching**: within each segment of interest, fit a probit model with the treatment indicator as the outcome variable and specified matching variables as explanatory variables and then predict the outcome. This predicted result represents the propensity, given the characteristics included in the model, to participate. Match participants to non-participants with similar propensities to participate in the EE program.
3. **Euclidian distance matching**: within each segment of interest, compute the Euclidian distance between each participant and each non-participant on the basis of all the matching characteristics of interest. Find the non-participant that has the smallest Euclidian distance to the participant.

For each of these methods, one-to-one matching was employed, where a single non-participant was matched to each participant. Non-participants were allowed to be matched to multiple participants if that non-participant was the best match for more than one participant.

In all three matching methods, matches were constructed within pre-defined customer segments, or groups of customers with similar characteristics. For both propensity score matching and Euclidian distance matching, other characteristics formed the basis of the selection of the match within each segment (non-participants are randomly assigned in the stratified random sample approach). Table 1 summarizes the segmentation and matching characteristics for each customer sector tested. In all cases, customers were matched within their climate zone and solar status at a minimum. Matching on climate zone ensures that pseudo-participants get matched to controls that experience similar weather conditions. Further study could explore whether matching

customers within their weather station catchment provides a higher-quality match as controls would then experience the exact same weather rather than similar weather to participants.

Table 1: Matching Characteristics

Sector	Segmentation Strategies Tested	Matching Characteristics Tested
Residential Electric	Climate Zone, Solar Onsite, and: • Annual Consumption (4 Bins) • Summer Consumption & Winter Consumption (4 Bins Each) • 24-hour Load Shape (2 Bins) and 12-Month Consumption Profile (4 Bins) ▪ DER System Size	• Mean Summer Demand 4pm-9pm • Load Factor • Peak:Off Peak Ratio • Load Factor and Cooling Weather Sensitivity • Load Factor and Bins of Annual Consumption (100 Bins) • Load Factor and Bins of Peak Period Consumption (100 Bins)
Commercial Electric	Climate Zone, Solar Onsite, and: • Annual Consumption (50 Bins) • Annual Consumption (4 Bins) and Peak Load (4 Bins) • 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins) • 1st and 2nd Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	• Bins of Annual Consumption (100 Bins) • Bins of Peak Period Consumption (100 Bins)
Residential Gas	Climate Zone and: • Annual Consumption (4 Bins) • Summer Consumption & Winter Consumption (4 Bins Each) • Annual Consumption (4 Bins and 12-Month Consumption Profile (2 Bins) • Annual Consumption (4 Bins) and Low Income Flag	• Percentile ranking of annual consumption • Percentile ranking of summer and winter consumption • Bins of Heating Weather Sensitivity (4 Bins) • Bins of Heating Weather Sensitivity (100 Bins)
Commercial Gas	Climate Zone and: • Annual Consumption (50 Bins) • Summer Consumption & Winter Consumption (4 Bins Each) • 1st Digit NAICS Code, Annual Consumption (4 Bins and 12-Month Consumption Profile (2 Bins) • 1st and 2nd Digit NAICS Code, Annual Consumption (4 Bins and 12-Month Consumption Profile (2 Bins)	• Percentile ranking of annual consumption • Percentile ranking of summer and winter consumption • Bins of Heating Weather Sensitivity (4 Bins) Bins of Heating Weather Sensitivity (100 Bins)

3.4 COMBINING MATCHING AND REGRESSION METHODS

Once the comparison groups have been constructed, they are added to the regression methods. There are four methods by which the comparison group can be added to each regression model:

1. **No comparison group.** The method is run as-is with no comparison group
2. **Using a difference-in-differences approach with individual matched comparators.** In this approach, the regression model is run for all participants and their matched comparison groups. The participant and matched comparator observed consumption and counterfactual usage are averaged for each time interval and treatment status. The impact estimated in each time interval in the pre and post period for the comparison group is then subtracted from the impact in the corresponding interval for the participant group. This difference-in-differences impact is then the load reduction associated with treatment in the participant group.
3. **Constructing a synthetic control group.** This involves adding one or more 8,760 (or 365 for daily methods) aggregated control group profiles to the specification as a right-hand-side/explanatory variable. This approach relies on these added profiles to construct a synthetic control that exploits the relationship of consumption patterns between the aggregated control group loads and the participant loads during the pre-treatment period to predict participant usage during the post-treatment period. The synthetic control profiles were constructed using the same segmentation groups as discussed above. For each control profile, 100 customers in a given segment were randomly sampled and their 8,760 load profile was aggregated. These aggregated profiles were included as right-hand-side variables in pseudo-participant regression models.
4. **Using a difference-in-differences approach with an aggregated control group profile.** Instead of a specific control customer. Instead of relying on individual customer matches for the difference-in-differences approach described in #2, produce a counterfactual for an aggregated control profile as would be done for an individual customer. Use that estimate for the comparator profiles to net out exogenous changes in consumption.

At the direction of PG&E and other stakeholders, DSA investigated some or all of these methods of integrating a comparison group into the existing regression methods. Table 2 summarizes the options included in this analysis.

Table 2: Regression Methods with Comparison Options Tested

Regression Method	Without Control	Matched Comparator DID	Granular Profile on RHS	Granular Profile DID
CalTRACK Hourly	✓	✓	✓*	✓*
CalTRACK Daily	✓	✓		
Alternative Regression		✓	✓	

* These granular profile tests were done on a subset of the full accuracy assessment and are described in more detail in Section 4.6

3.5 BOOTSTRAPPING AND AGGREGATION OF ERRORS

Any accuracy assessment should deliver robust recommendations that work for a variety of populations and sample sizes. Said another way, an accuracy assessment should take care to ensure that the results obtained are not due to the unique characteristics of a particular subset of customers sampled but represent the performance of the model in general. To accomplish this, the DSA team ran this accuracy assessment while bootstrapping the sampled participants with their matched controls and aggregating the results at various

sample sizes. Bootstrapping involves repeatedly drawing random samples from the population in question, and then computing the impacts (for an accuracy assessment where no treatment was in place, impacts are equivalent to errors). Bootstrapping allows for a review of results:

- At different levels of aggregation (different sample sizes)
- For individual iterations of the bootstrap
- For the full population in general

DSA ran the bootstrapping exercise for 200 iterations for each sector (residential electric, residential gas, commercial electric and commercial gas[12]) and for each sample size (5, 10, 25, 50, 100, 500, 1,000 participants). Unless otherwise noted below, results are reported for the average across all bootstrapped results for a given sample size.

3.6 DEFINING THE BEST MODEL

It is often helpful to conceptualize the process of conducting an accuracy assessment as a tournament: the candidates are defined up front and the rules for how the contest will be conducted and judged are not changed after the fact. As noted in the introduction, both quantitative and qualitative factors will influence which proposed method will ultimately be recommended. In this section, we discuss the evaluation criteria for defining the best model(s) for each sector.

Quantitative Considerations

The quantitative assessment itself results in measures of accuracy and precision that are relatively straightforward to compute and interpret. We recommend measuring both the bias and precision of the energy savings estimates. Bias refers to the tendency to over or under predict savings, while precision refers to how close the savings predictions are to actual answers (regardless of direction). The figure below illustrates the difference between bias and precision. An ideal method is both unbiased and precise (example #1). Estimates can be accurate but imprecise when errors are large but cancel each other out (#2). They can also exhibit false precision when the results are very similar but are biased (#3). The worst estimates are both biased and imprecise (#4).

[12] The total number of pseudo participants available for the bootstrap was: 22,642 for residential electric, 18,379 for commercial electric, 74,251 for residential gas, and 9,006 for commercial gas.

Figure 7: Conceptual Demonstration of Accuracy and Precision

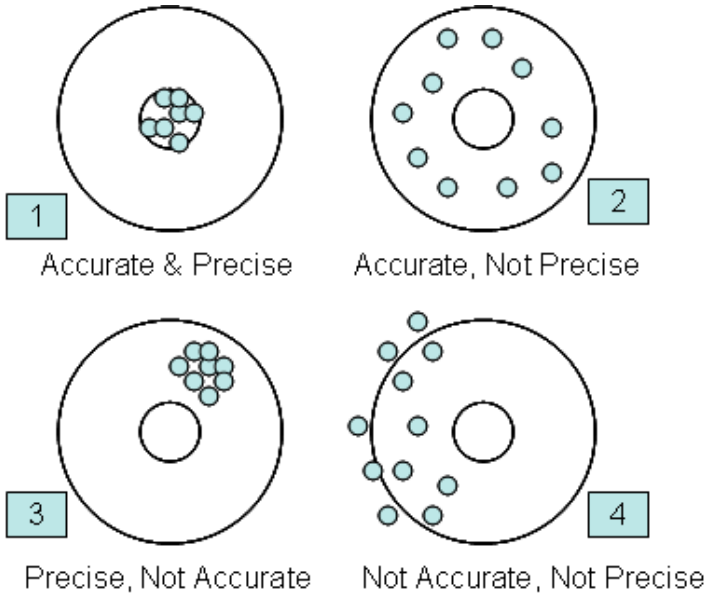


Table 3 summarizes metrics for accuracy (bias) and precision that were used for assessing performance. Assessing both accuracy and precision is clearly useful for quantifying errors in each method. It is important to distinguish the level at which these values can be computed, however. For example, bias and precision can be calculated for an individual site, where the % Bias indicates the percent by which the method tends to overstate or understate the savings for that site and the relative RMSE (CVRMSE) represents the relative "noisiness" of errors for an individual hour. Nevertheless, the % Bias and CVRMSE should be computed at the aggregate portfolio level, where the % Bias represents the average percentage above or below the observed consumption across all the aggregated post-treatment period consumption and the CVRMSE represents the variation in the aggregated post-treatment period error for each bootstrapped iteration. Assessing accuracy and precision at the aggregate portfolio level does mean that any errors associated with each method at large individual sites within that group will have a larger effect on the overall portfolio performance for that group. However, as settlement and program performance are estimated at the portfolio level, we believe that this level of aggregation should be prioritized in the accuracy assessment.

In this assessment, we will focus primarily on investigating the accuracy and precision in this second, aggregate, meaning. Investigating the distribution of errors at the individual site level will provide incremental, but secondary, information to our recommendation. This is because both PG&E's claimable and payable savings are calculated at the portfolio (aggregate) level. Finally, while the error associated with annual consumption is of primary interest in this use case, we will also report accuracy and precision associated with peak period savings where appropriate given that certain PG&E contracts provide for payable savings bonuses for savings achieved during certain hours of the week during the summer.

Table 3: Accuracy and Precision Metrics

Type of Metric	Metric	Description	Mathematical Expression
Bias	% Bias	Indicates the percentage by which the measurement, on average, over or underestimates the true energy savings.	$\% \, Bias = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)}{\bar{y}}$
Precision	Relative RMSE or CVRMSE	Measures the relative magnitude of errors, weighting more extreme errors more heavily.	$RRMSE = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\bar{y}}$

Qualitative Considerations

While the quantitative considerations listed above are relatively straightforward to compute and interpret, they must be subject to some qualitative interpretation as well. As much as possible, the best methods should adhere to the following principles to be considered a viable candidate:

- **Privacy:** disclosure of individual non-participant load data should be limited and shared only with authorized parties that are capable of maintaining its confidentiality. Reporting of results should not violate the 15/15 rule or any other PG&E or California rules or regulations surrounding customer personally identifiable information (PII).
- **Transparency:** PG&E, implementers, regulators, and evaluators should be able to understand the composition of the comparison group and how the methods used to calculate impacts were applied. Comparison groups, regression methods and results should be replicable and well documented. Moreover, Implementers and other interested third parties should be able to replicate the analysis given that their performance payments are based on the results.
- **Computationally Straightforward:** Methods that rely on straightforward regression methods are preferred over those that require excessive explanatory variables or multi-stage algorithms. This principle of parsimony also holds for the construction of the comparison groups, if applicable.

4 RESULTS

The following section reviews the results of the bootstrapped accuracy assessment for each sector: residential electric consumption, commercial electric consumption, and residential and commercial gas consumption. Further details surrounding each model and specific results can be found in the appendix. Each section summarizes the accuracy of all models, the accuracy of the subset of best models, and the results for the best model from each of the main regression frameworks. Results are summarized at different levels of aggregation and with different levels of exposure to COVID conditions. Finally, because the accuracy of different models may vary by customer segment, some results are shown for a variety of customer segments, as well as for the population overall.

4.1 ACCURACY OF EXISTING CALTRACK MODEL

The first research question in this study is to quantify the degree to which the existing Population NMEC methods (CalTRACK V2.0 models without comparison groups) are biased or unbiased during pandemic-related economic conditions. Figure 8 and Figure 10 show the bias observed in the CalTRACK Daily and Hourly models, for residential and commercial gas and electric consumption as a function of bootstrapped sample size and the amount of the post-treatment period impacted by the COVID-19 pandemic. In general, both CalTRACK models showed bias that was consistent across sample sizes, meaning that the observed values are not attributable to the level of aggregation but to the models themselves. There are clear trends in CalTRACK model performance with increasing exposure to the pandemic, however.[13] For commercial customers, more exposure to pandemic conditions in the post-treatment period leads to higher model bias. In Figure 8, a positive error indicates that the counterfactual consumption is higher than the observed consumption, such that the method records a positive energy "savings" in this pseudo-experiment. This aligns with our expectations that the CalTRACK model, when applied in the context of pandemic-related shutdowns, which on average reduced energy use in the commercial sector relative to the prior year, would yield upward bias in the absence of a comparison group and therefore overestimate energy savings in this sector. For the residential sector, the opposite trend holds: in the absence of any pandemic-related conditions, there is an upward bias of

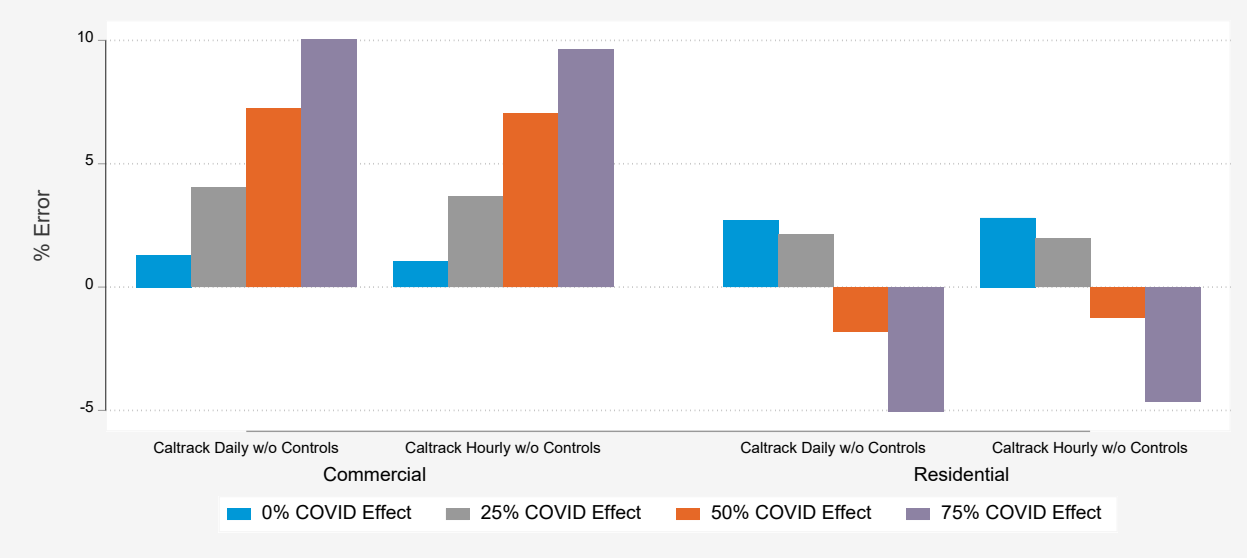
Key Finding

Population NMEC methods without comparison groups cannot account for the effects of the COVID-19 pandemic

[13] Treatment of the pseudo participants was randomly assigned by quarter of the year, with pseudo-treatment beginning in January of 2019. The results in these tables therefore represent different levels of exposure to pandemic conditions in the full year post-treatment. Customers treated in January-March of 2019 have essentially no exposure to pandemic conditions as their year of post-treatment runs through March of 2020 at the latest. Customers with 75% COVID exposure had pseudo-treatment starting in October-December of 2019, meaning that their post-treatment period completed by December 2020 at the latest.

approximately 3%, regardless of sample size in the residential electric sector. With additional pandemic exposure, the bias moves from positive to negative, due to the increased residential consumption associated with stay-at-home orders in California. This aligns with our expectations that the CalTRACK model, when applied in the context of pandemic-related shutdowns, which on average raised energy use in the residential sector relative to the prior year, would yield downward bias in the absence of a comparison group, and therefore underestimate energy savings in this sector.

Figure 8: Errors for CalTRACK without Controls on Electric Consumption



In summary, both residential and commercial CalTRACK models shows savings when there aren't any, even in the absence of pandemic-related conditions. This result is consistent across both sectors and is attributable to the lack of comparison group used to control for exogenous changes that can affect energy consumption.

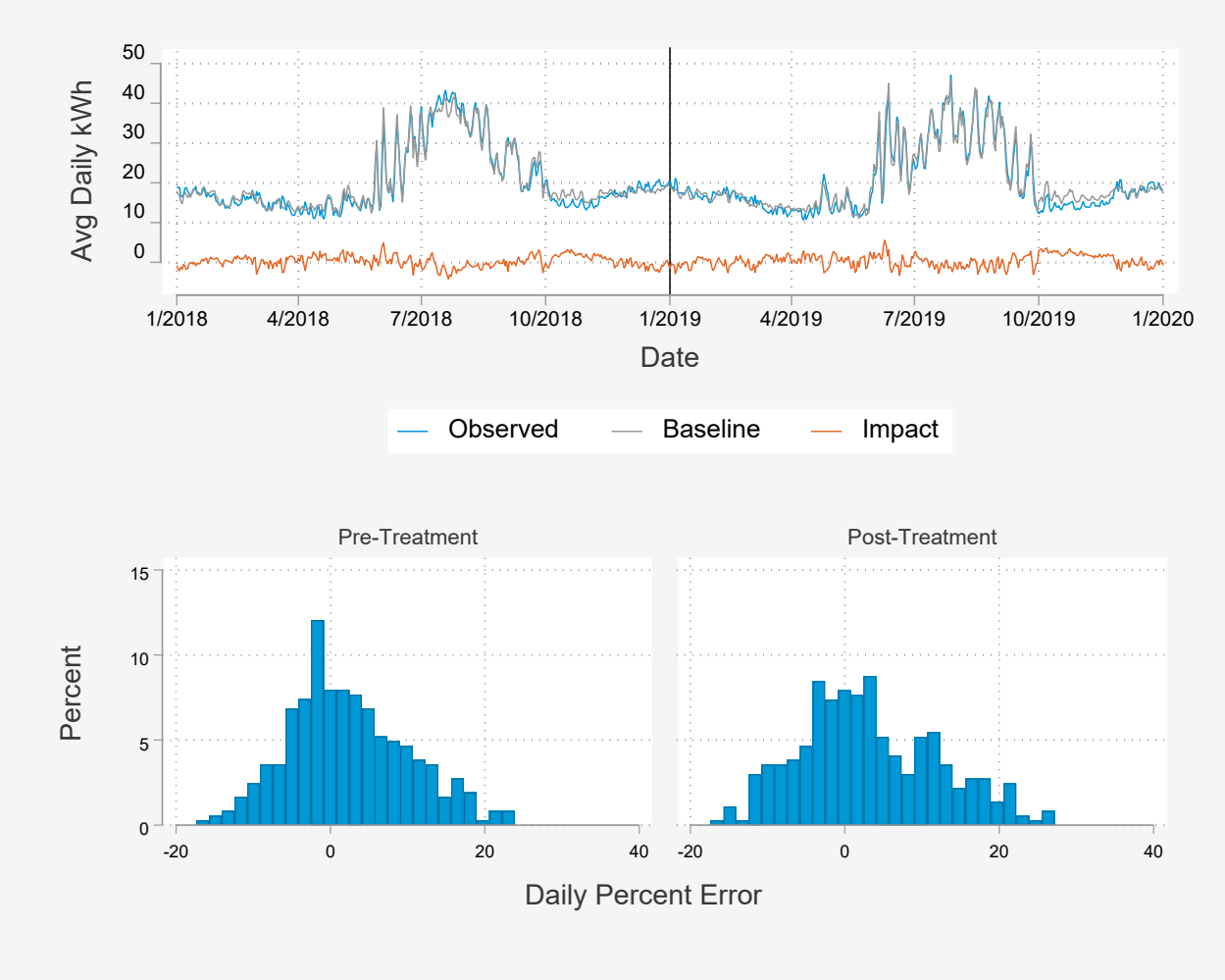
Figure 9 shows an alternate view of the results presented in Figure 8 for a random subset of 1,200 residential electric customers using the CalTRACK Hourly model without controls. These customers were assigned a pseudo-treatment date of January 1, 2019, meaning that their first-year post-treatment period (calendar year 2019) did not include any pandemic effects. The top pane of the graph shows daily average consumption for each day in the pre- and post-treatment periods. The orange line shows the actual or observed consumption for these customers, while the grey line shows the counterfactual calculated by the CalTRACK Hourly model. The orange line is the difference

Key Finding

The existing population NMEC methods without comparison groups show upward bias even prior to the effects of the pandemic

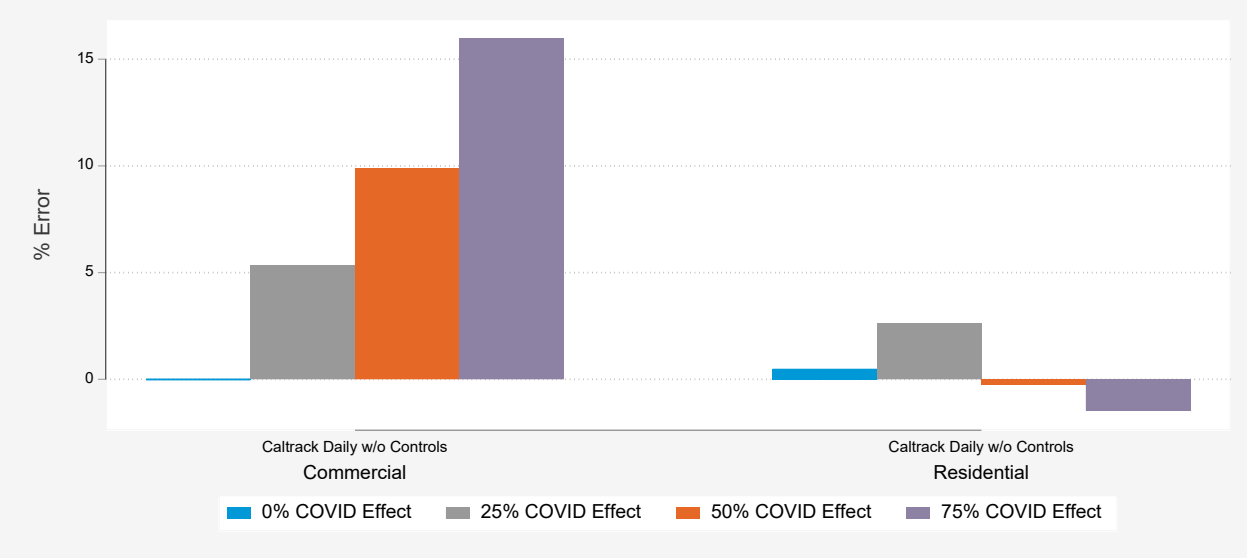
between the two, or the impact (where a reduction in energy use relative to the counterfactual is a positive value). The histograms in the bottom panes of the graph shows the distribution of percent impacts – equivalent to the percent error of the baseline for each day in the pre- and post-treatment periods. The further the bars in the bottom graphs are from zero indicates how much error there is in a given day, with the height of each bar indicating how frequently that level of error occurs in either the pre- or post-treatment periods.

Figure 9: Example Annual Profile for Residential Electric Consumption



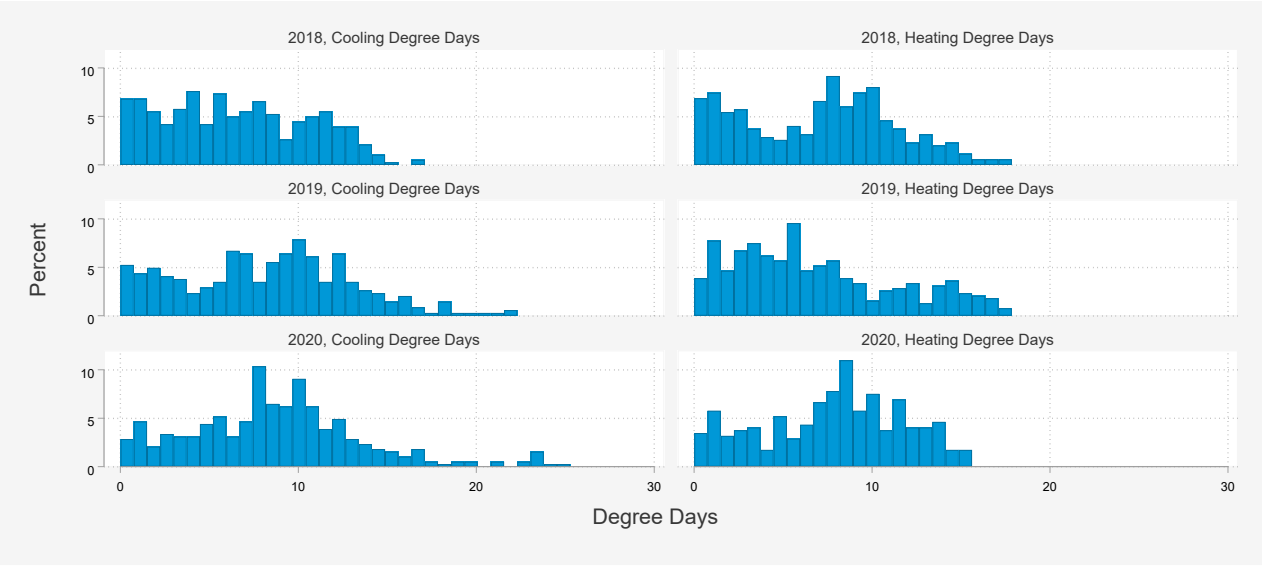
As shown in Figure 10, the trend of the pre-post CalTRACK model to overestimate the observed consumption for the commercial gas sector is the same as it is on the electric side, which is consistent with the hypothesis that pandemic conditions were the primary driver of reductions in gas consumption due to business shut downs. For residential gas consumption, the general trend of higher consumption with more exposure to the pandemic continues to hold. The magnitudes of the differences, especially in the residential sector, are smaller than for electric customers. This may be because of the timing of pandemic-related changes in gas consumption interacting with the times of year where gas tended to be consumed.

Figure 10: Errors for CalTRACK without Controls on Gas Consumption



As discussed in the introduction, CalTRACK models on their own are pre-post models that assume that the only change influencing energy use between the baseline and reporting periods, besides the introduction of the EEMs, is weather. It therefore stands to reason that any assessment of the accuracy of CalTRACK models also should look at differences in weather during the analysis period. Figure 11 shows the distribution of heating and cooling degree days, by year, for a random sample of non-participants, on the x (horizontal) axis. The heatwave of August and September of 2020 is clearly shown in the bottom left panel of the graph by the relatively higher incidence of cooling degree days in that year relative to the other two years. In general, the trend of warmer summers and warmer winters across these three years is clear, with a higher number of cooling degree days and a fewer number of heating degree days seen in the progressing years.

Figure 11: Distributions of Heating and Cooling Degree Days by Year

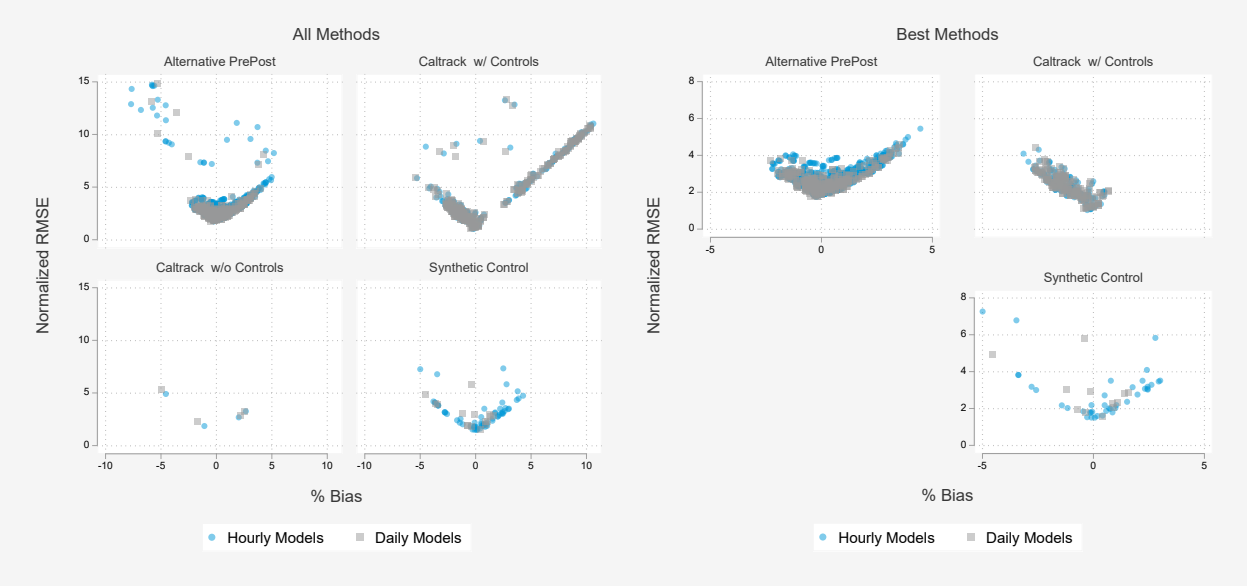


Of course, relatively warmer weather was not the only difference between 2020 and the prior years. The way in which heating- and cooling-related energy use interacted with pandemic-related business closures and stay-at-home orders in PG&E's service territory had a profound effect on the ability to measure the impacts of energy efficiency.

4.2 ACCURACY ASSESSMENT RESULTS – RESIDENTIAL ELECTRIC

Figure 12 shows the distribution of accuracy metrics for all models tested, as well as a subset of the best models. The selection of these 'best' models in this figure were chosen because their mean absolute bias across all sample sizes and treatment periods was less than 2% and their normalized RMSE was less than 20% at the aggregate portfolio level. To improve readability, the models shown in the figure are for aggregations of 500 participants. More stringent selection criteria were used to determine the overall recommendations shown in the subsequent sections. There were many models that performed well, with low bias and high precision, though there were many models that overstated the observed usage, leading to high upward bias. As discussed in the previous section, warmer summers in 2019 and 2020, compared to the pre-treatment 2018 and 2019 summers, could explain why counterfactual or baseline estimates of loads are higher than the observed loads. The CalTRACK Daily and Hourly methods with matched comparison groups were generally among the best-performing models.

Figure 12: Overall Accuracy and Precision for Residential Electric Models



BEST MODEL

The best model among those tested for residential electric consumption is the CalTRACK Daily model with a matched control group. This model has a bias of -0.17% and a normalized RMSE of 0.86%. This model is shown in Table 4, along with the best models for all other frameworks, summarized across the bootstrapped random samples of 1,000, and averaged across all treatment periods. The next best model is the CalTRACK Hourly model with the same matched control group. In most cases, the

segmentation strategy that produced the best results was a combination of climate zone, solar status, DER system size and bins of annual consumption.

Table 4: Best Model by Framework for Residential Electric Usage

Framework	Matching Method	Segmentation	Matching Characteristics	Model	% Bias	Normalized RMSE (%)
CalTRACK Daily w/ Controls	Euclidian	Climate Zone, Solar Onsite, DER System Size	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Daily	-0.17	0.87
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, DER System Size	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	-0.17	0.86
Alternative Pre-Post - Hourly	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	-0.11	1.50
Alternative Pre-Post - Daily	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	0.06	1.51
Synthetic Control - Daily	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	0.49	1.66
Synthetic Control - Hourly	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	N/A	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	0.91	1.78
CalTRACK Hourly w/o Controls	Participants Only	N/A	N/A	CalTRACK Hourly	-0.27	2.95
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	-0.51	3.20

Best Models as a Function of Pandemic Exposure and Sample Size

Table 5 and Table 6 show results for the top three overall models for this sector, along with results for the best performing models in each framework. The best overall models are highlighted in blue. The results in Table 5 are shown for an average across all sample sizes (from 5 customers to 1000) for each treatment period. Treatment of the pseudo participants was randomly assigned by quarter of the year, with pseudo-treatment beginning in January of 2019. The results in this table therefore represent different levels of exposure to pandemic conditions in the full year post-treatment. Customers treated in January-

Key Finding

Comparison groups improve accuracy of the NMEC method.

March of 2019 have essentially no exposure to pandemic conditions as their year of post-treatment runs through March of 2020 at the latest. Customers with 75% COVID exposure had pseudo-treatment starting in October-December of 2019, meaning that their post-treatment period completed by December 2020 at the latest. The top three models relied on CalTRACK daily or hourly models with essentially the same matched control group. The bias of each model varied with exposure to pandemic conditions, with positive bias in pre-pandemic simulations compared to negative bias with exposure to the pandemic. For the best model, however, the magnitude of the bias does not change substantially.

Table 6 shows the performance of each method as a function of sample aggregation. These results are averaged across COVID exposure quarters. As expected, both bias and precision improved with higher levels of aggregation.

For the models that relied on a matched control group, the best models relied on Euclidian distance matching with load factor and bins of annual consumption. Most of the best models relied either on segmentation of solar status, climate zone, and either bins of annual consumption or solar system size.

Table 5: Residential Electric Model Results for Different Levels of Pandemic Exposure

Framework	Matching Method	Segmentation	Matching Characteristics	Model	Value	% COVID Effect			
						0%	25%	50%	75%
CalTRACK Daily w/ Controls	Euclidian	Climate Zone, Solar Onsite, DER System Size	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Daily	% Bias	0.03	-0.18	-0.22	-0.18
					Norm. RMSE	6.30	5.62	4.34	4.81
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, 24-hour Load Shape (2 Bins) and 12-Month Consumption Profile (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	% Bias	0.27	-0.16	-0.06	0.01
					Norm. RMSE	6.24	5.55	4.36	4.85
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	% Bias	0.25	-0.17	-0.04	0.03
					Norm. RMSE	6.26	5.55	4.34	4.83
Alternative Pre-Post - Daily	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	0.57	0.11	-0.06	-0.31
					Norm. RMSE	11.31	8.45	7.15	7.98
Alternative Pre-Post - Hourly	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	% Bias	0.08	-0.71	0.17	-0.34
					Norm. RMSE	11.28	9.93	7.35	7.72
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	% Bias	2.85	2.06	-1.71	-4.89
					Norm. RMSE	9.51	8.87	6.11	8.52
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, DER System Size	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	% Bias	0.03	-0.44	-0.02	-0.15
					Norm. RMSE	6.17	5.67	4.33	4.93
CalTRACK Hourly w/o Controls	Participants Only	N/A	N/A	CalTRACK Hourly	% Bias	2.91	1.89	-1.10	-4.44
					Norm. RMSE	9.32	8.83	5.90	8.30
Synthetic Control - Daily	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	1.69	0.93	0.27	-0.67
					Norm. RMSE	11.32	10.01	5.99	6.67
Synthetic Control - Hourly	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	N/A	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	% Bias	1.45	2.30	-0.13	-0.05
					Norm. RMSE	8.07	9.17	5.74	6.32

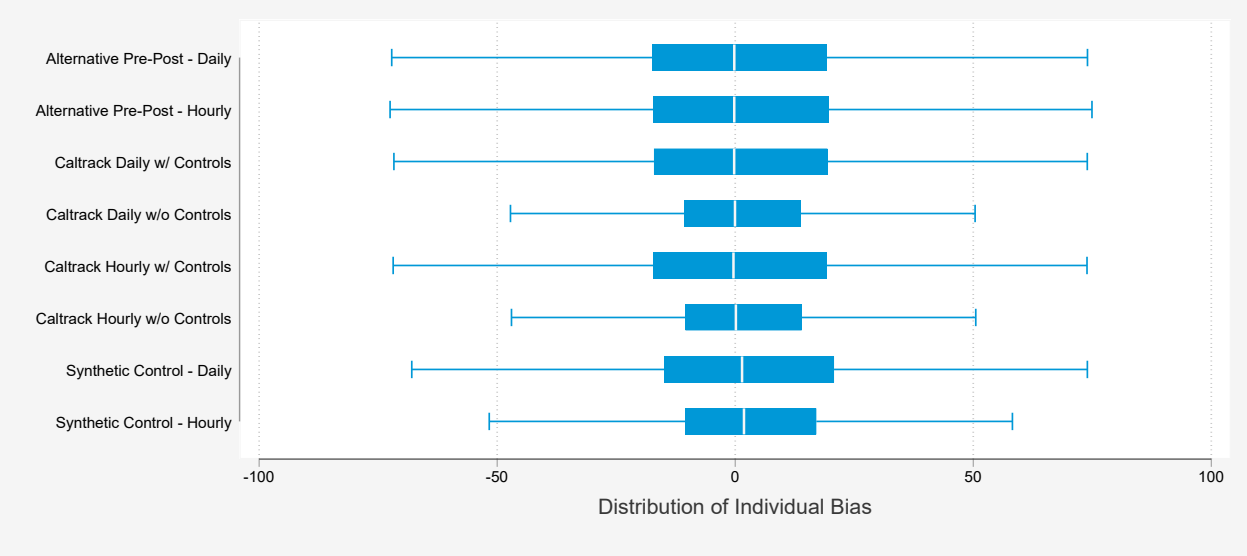
Table 6: Residential Electric Model Results for Different Sample Sizes

Framework	Matching Method	Segmentation	Matching Characteristics	Model	Value	Sample Aggregation						
						5	10	25	50	100	500	1000
CalTRACK Daily w/ Controls	Euclidian	Climate Zone, Solar Onsite, DER System Size	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Daily	% Bias	-0.07	-0.32	0.18	-0.41	-0.06	-0.12	-0.17
					Norm. RMSE	13.58	8.54	5.57	4.18	2.87	1.25	0.87
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, 24-hour Load Shape (2 Bins) and 12-Month Consumption Profile (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	% Bias	-0.01	-0.08	0.39	-0.23	0.02	0.02	-0.02
					Norm. RMSE	13.48	8.56	5.71	4.05	2.87	1.23	0.84
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	% Bias	-0.03	-0.08	0.39	-0.21	0.03	0.03	-0.01
					Norm. RMSE	13.48	8.58	5.68	4.03	2.87	1.23	0.85
Alternative Pre-Post - Daily	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	-0.41	0.24	0.51	-0.12	0.08	0.15	0.06
					Norm. RMSE	20.06	15.40	10.32	6.87	4.64	2.22	1.51
Alternative Pre-Post - Hourly	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	% Bias	-0.94	-0.38	0.30	-0.19	-0.09	0.00	-0.11
					Norm. RMSE	21.89	16.09	10.24	6.85	4.78	2.16	1.50
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	% Bias	-0.51	-0.40	-0.23	-0.46	-0.37	-0.48	-0.51
					Norm. RMSE	19.29	12.04	7.94	6.79	5.12	3.40	3.20
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, DER System Size	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	% Bias	-0.08	-0.32	0.18	-0.41	-0.09	-0.13	-0.17
					Norm. RMSE	13.43	8.71	5.61	4.18	2.87	1.26	0.86
CalTRACK Hourly w/o Controls	Participants Only	N/A	N/A	CalTRACK Hourly	% Bias	-0.28	-0.15	0.01	-0.23	-0.13	-0.24	-0.27
					Norm. RMSE	18.99	12.10	7.84	6.61	4.94	3.18	2.95
Synthetic Control - Daily	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	0.23	0.67	0.97	0.43	0.57	0.54	0.49
					Norm. RMSE	22.71	13.45	8.07	6.83	4.63	2.16	1.66
Synthetic Control - Hourly	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	N/A	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	% Bias	0.25	0.93	1.39	0.92	0.94	0.93	0.91
					Norm. RMSE	17.15	12.46	7.63	5.86	4.23	2.18	1.78

Distribution of Individual Errors

Figure 13 shows the distribution of individual participant errors (in a box plot) for the best method for each framework for the full pseudo-participant population. For each row, the blue box indicates the range of outcomes associated with the 25th-75th percentile, while the white line indicates the median value. The whiskers (or lines) are 1.5 times the interquartile range.[14] To enhance readability, outliers have been omitted from this plot. The best method at the aggregate, bootstrapped level is the CalTRACK Daily model with a matched comparison group. The distributions of the best frameworks tend to be unbiased when looked at in this manner, as the box plot omits outliers from the graphic, however. This indicates that for the majority of customers – those in the blue rectangles for each method – all methods perform reasonably well at explaining energy consumption. For all methods that do not rely on a differencing strategy – the CalTRACK without controls and Synthetic Control models – the error bands tend to be narrower. This finding suggests that, while comparison groups perform better in aggregate, individual customers may not be matched to their optimal matched comparison customer, leading to error. Nevertheless, as long as the method is unbiased in aggregate, the errors for individual customers will cancel each other out.

Figure 13: Residential Electric Individual Error Distribution



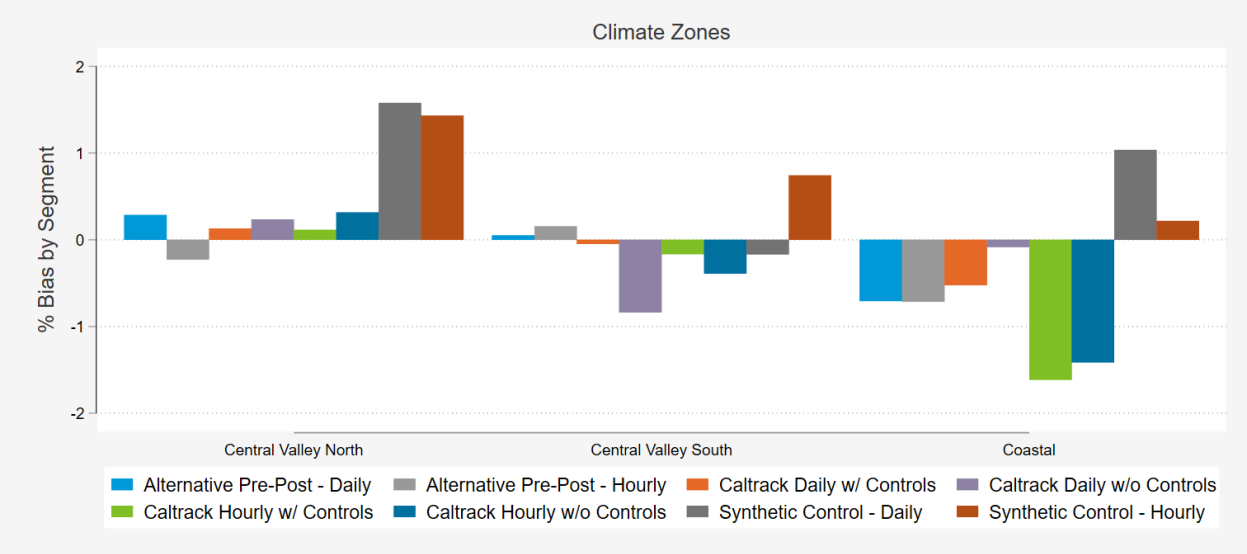
Best Model Results for Segments of Interest

This section reviews results by key customer segments, including groups of CEC climate zones (Figure 14), low income status (Figure 15), rooftop solar status (Figure 16), and customer size (Figure 17). Most models performed similarly for the North and South-Central Valley climate zones, with minimal bias in the results. The Coastal region tended to have higher and consistent negative bias compared to the

[14] The interquartile range is the difference in values from the 25th to the 75th percentile, or a measure of the breadth of individual customer bias.

other regions. This lends credence to the idea that hotter weather – which tends to be more pronounced inland – was a significant factor in these methods.

Figure 14: Residential Electric Model Results by CEC Climate Zone[15]



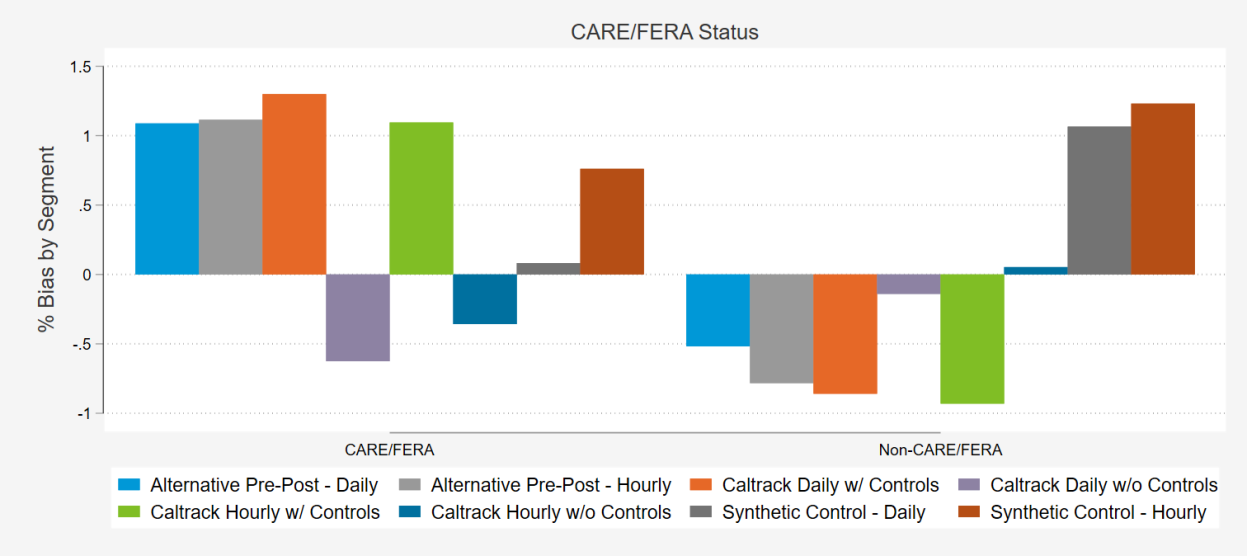
Low income customers tended to have slightly higher bias across most models compared to non-CARE rates. The reason for this is not necessarily clear but may have to do with smaller dwelling sizes or smaller air conditioning units associated with this customer type. Note that these results are reported on a percentage basis. Smaller premises with lower electricity consumption are more represented among customers with low income rates, and therefore the denominator of the total error will be smaller which leads to higher percent bias measures.

Key Finding

Synthetic controls may perform well but are highly sensitive to the choice of segmentation used

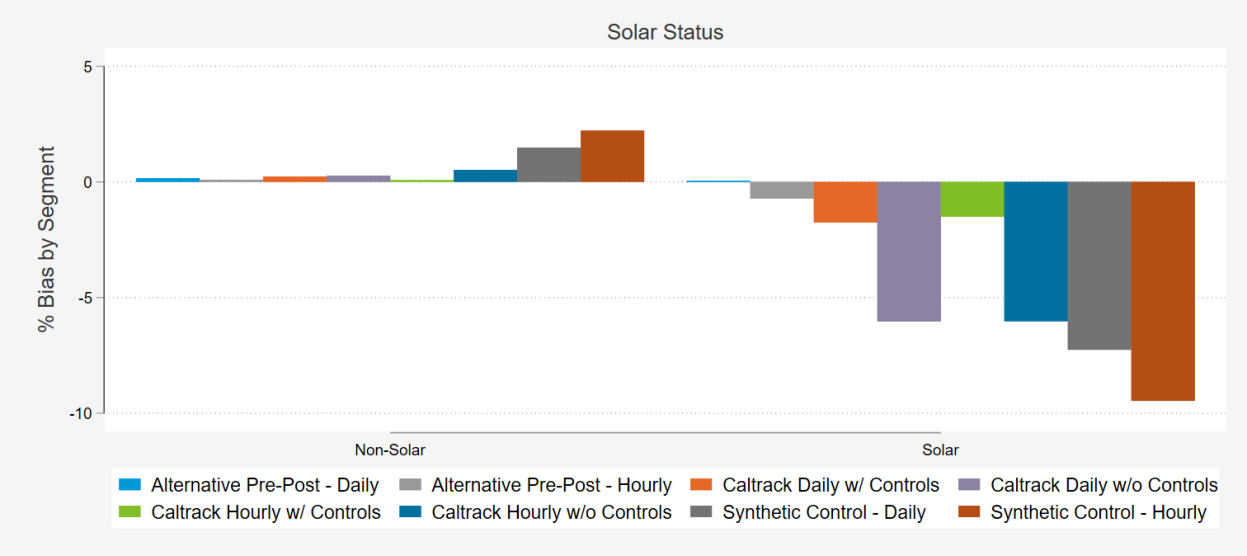
[15] Coastal included climate zones 1-5, Inland North included climate zones 11 and 12, and Inland South included climate zone 13

Figure 15: Residential Electric Model Results by Low Income Status



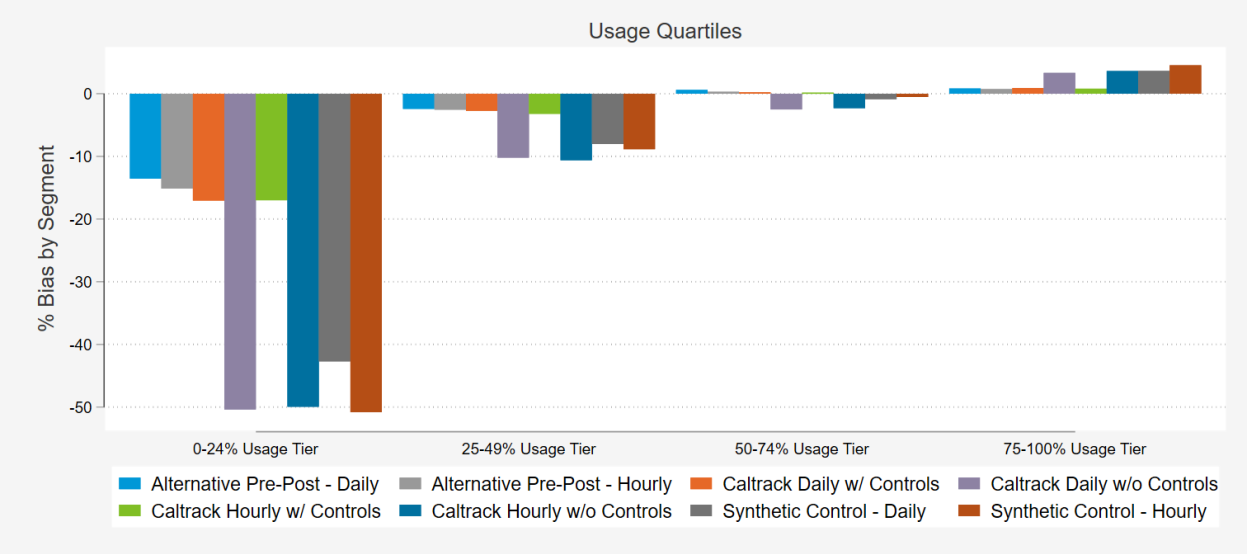
Savings estimates for residential customers with rooftop solar, comprising about 20% of the total number of residential EE program participants, had much higher bias across all models compared to non-solar customers. The analysis did account for solar customers; matching solar customers to solar customers in the comparison groups and removing customers who installed solar during the post-installation period. Nevertheless, these procedures do not seem to have controlled for all the ways in which installing and using solar energy affects customer consumption. It is important to note that the percent bias – a measure that requires dividing by the total consumption in each segment, may not appropriately capture the magnitude of errors for solar customers, as by dint of their rooftop solar status, the denominator in this ratio is smaller than non-solar customers. This magnifies the percent errors seen in this segment.

Figure 16: Residential Electric Model Results by Solar Status



The accuracy of each method for different sizes of customers shows a clear trend. All methods tended to underestimate loads for low users and overestimate loads for high users. While these groups are roughly equal in size, an upward bias among large consumers is more concerning because a 5% upward bias for a large customer can overcome a large negative bias among small customers on an absolute kWh consumption basis and result in an overestimation of residential electric program savings. Note that these results are reported on a percentage basis. In smaller premises with lower electricity consumption, the denominator of the total error will be smaller, leading to higher percent bias measures. In general, the average customer size in the smallest tier is between 10% to 20% of the customers in the largest tier.

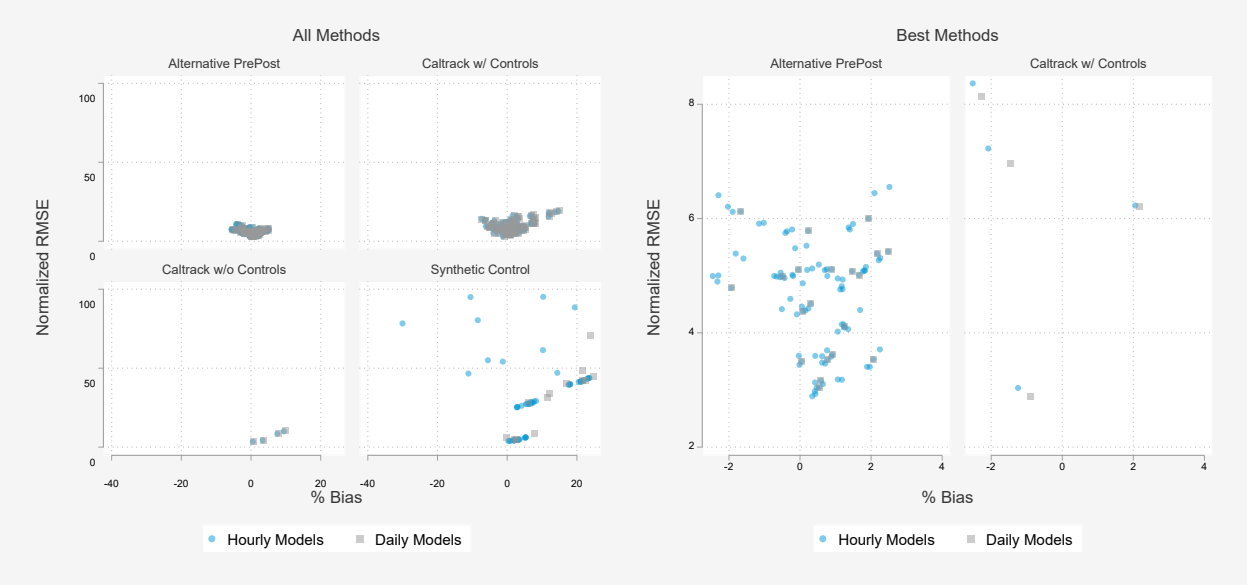
Figure 17: Residential Electric Model Results by Customer Size



4.3 ACCURACY ASSESSMENT RESULTS – COMMERCIAL ELECTRIC

Commercial electric accuracy results are shown in Figure 18. Unlike the residential electric results, the overall trend of bias is more balanced; with some models skewing upwards and some skewing downwards. This is likely because commercial customers, in general, are less weather sensitive than residential customers. Nevertheless, among the best models, there was a slight upward bias in most of the results. Among the winning models were the Alternative Pre-Post Hourly and Daily models.

Figure 18: Overall Accuracy and Precision for Commercial Electric Models



BEST MODELS FOR COMMERCIAL ELECTRIC CONSUMPTION

The best model among those tested for commercial electric consumption is the Alternative Pre-Post Hourly model with a matched control group. This model has a bias of 0.36% and a normalized RMSE of 3.03%. This model is shown in Table 7, along with the best models for all other frameworks, summarized across the bootstrapped random samples of 1,000, and averaged across all treatment periods. The next best model is the Alternative Pre-Post Daily model with the same matched control group and slightly different regression model. Bias and precision in the commercial sector are both slightly higher than the best models in the residential sector, likely a result of higher heterogeneity within the commercial sector. Segmentation that included bins of annual consumption and peak load performed best overall.

Table 7: Best Model by Framework for Commercial Electric Usage

Framework	Matching Method	Segmentation	Matching Characteristics	Model	% Bias	Normalized RMSE (%)
Alternative Pre-Post - Hourly	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 3-Hour Moving Avg Temp Spline	0.36	3.03
Alternative Pre-Post - Daily	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	0.54	3.11
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	1.32	4.45
Synthetic Control - Hourly	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)	N/A	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	2.62	3.73
CalTRACK Daily w/ Controls	Euclidian	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	CalTRACK Daily	1.60	4.47
CalTRACK Hourly w/o Controls	Participants Only	N/A	N/A	CalTRACK Hourly	5.35	6.10
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	5.65	6.39
Synthetic Control - Daily	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	3.34	4.96

Best Models as a Function of Pandemic Exposure and Sample Size

Table 8 and Table 9 show results for the top three overall models for the commercial sector, along with results for the best performing models in each framework. The best overall models are highlighted in blue. The results in Table 8 are shown for an average across all sample sizes (from 5 customers to 1000) for each treatment period. Treatment of the pseudo participants was randomly assigned by quarter of the year, with pseudo-treatment beginning in January of 2019. The results in this table therefore represent different levels of exposure to pandemic conditions in the full year post-treatment. Customers treated in January-March of 2019 have essentially no exposure to pandemic conditions as their year of post-treatment runs through March of 2020 at the latest. Customers with 75% COVID exposure had pseudo-treatment starting in October-December of 2019, meaning that their post-treatment period completed by December 2020 at the latest. The top three models relied on Alternative Pre-Post Hourly Models with similar matched control groups. The bias of each model varied with exposure to pandemic conditions, but with no clear trend. This result might be due to the interaction between changing COVID conditions and seasonal business schedules which left different holidays and busy seasons with different restrictions compared to the prior year.

Table 9 shows the performance of each method as a function of sample aggregation. These results are averaged across COVID exposure quarters. As expected, both bias and precision improved with higher levels of aggregation, however the RMSE of commercial electric consumption remains consistently higher than that of the residential sector. The improvement in model bias was limited for the best performing models – likely because event results for small aggregations of customers was quite good.

Winning segmentation strategies relied on matching within climate zone, solar status, bins of annual consumption and peak load. In all matching methods, the best matching variable was percentiles of annual consumption. There was a mix of matching methods represented amongst the best options, however. This suggests that the matching method is less important than the segmentation strategy for model performance.

Key Finding

When constructing a matched control group, the choice of segmentation and matching characteristics matter more than the matching method

Table 8: Commercial Electric Models Results for Different Levels of Pandemic Exposure

Framework	Matching Method	Segmentation	Matching Characteristics	Model	Value	% COVID Effect			
						0%	25%	50%	75%
Alternative Pre-Post - Hourly	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	% Bias	0.15	1.23	-0.56	0.21
					Norm. RMSE	12.39	11.52	19.69	21.57
Alternative Pre-Post - Hourly	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 3-Hour Moving Avg Temp Spline	% Bias	0.25	1.20	-0.39	0.67
					Norm. RMSE	12.24	11.21	19.90	20.91
Alternative Pre-Post - Hourly	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Temp Spline	% Bias	0.80	1.55	-0.51	0.46
					Norm. RMSE	12.75	12.14	19.76	21.34
Alternative Pre-Post - Daily	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	0.22	1.33	0.19	0.85
					Norm. RMSE	12.38	11.45	21.76	21.41
CalTRACK Daily w/ Controls	Euclidian	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	CalTRACK Daily	% Bias	1.80	1.61	0.76	2.39
					Norm. RMSE	15.76	18.40	28.07	28.54
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	% Bias	1.23	4.34	7.51	10.05
					Norm. RMSE	17.85	10.44	18.41	16.88
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	% Bias	1.74	1.34	0.47	2.22
					Norm. RMSE	15.46	18.56	29.01	28.83
CalTRACK Hourly w/o Controls	Participants Only	N/A	N/A	CalTRACK Hourly	% Bias	1.11	3.97	7.36	9.71
					Norm. RMSE	15.64	10.51	17.74	16.43
Synthetic Control - Daily	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	1.03	7.98	3.30	2.68
					Norm. RMSE	26.22	17.35	23.67	17.88
Synthetic Control - Hourly	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)	N/A	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	% Bias	1.21	5.52	1.71	3.46
					Norm. RMSE	17.38	13.01	14.25	12.59

Table 9: Commercial Electric Models Results for Different Sample Sizes

Framework	Matching Method	Segmentation	Matching Characteristics	Model	Value	Sample Aggregation						
						5	10	25	50	100	500	1000
Alternative Pre-Post - Hourly	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	% Bias	-0.04	0.60	0.82	-0.12	0.15	0.19	0.20
					Norm. RMSE	35.48	26.67	20.70	13.55	10.24	4.37	3.03
Alternative Pre-Post - Hourly	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 3-Hour Moving Avg Temp Spline	% Bias	-0.02	0.82	1.05	0.12	0.33	0.36	0.36
					Norm. RMSE	34.66	26.67	20.28	13.45	10.06	4.31	3.03
Alternative Pre-Post - Hourly	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Temp Spline	% Bias	0.45	0.96	1.14	0.14	0.43	0.44	0.44
					Norm. RMSE	36.81	27.05	20.49	13.55	10.15	4.36	3.06
Alternative Pre-Post - Daily	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	0.30	1.16	1.06	0.38	0.51	0.58	0.54
					Norm. RMSE	37.78	27.23	20.74	13.66	10.33	4.38	3.11
CalTRACK Daily w/ Controls	Euclidian	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	CalTRACK Daily	% Bias	1.09	1.90	3.48	0.96	1.19	1.28	1.60
					Norm. RMSE	43.80	46.91	22.64	20.86	13.67	6.50	4.47
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	% Bias	6.52	6.50	5.07	5.88	5.37	5.48	5.65
					Norm. RMSE	32.86	23.37	18.11	13.65	10.02	6.86	6.39
CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	% Bias	1.03	1.72	3.38	0.63	1.00	1.03	1.32
					Norm. RMSE	43.15	48.62	23.00	21.00	13.95	6.58	4.45
CalTRACK Hourly w/o Controls	Participants Only	N/A	N/A	CalTRACK Hourly	% Bias	6.58	6.13	4.85	5.54	5.09	5.21	5.35
					Norm. RMSE	30.07	22.44	17.56	13.14	9.66	6.57	6.10
Synthetic Control - Daily	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	4.33	4.08	3.55	4.47	3.35	3.12	3.34
					Norm. RMSE	48.17	37.55	22.45	17.32	12.27	6.25	4.96
Synthetic Control - Hourly	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)	N/A	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	% Bias	3.66	4.03	2.33	3.07	2.54	2.55	2.62
					Norm. RMSE	31.63	22.03	17.39	11.93	8.85	4.60	3.73

Distribution of Individual Errors

Figure 19 shows the distribution of individual participant errors in a box plot for the best method for each framework. For each row, the blue box indicates the range of outcomes associated with the 25th-75th percentile, while the white line indicates the median value. The whiskers are 1.5 times the interquartile range.[16] To enhance readability, outliers have been omitted from this plot. The best method at the aggregate, bootstrapped level is the Alternative Pre-Post Hourly model with a matched control group. Unlike the residential electric results, the median value of most of these models has an upward bias, though with a comparable distribution to that of the residential results. With the exception of the Synthetic Control Daily model, the error bands are narrower in all methods that do not rely on a differencing strategy. This suggests that while comparison groups perform better in aggregate, individual customers may not be matched to their optimal matched comparator customer, leading to error.

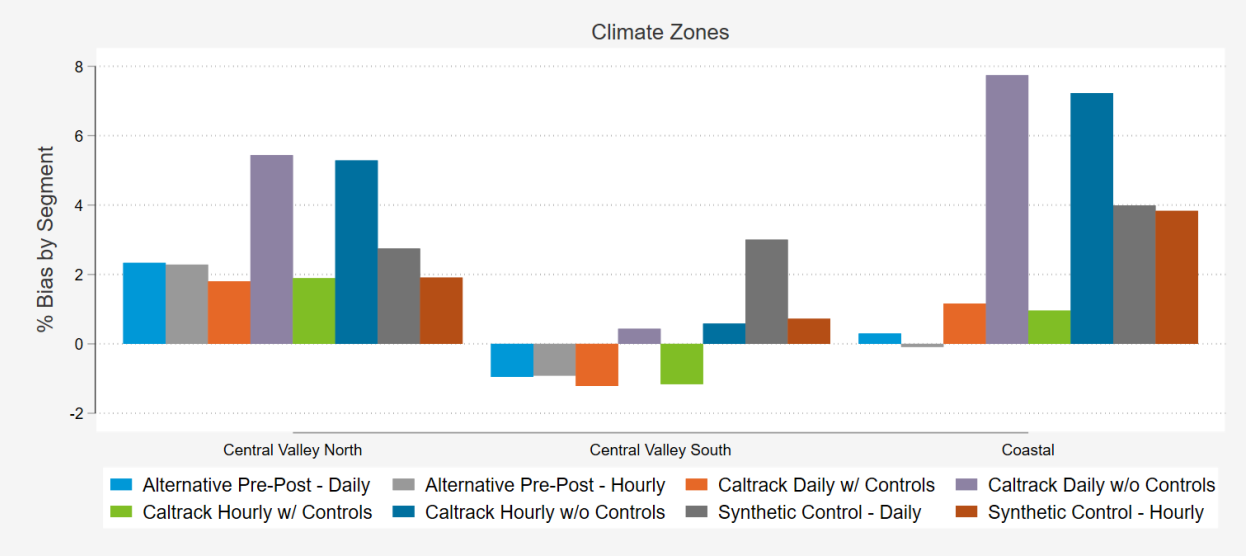
Figure 19: Commercial Electric Individual Error Distribution

Best Model Results for Segments of Interest

This section reviews results by key customer segments, including groups of CEC climate zones (Figure 14), rate (Figure 21) and solar status (Figure 22). Results by climate zone are more mixed than what was observed among residential customers, with positive bias in both the North Central Valley and Coastal areas. Again, CalTRACK models without any comparison groups generally had the highest bias of all the best models tested.

[16] The interquartile range is the difference in values from the 25th to the 75th percentile.

Figure 20: Commercial Electric Model Results by Climate Zone[17]

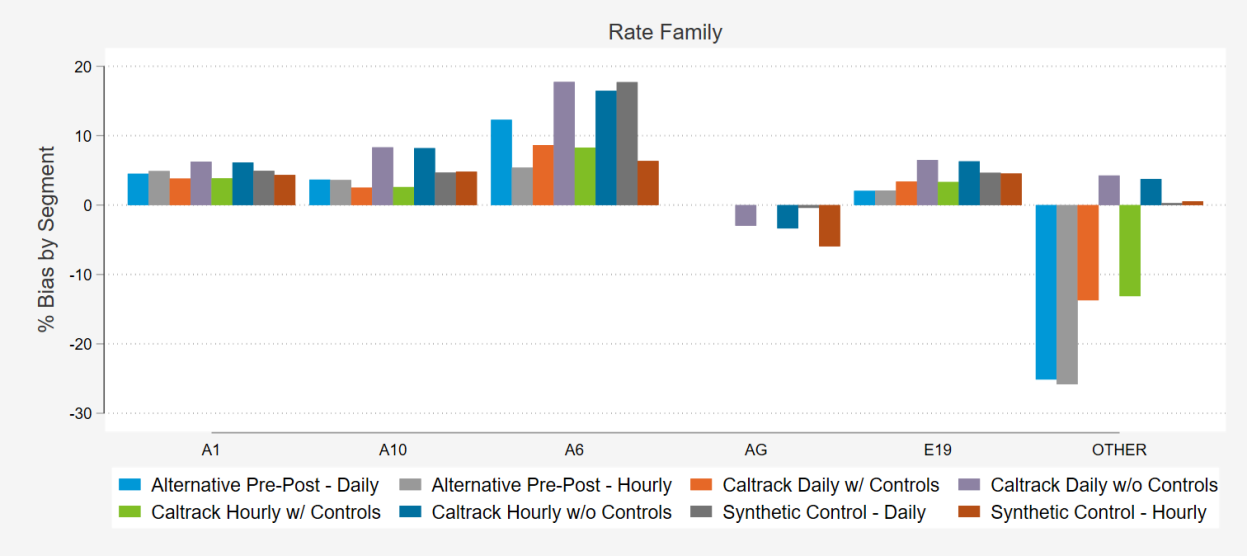


The bias for the best models in each framework were generally higher for A1, A10, A6, and E19 rates[18]. Together, these rates comprised about 90% of the sampled commercial customers, so it perhaps is not surprising that they look more consistent. The differences between the A1/A10/E19 and A6 groups may be related to the relative uptake of different rates within certain industries that have different sizes, more or less exposure to pandemic-related economic changes, or weather sensitivity.

[17] Coastal included climate zones 1-5, Inland North included climate zones 11 and 12, and Inland South included climate zone 13

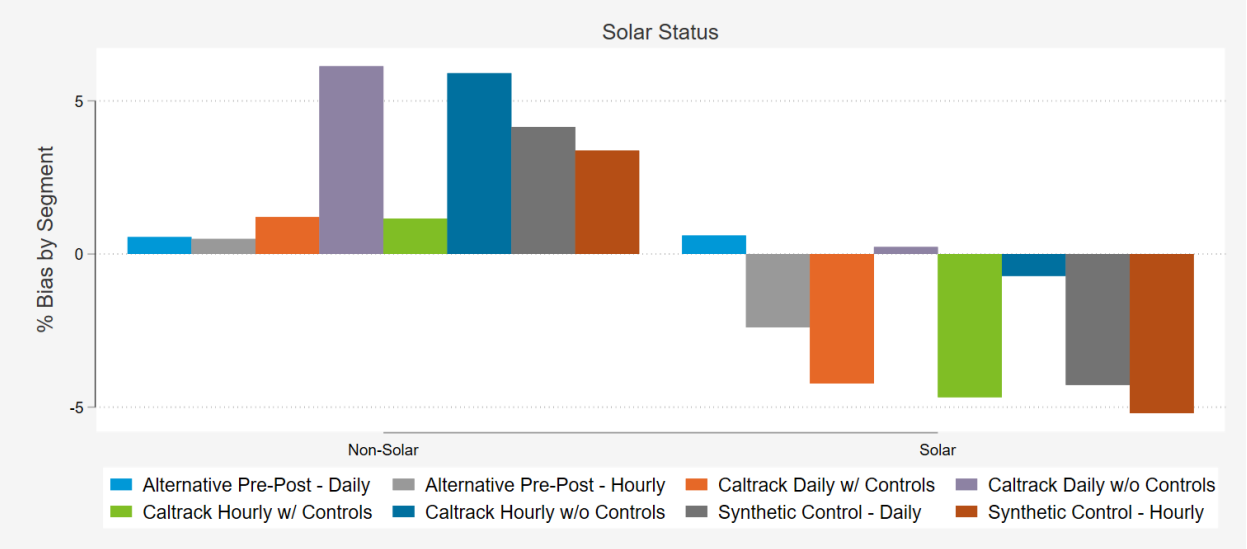
[18] Rate A1 represents small commercial customers (below 75kW maximum demand), A10 is the rate for medium commercial customers (less than 500kW maximum demand), A6 is the small commercial TOU rate, and E19 is the large customer tariff (greater than 500kW maximum demand). More detail about each rate can be found on PG&E's website at https://www.pge.com/tariffs/index.page

Figure 21: Commercial Electric Model Results by Rate



The bias of the best models for commercial solar customers varied quite a bit depending on which framework was used, with the Alternative Pre-Post Hourly model performing best across both segments. Solar customers only made up about 6% of the sampled participants, however, so this result should be interpreted with some caution. Again, the CalTRACK models without controls had consistently high bias. It is important to note that the percent bias – a measure that requires dividing by the total consumption in each segment, may not appropriately capture the magnitude of errors for solar customers, as by dint of their rooftop solar status, the denominator in this ratio is smaller than non-solar customers. This magnifies the percent errors seen in this segment.

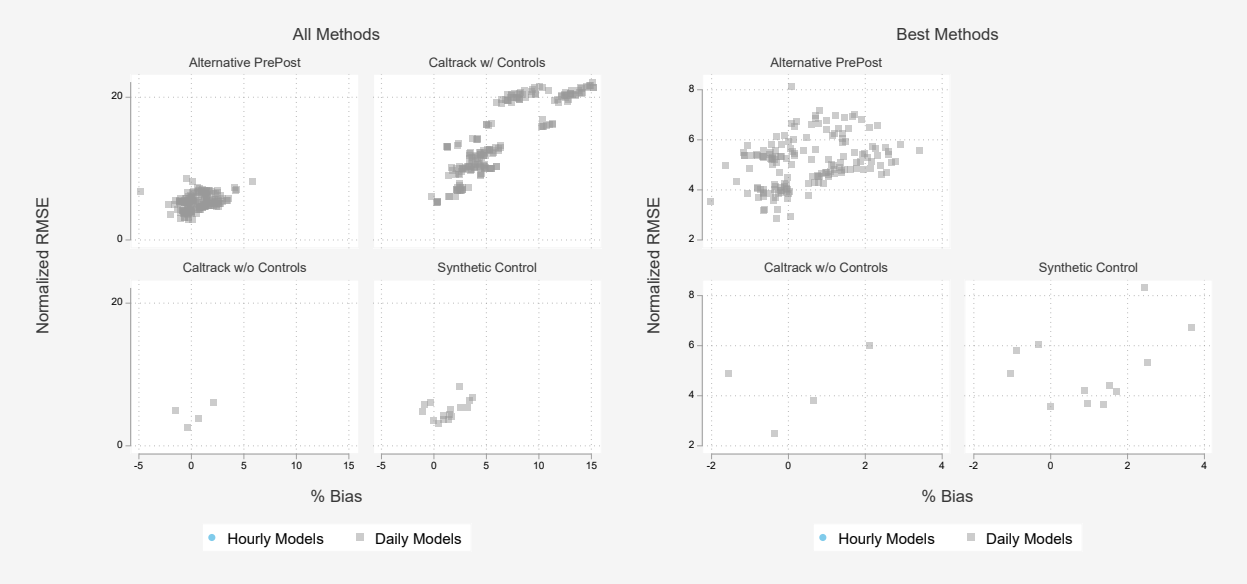
Figure 22: Commercial Electric Model Results by Solar Status



4.4 ACCURACY ASSESSMENT RESULTS – RESIDENTIAL GAS

Residential gas accuracy results are shown in Figure 23. Results again showed an upward bias trend, especially for the CalTRACK models with control groups. The overwhelming percent of best models, chosen as described for the residential electric participants, are the Alternative Pre-Post models. The winning model overall is the CalTRACK model without a control group, however. As shown in the figure below, the four different treatment quarters all displayed reasonable bias and precision, with some showing slightly higher bias and some lower; this seems to have averaged out to a low overall number in the final average. Many Alternative Pre-Post models also seem to perform well. The timing of gas consumption for residential customers, which is mainly used for heating, may interact with the effects of the pandemic as well as changes in holiday schedules compared to 2019.

Figure 23: Overall Accuracy and Precision for Residential Gas Models



BEST MODELS

The best model among those tested for residential gas consumption is the CalTRACK Daily model without a matched control group. This model has a bias of 0.34% and a normalized RMSE of 3.23%. Note that gas data is most commonly recorded in daily intervals only, therefore no hourly models were tested for this segment. This model is shown in Table 10 along with the best models for all other frameworks, for a sample size of 1,000, averaged across all treatment periods. This result is somewhat surprising and may be because many of the CalTRACK gas models simply use the daily average consumption as the counterfactual[19]. Because of the relatively temperate climate in California, using the pre-treatment average consumption value may act as a sufficiently good counterfactual. As noted in the prior sections, the frameworks using a differencing approach show higher individual-customer volatility in impacts; in this context it is possible that pseudo-participants are being matched to controls

[19] See section 3.4 of the CalTRACK methods for more detail

that are using a different counterfactual model (one that includes HDD and CDD compared to the intercept-only model for example). This could explain both the good performance of the CalTRACK without control group and the relatively worse performance of the differencing methods. More investigation of this result is needed. The next best model is the Daily Synthetic Control model with predictor variables comprised of profiles of average consumption segmented by climate zone, annual consumption bins and monthly consumption profile bins. There is not much similarity between segmentation methods in these models. It is interesting to note that heating weather sensitivity was not a winning matching characteristic; it was deliberately included to capture customers with electric heating. Ranking customers on their annual gas consumption proved to be a better option for improved performance.

Table 10: Best Model by Framework for Residential Gas Usage

Framework	Matching Method	Segmentation	Matching Characteristics	Model	% Bias	Normalized RMSE (%)
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	0.34	3.23
Alternative Pre-Post - Daily	Propensity	Climate Zone, Annual Consumption (4 Bins) and Low Income Flag	Percentile ranking of annual consumption	Hour, DOW, Month, Daily Avg Temp Spline	0.89	3.58
Synthetic Control - Daily	N/A	Climate Zone, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	0.66	2.97
CalTRACK Daily w/ Controls	Propensity	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	Percentile ranking of annual consumption	CalTRACK Daily	2.63	6.94

Best Models as a Function of Pandemic Exposure and Sample Size

Table 11 and Table 12 show results for the top three overall models for this sector, along with results for the best performing models in each framework. The best overall models are highlighted in blue. The results in Table 11 are shown for an average across all sample sizes (from 5 customers to 1000) for each treatment period. Treatment of the pseudo participants was randomly assigned by quarter of the year, with pseudo-treatment beginning in January of 2019. The results in this table therefore represent different levels of exposure to pandemic conditions in the full year post-treatment. Customers treated in January-March of 2019 have essentially no exposure to pandemic conditions as their year of post-treatment runs through March of 2020 at the latest. Customers with 75% COVID exposure had pseudo-treatment starting in October-December of 2019, meaning that their post-treatment period completed by December 2020 at the latest. The top three models relied on Alternative Pre-Post Hourly Models or a CalTRACK Daily model without a control. The bias of each model varied with exposure to pandemic conditions, with slightly positive bias in pre-pandemic simulations compared to minimal or negative

bias with exposure to the pandemic. This is in contrast to the residential electric results which showed the opposite trend. This supports the evidence that differences in weather between years cannot be fully accounted for by any of these models for residential customers; a cooler post-treatment period would have less air conditioning and more electric heating on average.

Table 12 shows the performance of each method as a function of sample aggregation. These results are averaged across COVID exposure quarters. As expected, both bias and precision improved with higher levels of aggregation.

There does not appear to be any specific winning segmentation strategy among residential gas models. Matching methods relied on matching customers within segments according to similar annual consumption percentiles using propensity score matching, however the segmentation methods themselves varied.

Table 11: Residential Gas Models Results for Different Levels of Pandemic Exposure

Framework	Matching Method	Segmentation	Matching Characteristics	Model	Value	% COVID Effect			
						0%	25%	50%	75%
Alternative Pre-Post - Daily	Propensity	Climate Zone, Annual Consumption (4 Bins) and Low Income Flag	Percentile ranking of annual consumption	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	2.53	0.97	0.33	0.39
					Norm. RMSE	13.26	15.71	13.09	18.48
Alternative Pre-Post - Daily	Propensity	Climate Zone, Annual Consumption (4 Bins)	Percentile ranking of annual consumption	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	1.31	0.60	-0.62	0.07
					Norm. RMSE	13.00	15.41	11.12	16.39
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	% Bias	0.57	1.96	-0.36	-1.52
					Norm. RMSE	11.45	14.88	7.66	12.14
CalTRACK Daily w/ Controls	Propensity	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	Percentile ranking of annual consumption	CalTRACK Daily	% Bias	5.30	2.58	3.44	0.73
					Norm. RMSE	49.33	41.30	25.74	29.71
Synthetic Control - Daily	N/A	Climate Zone, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	1.07	0.95	1.80	-1.49
					Norm. RMSE	10.73	12.99	12.82	12.13

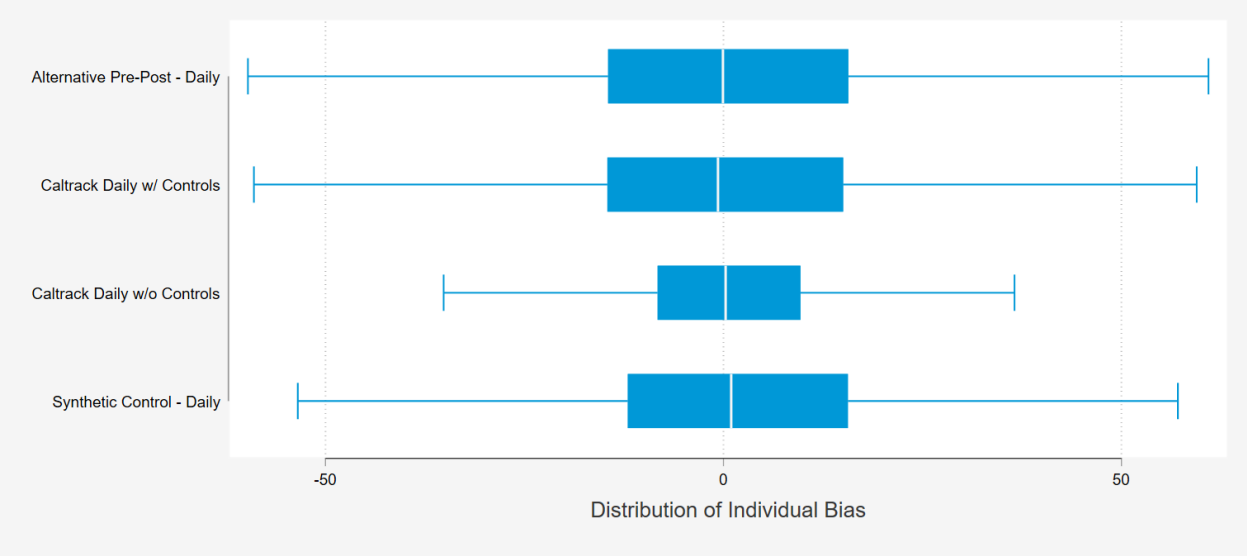
Table 12: Residential Gas Models Results for Different Sample Sizes

Framework	Matching Method	Segmentation	Matching Characteristics	Model	Value	Sample Aggregation						
						5	10	25	50	100	500	1000
Alternative Pre-Post - Daily	Propensity	Climate Zone, Annual Consumption (4 Bins) and Low Income Flag	Percentile ranking of annual consumption	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	1.29	1.79	0.69	1.42	0.62	0.69	0.89
					Norm. RMSE	33.00	26.61	14.80	14.29	8.97	4.70	3.58
Alternative Pre-Post - Daily	Propensity	Climate Zone, Annual Consumption (4 Bins)	Percentile ranking of annual consumption	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	-0.19	0.40	0.17	0.90	0.35	0.30	0.44
					Norm. RMSE	26.75	24.85	14.60	14.31	9.54	4.59	3.22
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	% Bias	0.15	-0.81	0.06	1.00	0.16	0.21	0.34
					Norm. RMSE	24.35	17.01	10.46	13.85	7.51	4.31	3.23
CalTRACK Daily w/ Controls	Propensity	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	Percentile ranking of annual consumption	CalTRACK Daily	% Bias	8.23	-0.36	2.28	3.72	2.20	2.39	2.63
					Norm. RMSE	107.27	56.38	29.16	30.43	16.63	8.84	6.94
Synthetic Control - Daily	N/A	Climate Zone, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	0.14	0.04	0.40	1.03	1.05	0.75	0.66
					Norm. RMSE	22.42	25.54	9.80	11.91	8.46	4.10	2.97

Distribution of Individual Errors

Figure 24 shows the distribution of individual participant errors in a box plot for the best method for each framework. For each row, the blue box indicates the range of outcomes associated with the 25th-75th percentile, while the white line indicates the median value. The whiskers are 1.5 times the interquartile range[20]. To enhance readability, outliers have been omitted from this plot. The best method at the aggregate, bootstrapped level is the CalTRACK Daily model without a matched control group. In general, however, the distributions of all of the best frameworks tend to be unbiased when looked at in this manner, as the box plot omits outliers from the graphic. The CalTRACK Daily model also has a narrower distribution of individual customer errors, meaning fewer customers have large upward or downward bias in their results.

Figure 24: Residential Gas Individual Error Distribution

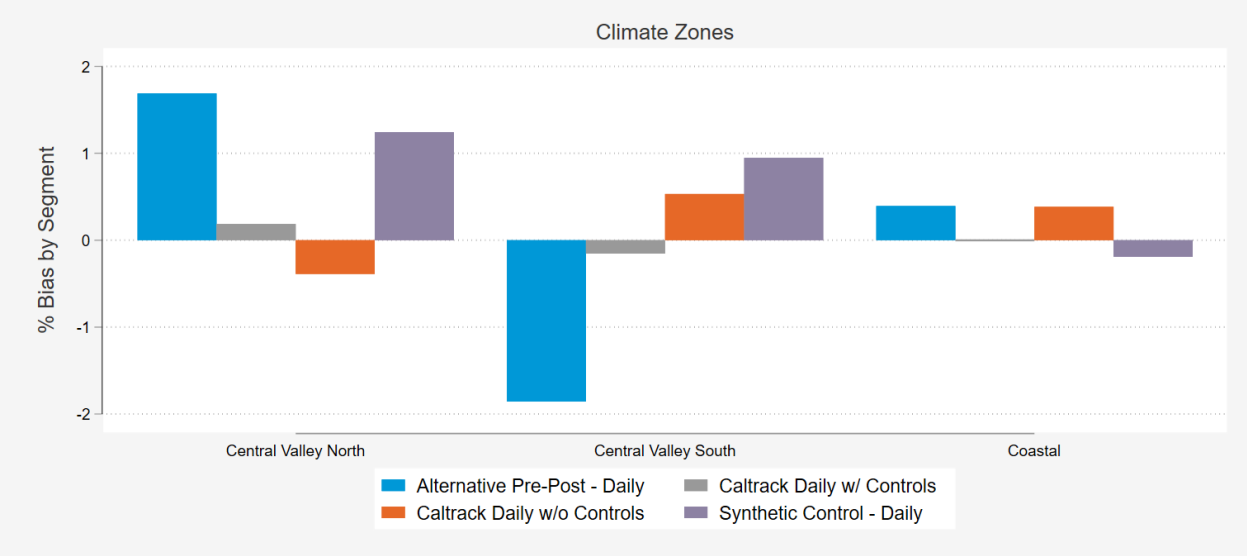


Best Model Results for Segments of Interest

This section reviews results by key customer segments, including groups of CEC climate zones (Figure 25), low income status (Figure 26), and customer size (Figure 27). Results by climate zone are more mixed than what was observed among residential electric customers, both between climate zone groups and between frameworks. Generally speaking, however, the magnitude of all of these values is quite low compared to what was observed in the electric sector, meaning that all models were quite accurate.

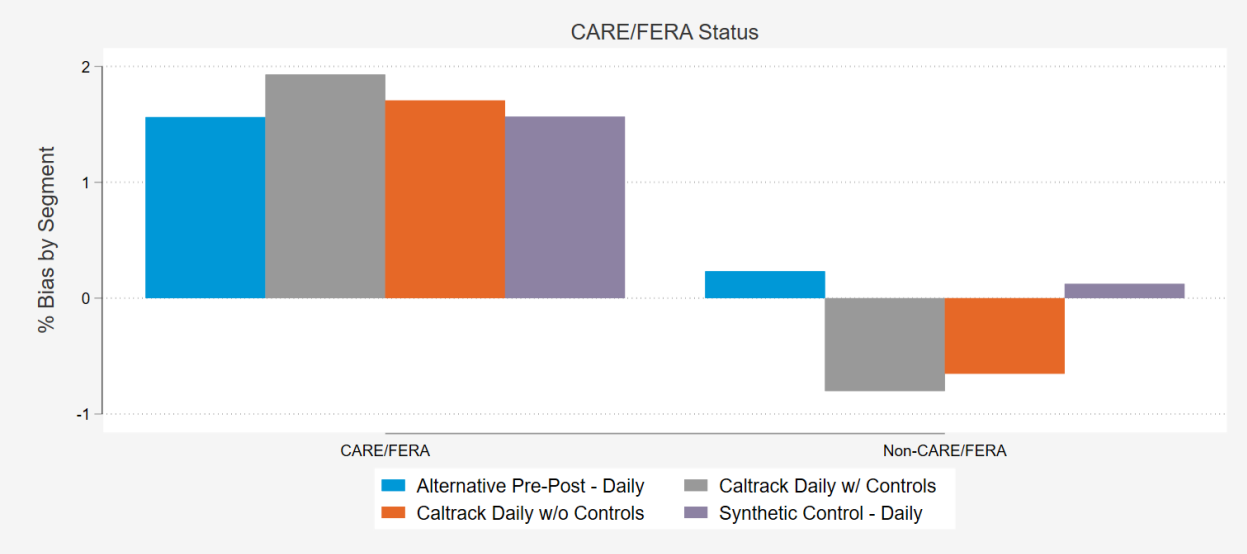
[20] The interquartile range is the difference in values from the 25th to the 75th percentile.

Figure 25: Residential Gas Model Results by Climate Zone[21]



In contrast to the residential electric results, higher bias was observed in the low-income group for gas customers. These results were relatively consistent among all methods tested. The timing of pandemic-related shutdowns, work-from-home schedules, and gas consumption may mean that low-income households may have used less gas than these models predicted.

Figure 26: Residential Gas Model Results by Low Income Status

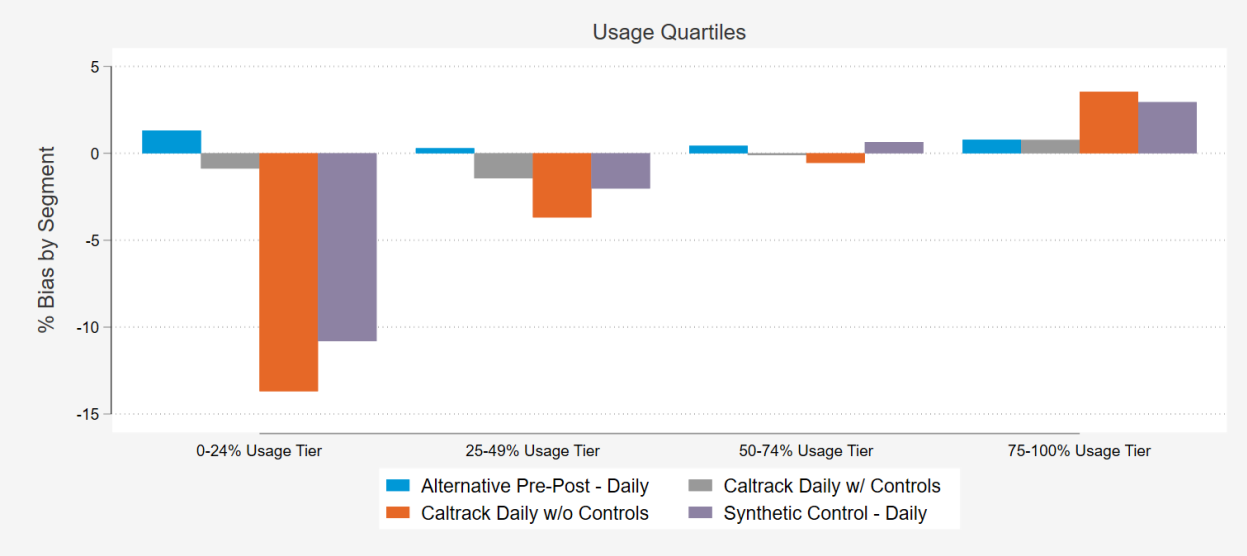


The trend of model bias across customer sizes does appear to be similar to that of the residential electric customers, with smaller household generally having negative bias for the best models while

[21] [21] Coastal included climate zones 1-5, Inland North included climate zones11 and 12, and Inland South included climate zone 13

larger customers have higher bias. This is most pronounced in the CalTRACK without matched control model and the Synthetic Control model.

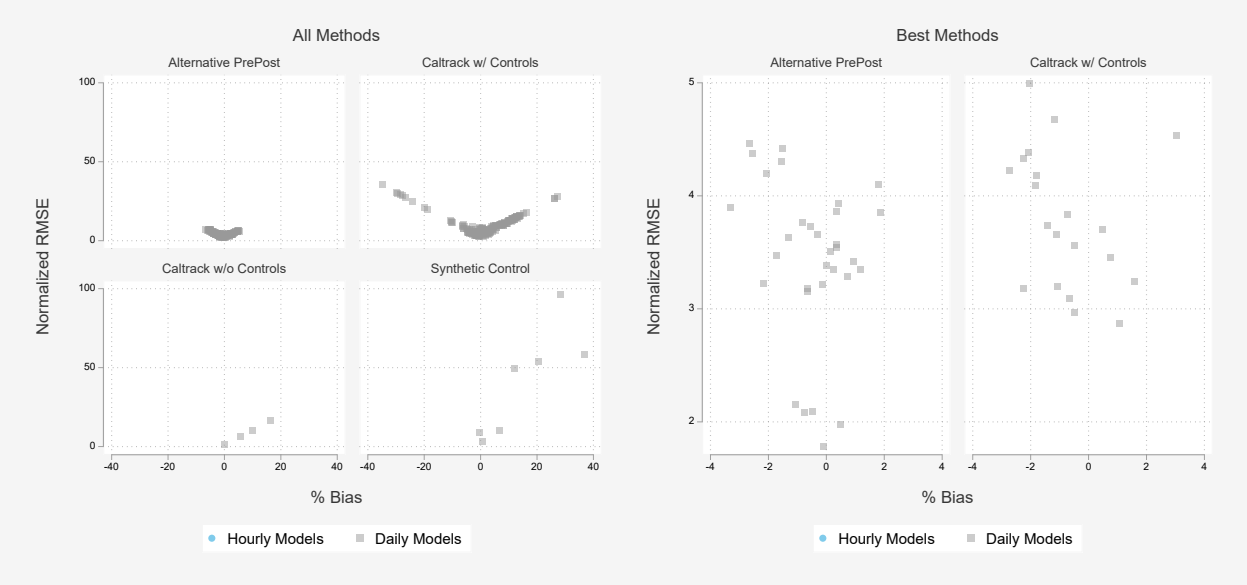
Figure 27: Residential Gas Model Results by Customer Size



4.5 ACCURACY ASSESSMENT RESULTS – COMMERCIAL GAS

Commercial electric accuracy results are shown in Figure 28. Unlike the residential electric results, the overall trend of bias is more balanced; with some models skewing upwards and some skewing downwards. This is likely because commercial customers, in general, are less weather sensitive than residential customers. While some CalTRACK Daily models with matched controls had substantial errors in this test, others were among the best models available. As with all the other sectors discussed in this report, it is clear that the choice of how a matched control group is constructed can matter a great deal more than the method of estimating a customer baseline.

Figure 28: Overall Accuracy and Precision for Commercial Gas Models



Best Models

The best model among those tested for commercial gas consumption is the Alternative Pre-Post Daily model with a matched control group. This model has a bias of -0.60% and a normalized RMSE of 1.94%. Note that gas data is most commonly recorded in daily intervals only, therefore no hourly models were tested for this segment. This model is shown in

Table 13Table 4, along with the best models for all other frameworks, for a sample size of 1,000, averaged across all treatment periods. The next best model is the CalTRACK Daily model with a similar matched control group. The two best models significantly outperformed the other two candidates. This is likely driven by the matched control group selection in this sector. It is interesting to note that heating weather sensitivity was not a winning matching characteristic; it was deliberately included to capture customers with electric heating. Ranking customers on their annual gas consumption proved to be a better option for improved performance.

Table 13: Best Model by Framework for Commercial Gas Usage

Framework	Matching Method	Segmentation	Matching Characteristics	Model	% Bias	Normalized RMSE (%)
Alternative Pre-Post - Daily	Propensity	Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	Hour, DOW, Month, Daily Avg Temp Spline	-0.60	1.94
CalTRACK Daily w/ Controls	Euclidian	Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	CalTRACK Daily	0.55	2.37
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	7.78	8.16
Synthetic Control - Daily	N/A	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	10.01	14.40

Best Models as a Function of Pandemic Exposure and Sample Size

Table 14 and Table 15 show results for the top three overall models for this sector, along with results for the best performing models in each framework. The best overall models are highlighted in blue. The results in Table 14 are shown for an average across all sample sizes (from 5 customers to 1000) for each treatment period. Treatment of the pseudo participants was randomly assigned by quarter of the year, with pseudo-treatment beginning in January of 2019. The results in this table therefore represent different levels of exposure to pandemic conditions in the full year post-treatment. Customers treated in January-March of 2019 have essentially no exposure to pandemic conditions as their year of post-treatment runs through March of 2020 at the latest. Customers with 75% COVID exposure had pseudo-treatment starting in October-December of 2019, meaning that their post-treatment period completed by December 2020 at the latest. The top three models relied on Alternative Pre-Post Daily Models with matched control groups. Similar to the commercial electric sector, the bias of each model varied with exposure to pandemic conditions, but with no clear trend. This might be due to the interaction between changing COVID conditions and seasonal business schedules which left different holidays and busy seasons with different restrictions compared to the prior year.

Table 15 shows the performance of each method as a function of sample aggregation. These results are averaged across COVID exposure quarters. As expected, both bias and precision improved with higher levels of aggregation, however the RMSE of commercial electric consumption remains consistently higher than that of the residential sector. Compared to the commercial electric sector, however, precision improved at a faster rate as sample sizes increased.

Winning segmentation strategies relied on matching within climate zone, bins of annual consumption and similar monthly consumption profiles, with either the inclusion of the first digit or first two digits of the premise's NAICS code. Inclusion of the bins of 12-month consumption profiles in winning methods makes sense, as that variable reflects the annual shape of gas consumption for commercial customers. Generally, summer and winter consumption percentiles were the winning matching strategies.

Table 14: Commercial Gas Models Results for Different Levels of Pandemic Exposure

Framework	Matching Method	Segmentation	Matching Characteristics	Model	Value	% COVID Effect			
						0%	25%	50%	75%
Alternative Pre-Post - Daily	Euclidian	Climate Zone, 1st and 2nd Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	-1.05	-1.33	-0.60	2.09
					Norm. RMSE	9.54	15.81	13.16	15.51
Alternative Pre-Post - Daily	Propensity	Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	-0.33	-1.69	-0.48	1.14
					Norm. RMSE	8.91	18.68	13.42	17.19
Alternative Pre-Post - Daily	Propensity	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	Bins of Heating Weather Sensitivity (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	0.36	-0.86	0.20	-1.40
					Norm. RMSE	9.56	19.81	16.60	20.51
CalTRACK Daily w/ Controls	Euclidian	Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	CalTRACK Daily	% Bias	0.83	0.03	1.48	-0.73
					Norm. RMSE	16.03	14.53	12.93	18.38
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	% Bias	0.20	4.58	10.36	16.89
					Norm. RMSE	7.66	17.85	17.01	23.11
Synthetic Control - Daily	N/A	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	0.94	-1.85	36.73	8.47
					Norm. RMSE	14.36	39.14	205.58	39.82

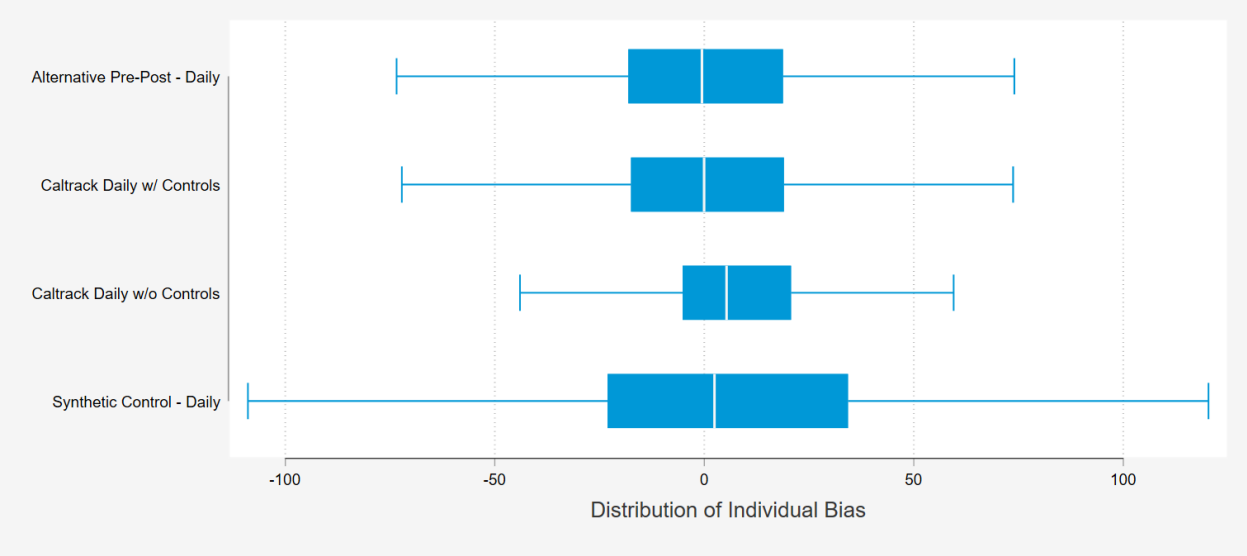
Table 15: Commercial Gas Models Results for Different Sample Sizes

Framework	Matching Method	Segmentation	Matching Characteristics	Model	Value	Sample Aggregation						
						5	10	25	50	100	500	1000
Alternative Pre-Post - Daily	Euclidian	Climate Zone, 1st and 2nd Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	-0.05	0.69	-0.34	-0.55	-0.49	-0.24	-0.59
					Norm. RMSE	36.42	20.08	15.15	10.71	7.20	2.99	1.99
Alternative Pre-Post - Daily	Propensity	Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	-0.60	1.36	-0.74	-0.88	-0.67	-0.24	-0.60
					Norm. RMSE	39.63	22.50	15.67	11.52	7.54	3.06	1.94
Alternative Pre-Post - Daily	Propensity	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	Bins of Heating Weather Sensitivity (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	-1.19	0.72	-0.88	-0.54	-0.37	-0.22	-0.50
					Norm. RMSE	46.00	26.76	17.06	12.69	8.22	3.37	2.24
CalTRACK Daily w/ Controls	Euclidian	Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	CalTRACK Daily	% Bias	1.75	1.02	-0.24	-0.07	-0.49	0.28	0.55
					Norm. RMSE	35.87	27.45	17.59	12.88	8.58	3.52	2.37
CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	% Bias	7.20	9.21	8.47	7.66	7.77	7.96	7.78
					Norm. RMSE	34.53	24.05	15.75	12.98	10.73	8.65	8.16
Synthetic Control - Daily	N/A	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	% Bias	14.45	8.79	8.59	11.37	13.30	11.02	10.01
					Norm. RMSE	178.59	120.33	79.91	67.28	42.54	20.05	14.40

Distribution of Individual Errors

Figure 29 shows the distribution of individual participant errors in a box plot for the best method for each framework. For each row, the blue box indicates the range of outcomes associated with the 25th-75th percentile, while the white line indicates the median value. The whiskers are 1.5 times the interquartile range[22]. To enhance readability, outliers have been omitted from this plot. The best method at the aggregate, bootstrapped level is the Alternative Pre-Post Daily model with a matched control group. Both the CalTRACK Daily and Alternative Pre-Post Daily models have very similar distributions of site-level error, with median errors very close to 0%. While the other two models both show some upward bias in their results, the CalTRACK Daily model without controls has an error distribution that is comparatively narrow.

Figure 29: Commercial Gas Individual Error Distribution

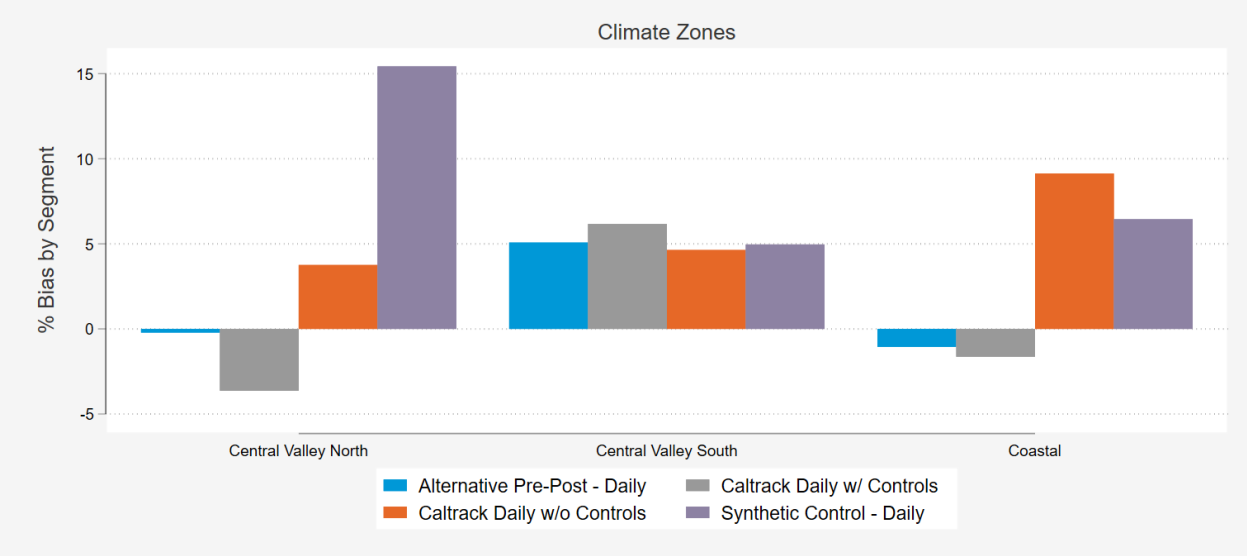


Best Model Results for Segments of Interest

This section reviews results by key customer segments, including groups of CEC climate zones (Figure 30), customer size (Figure 31). The results by climate zone are not necessarily conclusive and may have more to do with the distributions of industry in each region than weather. In general, the Alternative Pre-Post model had the least bias across the three groups of climate zones, while the best Synthetic Control model remained quite variable.

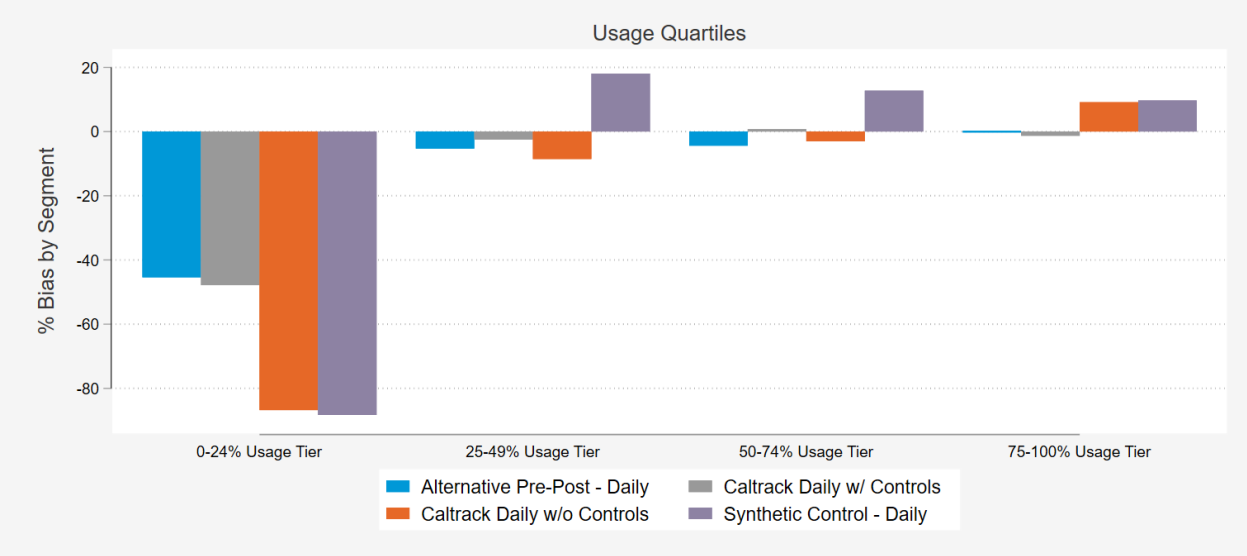
[22] The interquartile range is the difference in values from the 25th to the 75th percentile.

Figure 30: Commercial Gas Model Results by Climate Zone23



Results by customer size again follow a trend of smaller customers having negative bias. As the percent bias is calculated by dividing by the observed number of consumed therms, small denominators can mean that the magnitude of the bias on a percentage basis is quite large. Nevertheless, the Alternative Pre-Post and CalTRACK with a matched control group method performed the best, with lower absolute bias in most size groups.

Figure 31: Commercial Gas Model Results by Customer Size



[23] Coastal included climate zones 1-5, Inland North included climate zones 11 and 12, and Inland South included climate zone 13

4.6 ACCURACY OF GRANULAR PROFILES

A request was made of Demand Side Analytics, to test the accuracy and precision of two alternative methods of incorporating comparison groups into the existing CalTRACK methods. Due to the timing of the request, towards the end of the simulation period that generated the results for this report, a smaller-scale study was completed to ascertain whether these alternative methods would be viable options. The two models tested relied on granular profiles to be used in place of matched comparison groups, first as a synthetic control – simply adding the profile to the right hand side of the CalTRACK regression (CalTRACK + Synthetic Control model), and second as a CalTRACK counterfactual to difference out of individual participant CalTRACK counterfactual (Difference-in-Differences approach). The benefits of such models are straightforward:

- Profiles can be produced in a range of segmentations and aggregations
- Queries to sample non-participants and aggregate their consumption data can be automated and run on a regular basis
- Because aggregated data can be shared publicly, these profiles can be published online, and all interested parties can access without extensive security reviews

The steps to construct these methods are relatively straightforward. First, construct granular (8,760 or daily) profiles of average temperature and consumption for a sample of non-participants in defined segmentations of interest. To run the CalTRACK + Synthetic Control model, simply add the profile that matches the participant's characteristics as a right-hand-side variable to the existing CalTRACK model. For the Difference-in-Differences (DiD) approach, run CalTRACK on each: the participant consumption profile and the same matched granular profile. Difference the matched comparator impact, estimated from CalTRACK, from the participant's impact to get the DiD impact.

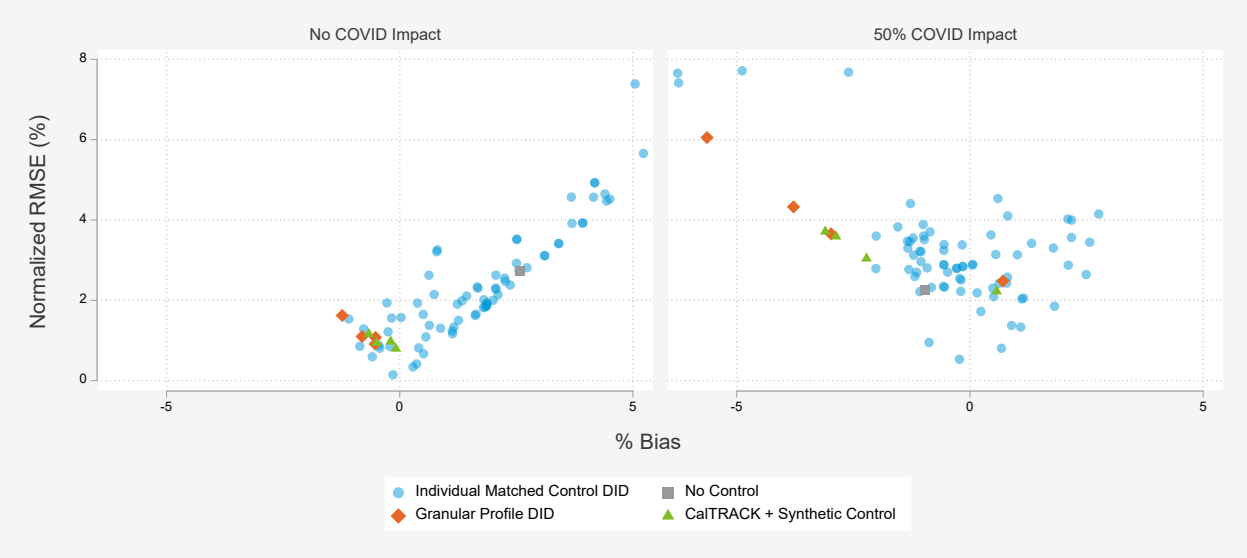
To assess whether these options could be viable in an efficient manner, a small-scale accuracy assessment was completed on residential and commercial electric loads, using approximately 2,000 premises in each sector. These 2,000 premises represented two bootstrapped iterations of 5, 50, and 500 sample aggregations and two separate treatment periods representing different exposure levels to pandemic conditions. The segmentation strategies used to construct the granular profiles were the same as what were used to construct the matched comparison groups. The results shown in this section are compared against the equivalent matched comparison CalTRACK results for the same set of participants.

RESIDENTIAL PERFORMANCE

Figure 32 shows the results for the assessment of residential CalTRACK with granular profiles. As in the other figures, bias is on the x-axis while normalized RMSE is on the y-axis. Each dot represents one run of an aggregation for the same set of customers across all different methods: CalTRACK with no comparison group, CalTRACK with a variety of matched comparators, CalTRACK with a synthetic control, and CalTRACK with a granular profile DiD. Because the matched comparison options tested also had many different tests of matching strategy and matching methods, there are substantially

more results for that method. The CalTRACK without a comparison group has only one strategy, while the granular profile DiD and the CalTRACK with synthetic profiles each tested performance for four segmentation strategies. In general, these alternative granular profile methods performed comparably to the CalTRACK with matched comparison groups, and in some cases performed better. The granular profile DiD approach tended to perform comparably to the CalTRACK with synthetic controls. Generally, it seems that segmentation strategy is more important than approach between these methods as both of the new methods can be subject to bias with exposure to COVID conditions.

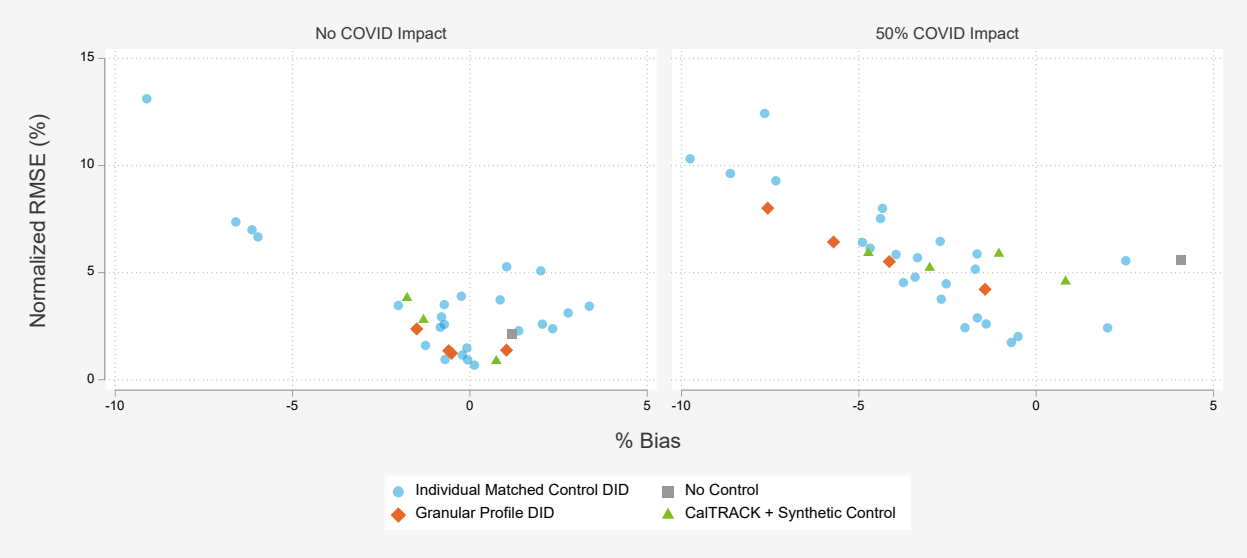
Figure 32: Residential Electric Results for Alternative Granular Profile Tests



Commercial Performance

Figure 33 shows the results for the assessment of commercial CalTRACK with granular profiles. As in the other figures, bias is on the x-axis while normalized RMSE is on the y-axis. Each dot represents one run of an aggregation for the same set of customers across all different methods: CalTRACK with no comparison group, CalTRACK with a variety of matched comparators, CalTRACK with a synthetic control, and CalTRACK with a granular profile DiD. Because the matched control options tested also had many different tests of matching strategy and matching methods, there are substantially more results for that method. The CalTRACK without comparators has only one strategy, while the granular profile DiD and the CalTRACK with synthetic controls each tested performance for four segmentation strategies. As with the residential results, the alternative models tested performed comparably to the CalTRACK with matched comparison groups, and certainly better than the CalTRACK model without a matched comparison group.

Figure 33: Commercial Electric Results for Alternative Granular Profile Tests



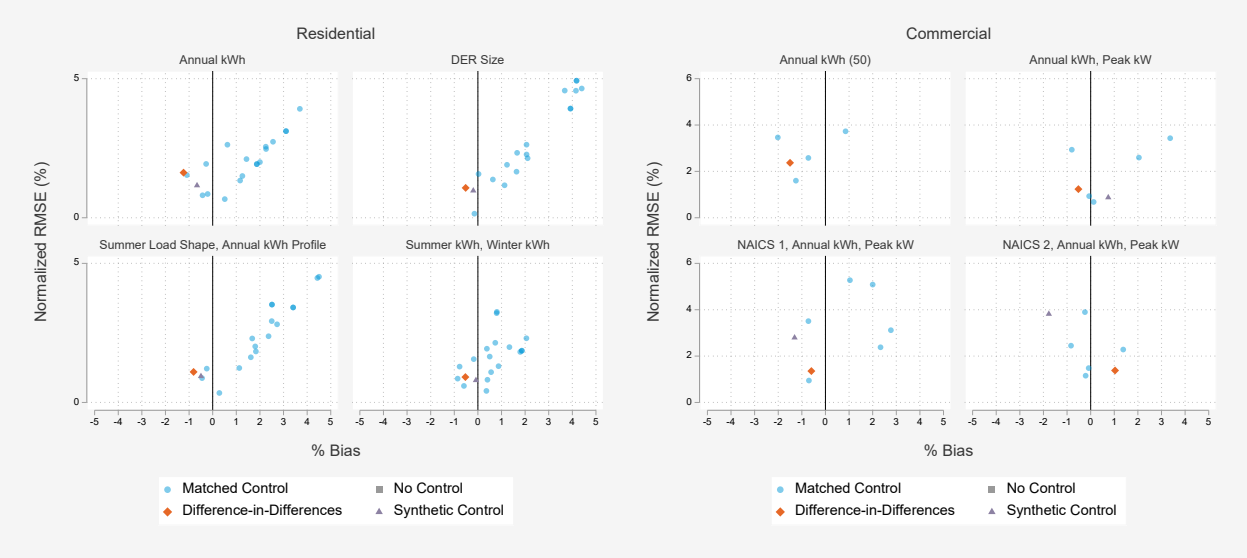
Results By Segmentation Strategy

Aggregated results can make it hard to assess the performance of each method relative to comparable alternatives. Figure 34 shows the results for each test, separated by segmentation strategy. Again, there are substantially more CalTRACK with matched comparator results as each matched group could be matched using stratified random sampling, propensity score matching, and Euclidian distance matching using a variety of matching characteristics. For residential results, the granular profile DiD approach fared better than most of the matched control methods in terms of performance and was generally consistent with the CalTRACK + Synthetic Control approach. It is clear from this assessment that certain matching methods can perform better for individual segments, but that matching on the wrong characteristics can also make things worse. On the commercial side, the granular profile DiD method outperformed the CalTRACK + Synthetic Control approach, but also generally fared better than most of the matched control methods.

Key Finding

Using aggregated granular profiles in the CalTRACK Difference-in-differences approach yields comparable results to using individual customer matched controls

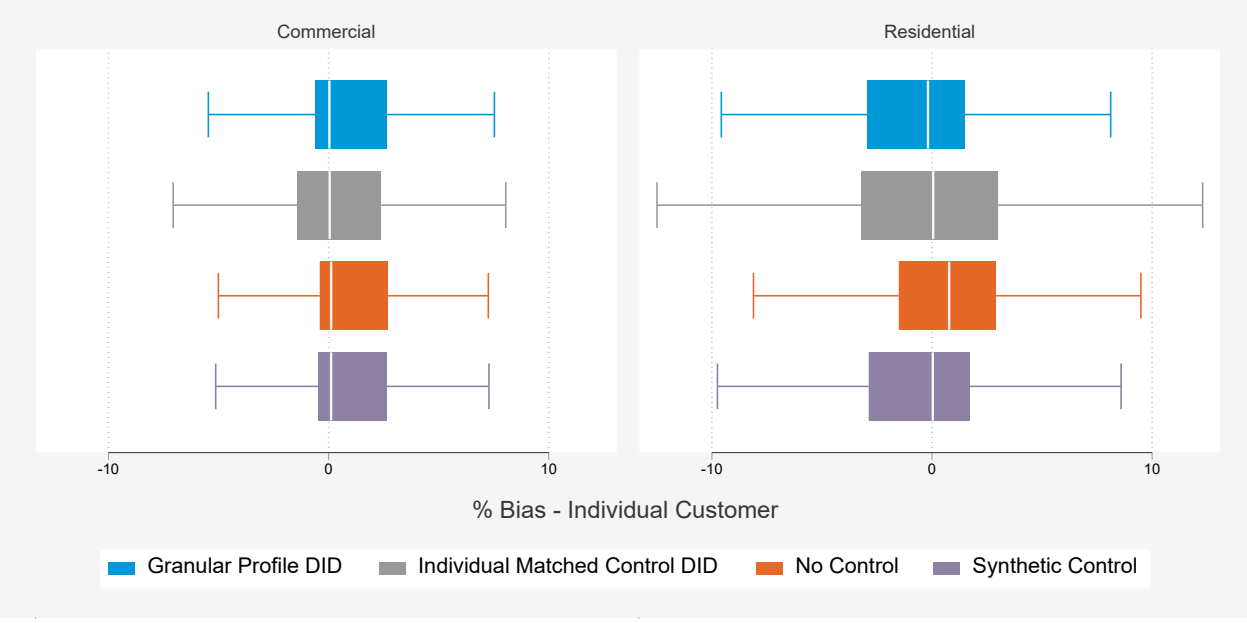
Figure 34: Comparison of Alternative Method Performance by Segmentation Strategy



Distribution of Individual Errors

As with the full accuracy assessment, results are also reported for individual level errors by framework in Figure 35. Note that these distributions should be compared cautiously with those shown in the prior section, as they represent fewer overall customers. Nevertheless, the same general trend emerges, where median errors remain close to 0%. As with the individual level results for the full accuracy assessment, outliers drive the majority of the bias in aggregate.

Figure 35: Distribution of Individual Customer Bias for Alternative Granular Profile Tests



4.7 OTHER DIMENSIONS OF ACCURACY

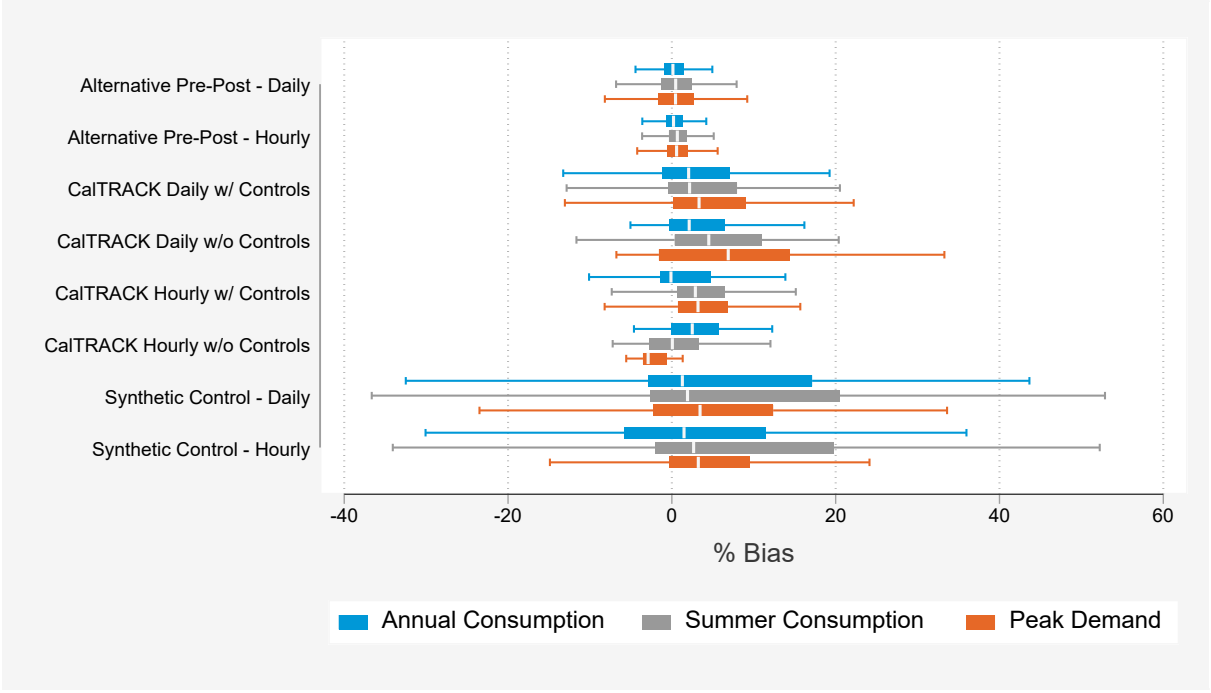
ACCURACY OF SUMMER CONSUMPTION AND PEAK DEMAND

Key Finding

No method is completely free of error

While the primary goal of this assessment is to find models that accurately estimate the counterfactual for the total consumption in the first post-treatment year, it is worth briefly investigating the accuracy of two additional metrics: summer loads and peak demand. Summer consumption is defined as the total consumption occurring from June to September in the 4-9pm period, while peak demand is defined as the average demand during the hottest three consecutive weekdays for each year in the 4-9pm period. Because daily models cannot capture peak hours, summer consumption and peak demand are the total kWh consumed on days that meet the criteria described above. Figure 36 shows the distribution of error across all the models tested for each framework for each metric, where a positive bias indicates a tendency to over-predict the value in question. In general, these results are reassuring: frameworks that perform well for annual consumption also tend to perform well for measures of summer consumption and peak demand. The Alternative Pre-Post models generally perform better for measures of summer consumption and peak demand compared to errors in their annual consumption, while most other models tend to perform worse for these subsets of interest.

Figure 36: Distribution of Errors for Each Framework by Metrics of Interest



Effects of Sample Size On Accuracy And Precision

The results in Section 4 clearly emphasize that many models tested are unbiased on average. Nevertheless, it is still important to clearly understand the distribution of possible errors when evaluating a given sample. Shown in Figure 37 to Figure 40 are the distribution of error for each bootstrapped iteration, for each of the best models in each framework at every tested sample size. A wide distribution means that a two samples of pseudo-participants, aggregated to a given level, will likely vary wildly in the estimate of the program's impacts. Narrow error bands mean that any two samples are likely to yield close to the same estimate of program performance. The closer the white line – the median error – is to zero in the plots below, the less biased the results are, but the width of the band indicates how precise all the bootstrapped results are. This is important to keep in mind, because all methods have a non-zero distribution of error. This means that, all else being equal, it will still be hard to detect small impacts using any of these models. At levels of 1,000 customers, individual estimates of error for a given bootstrapped iteration can still be biased up or down by approximately 5%, meaning that programs with expected savings of 5% or less may not be able to be estimated with statistical confidence. The ability to detect a 5% effect for smaller sample sizes is even more limited, because in many bootstrapped iterations, the error inherent to the method and that sample of participants is greater than 5%. For programs with small expected effects, population NMEC methods may not be able to detect the effect with the appropriate level of statistical confidence. While individual methods have wider or narrower error bars in the figures below, there is no one method that is error-free across all simulation runs at all sample sizes. Said another way, the choice of model can mitigate, but cannot overcome the difficulty of measuring small effects in small sample sizes. Therefore, program implementers and evaluators should keep this tradeoff between percent savings (effect size), sample size, and precision in mind when measuring savings using these metrics.

Key Finding

Accuracy and precision are dependent upon the number of sites aggregated together

Figure 37: Distribution of Error across Bootstrapped Iterations for Residential Electric

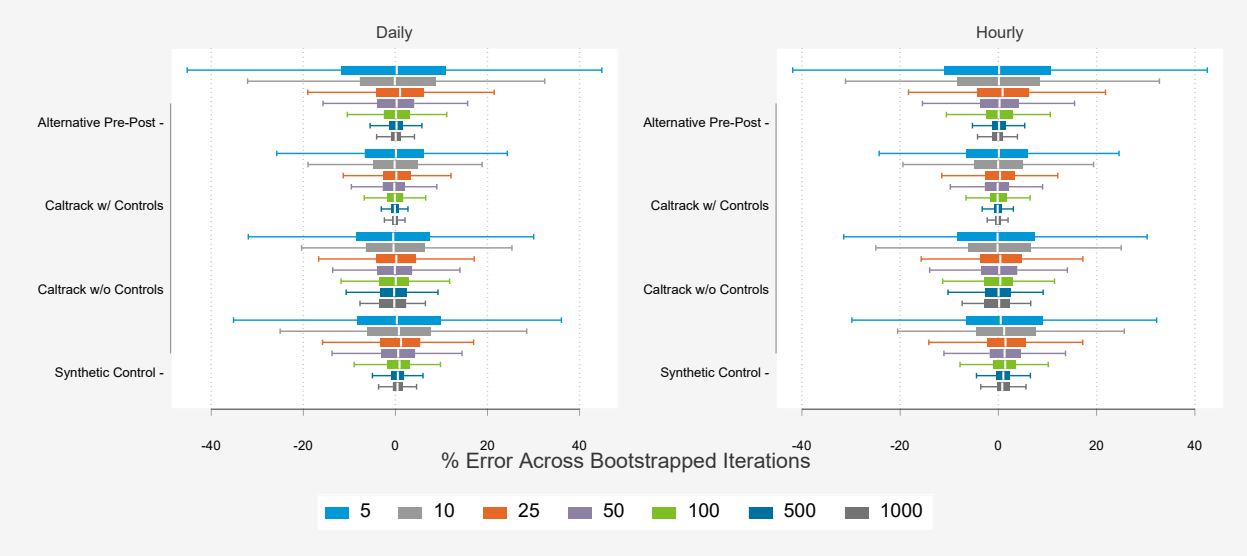


Figure 38: Distribution of Error across Bootstrapped Iterations for Commercial Electric

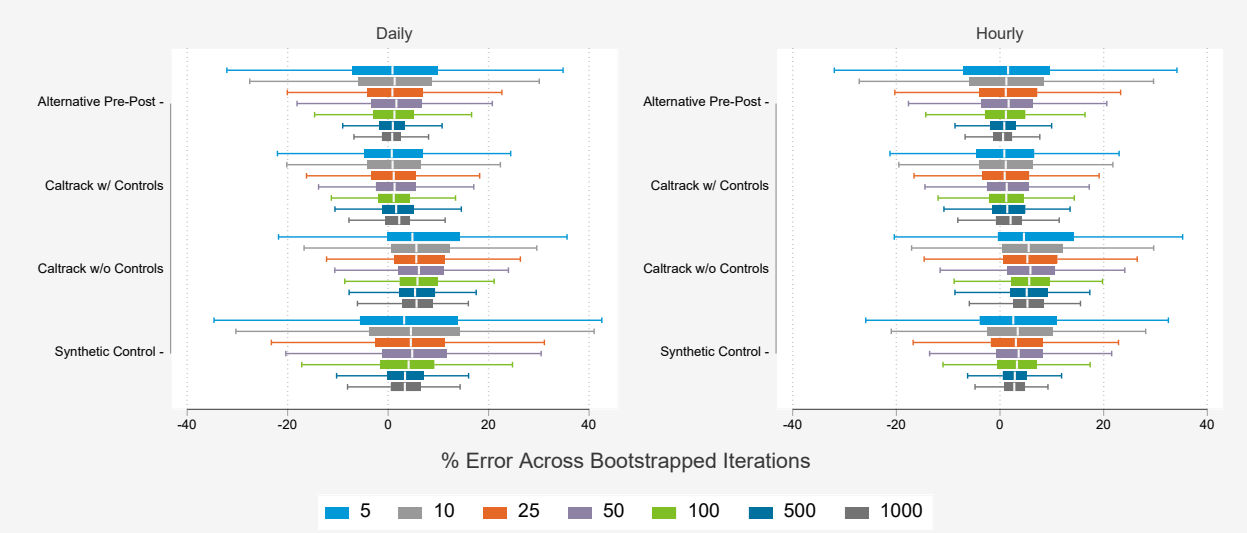


Figure 39: Distribution of Error across Bootstrapped Iterations for Residential Gas

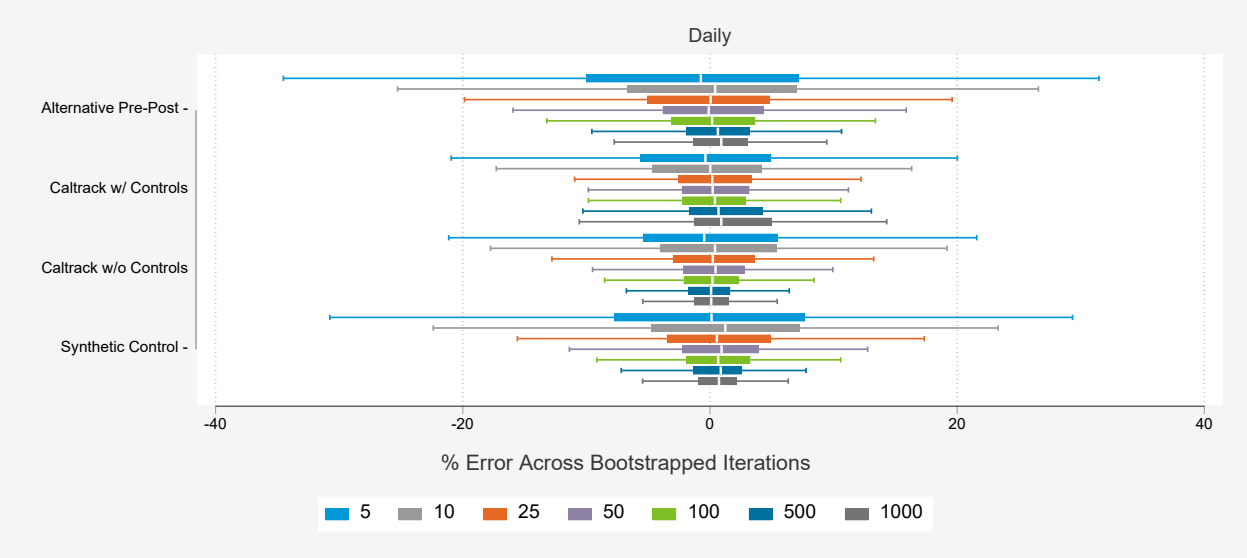
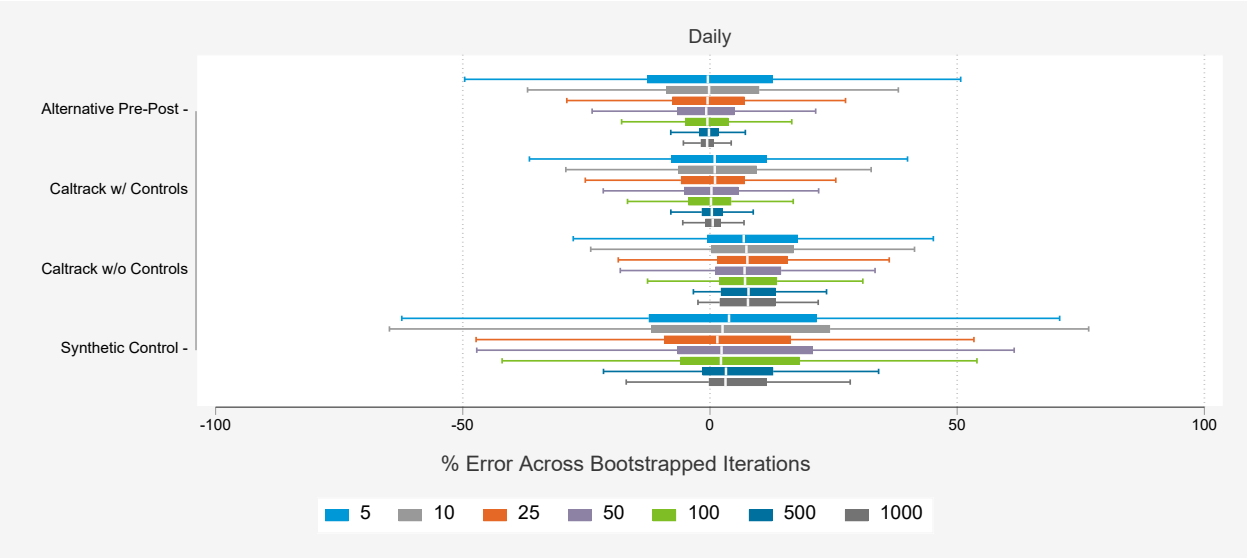


Figure 40: Distribution of Error across Bootstrapped Iterations for Commercial Gas



Factors that Influence Accuracy and Precision

Using the relative frequency of occurrence of each factor in the best-performing models can show specific drivers of accuracy and precision. Table 16 and Table 17 show these frequencies for each type of parameter: framework, matching variables, matching methods, regression models, and segmentation strategies. The best models in this case were the top 25 models for each sector (residential, commercial) and fuel type (electric, gas), across treatment periods and sample sizes. The frameworks that had the highest percentage of occurring in the best models were the Daily Alternative Pre-Post

model and the CalTRACK Daily without a comparison group. It's important to note that the surprisingly high frequency of success for the CalTRACK Daily model without a comparison group was exclusively due to its performance in the residential gas sector.

Matching variables that improved performance were percentiles of annual, summer and winter consumption, as well as heating weather sensitivity among gas models. Euclidian distance matching outperformed propensity score matching and stratified random sampling. Among models tested, the average daily temperature spline performed best, consistent with the Alternative Pre-Post Daily models being among the most common best models. Finally, there were many segmentation strategies that performed well. The strategies that performed best included combinations of 1-digit NAICS codes (for the commercial sector), four bins of annual consumption, four bins of peak load consumption, and monthly consumption load shapes.

Table 16: Frequency of Each Factor in Best-Performing Models for Electric

Type	Value	Total # of Models	% of Best Models
Framework	Alternative Pre-Post - Daily	2,688	10.4
	Alternative Pre-Post - Hourly	10,752	7.3
	CalTRACK Daily w/ Controls	2,688	6.3
	CalTRACK Daily w/o Controls	56	0.0
	CalTRACK Hourly w/ Controls	2,688	6.3
	CalTRACK Hourly w/o Controls	56	0.0
	Synthetic Control - Daily	476	0.0
	Synthetic Control - Hourly	1,904	0.0
Matching Variable	Load Factor	2,352	0.0
	Load Factor and Cooling Weather Sensitivity	2,352	1.2
	Load Factor and Bins of Peak Period Consumption (100 Bins)	2,352	1.2
	Load Factor and Bins of Annual Consumption (100 Bins)	2,352	22.6
	Mean Summer Demand 4pm-9pm	2,352	0.0
	Peak:Off Peak Ratio	2,352	4.8
	Bins of Peak Period Consumption (100 Bins)	2,352	4.8
	Bins of Annual Consumption (100 Bins)	2,352	25.0
Matching Method	Euclidian	6,272	12.1
	Propensity	6,272	5.8
	Strata	6,272	4.5
Regression Model	CalTRACK Daily	2,744	6.1
	CalTRACK Hourly	2,744	6.1
	Hour, DOW, Month, Daily Avg Temp Spline	6,328	7.1
	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	3,164	8.0
	Hour, DOW, Month, 3-Hour Moving Avg Temp Spline	3,164	8.0
	Hour, DOW, Month, Temp Spline	3,164	3.5
Segmen-tation	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	1,316	21.3
	Climate Zone, Solar Onsite, 1st and 2nd Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	1,316	6.4
	Climate Zone, Solar Onsite, DER System Size	3,668	3.8
	Climate Zone, Solar Onsite, 24-hour Load Shape (2 Bins) and 12-Month Consumption Profile (4 Bins)	3,668	3.8
	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	3,668	9.2
	Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)	1,316	12.8
	Climate Zone, Solar Onsite, Annual Consumption (50 Bins)	1,316	12.8
	Climate Zone, Solar Onsite, Summer Consumption & Winter Consumption (4 Bins Each)	3,668	2.3

Table 17: Frequency of Each Factor in Best-Performing Models for Gas

Type	Value	Total # of Models	% of Best Models
Framework	Alternative Pre-Post - Daily	2,688	47.9
	CalTRACK Daily w/ Controls	2,688	2.1
	CalTRACK Daily w/o Controls	56	50.0
	Synthetic Control - Daily	224	12.5
Matching Variable	Percentile ranking of annual consumption	1,344	27.1
	Percentile ranking of summer and winter consumption	1,344	27.1
	Bins of Heating Weather Sensitivity (100 Bins)	1,344	16.7
	Bins of Heating Weather Sensitivity (4 Bins)	1,344	29.2
Matching Method	Euclidian	1,792	28.1
	Propensity	1,792	26.6
	Strata	1,792	20.3
Regression Model	CalTRACK Daily	2,744	3.1
	Hour, DOW, Month, Daily Avg Temp Spline	2,912	45.2
Segmen-tation	Climate Zone, Annual Consumption (4 Bins) and Low Income Flag	700	16.0
	Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	700	32.0
	Climate Zone, 1st and 2nd Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	700	32.0
	Climate Zone, Annual Consumption (4 Bins)	700	32.0
	Climate Zone, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	700	20.0
	Climate Zone, Annual Consumption (50 Bins)	700	12.0
	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	1,400	26.0

5 DISCUSSION AND RECOMMENDATIONS

While summarizing the results of such a large-scale test of accuracy is never straightforward, there are several key insights that can be extracted from the findings. In brief:

1. **Population NMEC methods without comparison groups cannot account for the effects of the COVID-19 pandemic.** As discussed in Section 4.1, the effects of the pandemic can be easily observed in the bias of the current population NMEC (CalTRACK v2.0) models without any comparison groups. The magnitude and direction of the error varies by segment and interacts with the extreme weather that occurred in 2020.
2. **The existing population NMEC methods without comparison groups show upward bias even prior to the effects of the pandemic**. Across commercial and residential sectors, the models without comparison groups have a tendency to over-estimate program savings.
3. **Comparison groups improve accuracy of the CalTRACK method.** In essentially all cases, the inclusion of a comparison group, regardless of how it was incorporated into the counterfactual, improved accuracy and precision.
4. **When constructing a matched control group, the choice of segmentation and matching characteristics matter more than the method of matching customers**. In essentially all cases, Euclidian distance matching and propensity score matching performed similarly and outperformed stratified matching. Across frameworks and models, certain segmentation strategies and matching variables were consistently observed among the best-performing models.
5. **Synthetic controls may perform well but are highly sensitive to the choice of segmentation used.** Synthetic controls are a simple and straightforward method for incorporating aggregated granular profiles into counterfactual estimates. Nevertheless, the wrong choice of segmentation strategy can yield large errors for some segments and customers, which means they should be employed with caution. We believe this is because the inclusion of control profiles in addition to standard regression coefficients such as temperature splines, day of week indicators or other variables can introduce collinearity to the regression model that leads to coefficients that do not accurately capture effects of weather in the post-period.
6. **Using aggregated granular profiles in the CalTRACK Difference-in-Differences approach yields comparable results to using individual customer matched controls.** While individual matched controls groups may perform better or worse, depending on the matching method and matching variables, using a granular profile yields roughly the same accuracy and precision overall without the need to include individual non-participant data.
7. **Accuracy and precision are dependent upon the number of sites aggregated together.** Evaluators, regulators, and implementers should take care to note that the ability to detect small effects in small populations may not be possible even using the best models available.
8. **No method is completely free of error.** All methods tested had error and noise across the bootstrapped iterations and across individual participants. When estimating impacts for energy efficiency programs, evaluators and vendors should keep in mind that no model will be able to precisely distinguish small effects from background noise.

5.1 OVERALL RECOMMENDATIONS

Summarizing an analysis of this magnitude into a set of recommendations will necessarily require some degree of simplification. In general, there are several principles that should be relied upon beyond simply recommending the most accurate and precise method for each customer segment. The recommended method should:

1. Perform well for residential and commercial customers across fuel types.
2. Minimize the requirements for granular non-participant data to be divulged.
3. Be transparent and replicable by all interested parties.
4. Preserve the ability to do near real-time analysis by interested parties.

The findings summarized in Table 18 are the best performing model by framework. From the table it is clear that the frameworks that should be seriously considered are CalTRACK models with control groups and Alternative Pre-Post models with control groups. As discussed in Section 4.1, the existing Population NMEC methods (CalTRACK v2.0 models without comparison groups) fail to account for the effects of the pandemic (and, by extension, other systemic non-routine events) and are subject to systematic bias. Similarly, the Synthetic Control methods seem subject to over-specification and cannot reliably estimate consumption across a wide array of customer segments.

Both the CalTRACK models with matched controls and the Alternative Pre-Post models with matched controls generally perform well across all sectors and fuel types. Within these models, the Alternative Pre-Post Daily models, with an average daily temperature spline is a consistent performer. It is also important to keep in mind that while the CalTRACK plus Granular Profiles approach was not tested in the larger accuracy assessment, the results from the smaller-scale test conducted indicate that the Granular Profile Difference-in-Differences approach performed comparably to the CalTRACK with a matched control group.

Euclidian distance matching outperformed propensity score matching and stratified random sampling, though the choice of matching segments mattered more for producing a good matching group. Matching on percentiles of annual consumption and load factor were the preferred matching methods across models. Segmentation strategies that relied on climate zone, solar status, and bins of annual consumption and peak load for electric customers and 12-month consumption profile or summer and winter consumption bins for gas customers performed best overall.

Table 18: Summary of Best Models by Framework

Fuel	Framework	Matching Method	Segmentation	Matching Characteristics	Model	% Bias	Normalized RMSE
Residential Electric	CalTRACK Daily w/ Controls	Euclidian	Climate Zone, Solar Onsite, DER System Size	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Daily	-0.17	0.87
	CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, DER System Size	Load Factor and Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	-0.17	0.86
	Alternative Pre-Post - Hourly	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	-0.11	1.50
	Alternative Pre-Post - Daily	Euclidian	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	Load Factor and Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	0.06	1.51
	Synthetic Control - Daily	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	0.49	1.66
	Synthetic Control - Hourly	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	N/A	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	0.91	1.78
	CalTRACK Hourly w/o Controls	Participants Only	N/A	N/A	CalTRACK Hourly	-0.27	2.95
	CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	-0.51	3.20
Commercial Electric	Alternative Pre-Post - Hourly	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, 3-Hour Moving Avg Temp Spline	0.36	3.03
	Alternative Pre-Post - Daily	Propensity	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	Hour, DOW, Month, Daily Avg Temp Spline	0.54	3.11
	CalTRACK Hourly w/ Controls	Euclidian	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	CalTRACK Hourly	1.32	4.45
	Synthetic Control - Hourly	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)	N/A	Hour, DOW, Month, 12-Hour Moving Avg Temp Spline	2.62	3.73

Fuel	Framework	Matching Method	Segmentation	Matching Characteristics	Model	% Bias	Normalized RMSE
	CalTRACK Daily w/ Controls	Euclidian	Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)	Bins of Annual Consumption (100 Bins)	CalTRACK Daily	1.60	4.47
	CalTRACK Hourly w/o Controls	Participants Only	N/A	N/A	CalTRACK Hourly	5.35	6.10
	CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	5.65	6.39
	Synthetic Control - Daily	N/A	Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	3.34	4.96
Residential Gas	CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	0.34	3.23
	Alternative Pre-Post - Daily	Propensity	Climate Zone, Annual Consumption (4 Bins) and Low Income Flag	Percentile ranking of annual consumption	Hour, DOW, Month, Daily Avg Temp Spline	0.89	3.58
	Synthetic Control - Daily	N/A	Climate Zone, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	0.66	2.97
	CalTRACK Daily w/ Controls	Propensity	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	Percentile ranking of annual consumption	CalTRACK Daily	2.63	6.94
Commercial Gas	Alternative Pre-Post - Daily	Propensity	Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	Hour, DOW, Month, Daily Avg Temp Spline	-0.60	1.94
	CalTRACK Daily w/ Controls	Euclidian	Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)	Percentile ranking of summer and winter consumption	CalTRACK Daily	0.55	2.37
	CalTRACK Daily w/o Controls	Participants Only	N/A	N/A	CalTRACK Daily	7.78	8.16
	Synthetic Control - Daily	N/A	Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)	N/A	Hour, DOW, Month, Daily Avg Temp Spline	10.01	14.40

The key finding of this analysis, however, is that many models can perform accurately and with a high degree of precision. Rather than try to produce a single prescriptive method for Population NMEC analyses of energy efficiency programs, we instead recommend a framework by which proposed Population NMEC methods can be tested, certified, and used to estimate savings.

The benefit of such an approach is that it recognizes that, as broad as this accuracy study was, it was never possible to capture the specific characteristics of participants, regions, and measures of all varied Population NMEC-measured programs that are currently operating or will operate in California. Results from this study, specifically around regression methods, control group matching methods and segmentation strategies, can inform future model selection under this suggested framework, but the goal is to encourage additional innovation and transparency rather than select a single approach. Instead, this approach allows for flexibility in methods while requiring transparency in how the methods are implemented and in documenting how well the methods can isolate the effect of treatment from background noise. Once a method is tested and certified, it can be applied for estimating savings without the need for extensive explanation. Several key principles apply to certifying Population NMEC methods that perform well:

1. **Certification needs to be implemented by an independent party.** The party that develops a Population NMEC method cannot self-certify. We recommend that a party such as CALMAC or one of the National Laboratories be responsible for certification.

2. **Population NMEC methods need to be tested for reproducibility.** Reproducibility means obtaining consistent computational results using the same input data, computational steps, methods, and code.

3. **Population NMEC methods must meet pre-defined input analysis dataset structures and pre-defined output structures.** Defining the input data structure(s) ensures the method can be tested for reproducibility and also allows for an independent party to produce metrics for accuracy and precision. It also ensures that different NMEC methods can be applied to same datasets. At minimum, the input dataset must include AMI hourly data, public hourly weather data, and when energy efficiency measures where installed. Defining the output data structure allows utilities, vendor, and public entities to build dashboards and tools to display the results regardless of which underlying NMEC algorithm is applied.

4. **Population NMEC metrics of accuracy (bias) and precision should be calculated out-of-sample at a portfolio level.** Metrics of accuracy measure the tendency to over or underestimate the baseline. They are used to assess if a method or model is biased. Metrics for precision measure how close individual hourly or daily estimates are to the actual answer and measure notice. While evaluators and other interested parties may choose to pick a model that is accurate and precise for individual sites,

certification must be done on an aggregate program basis rather than for individual participants. Out of sample validation is critical since models that are over-fitted can perform well in sample and poorly out of sample

5

The measurement of accuracy (bias) and precision metrics should be calculated by the independent party certifying the method using a blind test. Using an independent party ensures consistency and independence of the metric calculations. We also recommend that test be a blind test, meaning that proponents of the Population NMEC method do not have access to dataset used to test the proposed method.

6

To be certified, an NMEC method must meet specific criteria for accuracy and precision. Accuracy and precision, as noted above, are dependent upon the size of the participant population of interest. Therefore, targets for model acceptability are similarly size-dependent. Table 19 shows the proposed targets as a function of sample size. The metrics for bias and precision may need to be modified for sites with solar to account for the fact that large energy users can have lower energy consumption at the meter. These cutoffs were chosen on the basis of the bootstrapped accuracy test results and were set such that 15% of all models tested in this study met the criteria.

Table 19: Proposed Out-of-Sample Accuracy and Precision Targets for Certification

Sample Size	Normalized RMSE		Mean % Bias	
	Residential	Commercial	Residential	Commercial
25	< 15%	< 20%	< 0.5%	< 0.5%
100	< 10%	< 10%	< 0.5%	< 0.5%
500	< 5%	< 5%	< 0.5%	< 0.5%

7

Population NMEC methods must be separately certified for residential, small and medium businesses, and large businesses and for sites with and without solar. The approach allows methods that work for specific segments to be applied.

8

The out-of-sample metrics for accuracy and precision of Population NMEC methods tested for certification should be posted on a public repository such as CALMAC. Public data on the performance of different models is useful for helping develop new methods and avoiding redundant efforts.

9

The code for estimating savings needs to be publicly available and include examples of how it is applied. A key goal of NMEC is transparency, which means everyone has access to the analysis code and examples for how to apply it to estimate savings. The code to estimate the savings needs to be in a standard statistical computing language – Python, R, SAS, Stata, Julia.

10

The method used must be selected and certified in advance of program implementation. Requiring up-front identification of the estimation procedure ensures that there is no post-hoc model selection that would produce more favorable results.

Despite the breadth of the accuracy assessment study, it was not possible to test all the current and future techniques and models. A framework for certification encourages innovation and competition while ensuring transparency and consistency in outputs. By contrast, prescribing a single method for NMEC analyses can limit innovation and reduce competition.

APPENDIX A: INDEX OF ALL MODELS TESTED

Table 20: Specifications and Control Group Strategies Tested

Control Group Construction Method	No Matching	Sampling within Segments	Matched Control
No Controls	■		
Matched Controls			■
Synthetic Controls		■	
Granular Aggregated Profiles		■	

Control Groups Tested with Regression Frameworks	CalTRACK	Alternative Pre-Post
No Controls	■	
Matched Controls	■	■
Synthetic Controls	■	■
Granular Aggregated Profiles	■	

Regression Specifications Tested	CalTRACK	Alternative Pre-Post
CalTRACK Daily	■	
CalTRACK Hourly	■	
Hour, DOW, Month, Temp Spline		■
Hour, DOW, Month, Daily Avg Temp Spline		■
Hour, DOW, Month, 3-Hour Moving Avg Temp Spline		■
Hour, DOW, Month, 12-Hour Moving Avg Temp Spline		■

Data Granularity Tested	Electric	Gas
Daily	■	■
Hourly	■	

Table 21: Segmentation and Matching Strategies Tested

Segmentation Strategies Tested	Res Elec	Com Elec	Res Gas	Com Gas
Climate Zone, Solar Onsite, Annual Consumption (4 Bins)	■			
Climate Zone, Solar Onsite, Summer Consumption & Winter Consumption (4 Bins Each)	■			
Climate Zone, Solar Onsite, 24-hour Load Shape (2 Bins) and 12-Month Consumption Profile (4 Bins)	■			
Climate Zone, Solar Onsite, DER System Size	■			
Climate Zone, Solar Onsite, Annual Consumption (50 Bins)		■		
Climate Zone, Solar Onsite, Annual Consumption (4 Bins) and Peak Load (4 Bins)		■		
Climate Zone, Solar Onsite, 1st Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)		■		
Climate Zone, Solar Onsite, 1st and 2nd Digit NAICS Code, Annual Consumption (4 Bins) and Peak Load (4 Bins)		■		
Climate Zone, Annual Consumption (4 Bins)			■	
Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)			■	
Climate Zone, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)			■	
Climate Zone, Annual Consumption (4 Bins) and Low Income Flag			■	
Climate Zone, Annual Consumption (50 Bins)				■
Climate Zone, Summer Consumption & Winter Consumption (4 Bins Each)				■
Climate Zone, 1st Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)				■
Climate Zone, 1st and 2nd Digit NAICS Code, Annual Consumption(4 Bins and 12-Month Consumption Profile (2 Bins)				■

Matching Strategies Tested	Res	Com	Res	Com
Mean Summer Demand 4pm-9pm	■			
Load Factor	■			
Peak:Off Peak Ratio	■			
Load Factor and Cooling Weather Sensitivity	■			
Load Factor and Bins of Annual Consumption (100 Bins)	■			
Load Factor and Bins of Peak Period Consumption (100 Bins)	■			
Bins of Annual Consumption (100 Bins)		■		
Bins of Peak Period Consumption (100 Bins)		■		
Percentile ranking of annual consumption			■	
Percentile ranking of summer and winter consumption			■	
Bins of Heating Weather Sensitivity (4 Bins)			■	
Bins of Heating Weather Sensitivity (100 Bins)			■	
Percentile ranking of annual consumption				■
Percentile ranking of summer and winter consumption				■
Bins of Heating Weather Sensitivity (4 Bins)				■
Bins of Heating Weather Sensitivity (100 Bins)				■

APPENDIX B: ACCURACY AND PRECISION FOR ALL MODELS

See Attached Workbood

APPENDIX C: ESTIMATING THE EFFECTS OF THE COVID-19 PANDEMIC

For this analysis, it was important to first understand the effects of the COVID-19 pandemic on electricity consumption. To estimate these changes, a Random Forest model was trained and used to relate variables such as weather, seasonal and other temporal variables to electricity demand for different residential and non-residential segments of PG&E's service territory. The analysis was completed at this level of granularity to understand the heterogeneity in response to stay-at-home orders, temporary closures of non-essential businesses, reduced operating hours and capacity restrictions and other behavior changes.

- The estimates were developed with historical hourly energy demand data, aggregated to various sectors of interest, hourly weather data, and information about when the pandemic began (March, 2020). An estimate of consumption was developed for each segment using pre-pandemic data. This data was then used to predict what consumption would have been absent the pandemic, given weather conditions, day of the week, month, and other seasonal factors. The model was developed using a machine learning method known as a Random Forest Regressor. The final model was selected on the basis of out of sample cross-validation (withholding 25% of the data). The model error is on the order of varies for each sector and is shown in Table 22 and Table 23. Any impact outside of this range can be attributed to external shocks such as the pandemic. Those shocks, by sector and month are shown in Table 24 and
-

Table 25.

Table 22: Model Out-of-Sample Errors for Modeling COVID Impacts in Residential Sectors

Segment	Count	% Bias	Norm RMSE
Coastal Bay Area, not low income, no solar	10,918	0.0%	1.4%
Coastal Bay Area, low income, no solar	3,357	0.0%	1.5%
Coastal Bay Area, not low income, with solar	564	-1.6%	10.5%
Coastal Bay Area, low income, with solar	91	0.7%	9.9%
Coastal Other, not low income, no solar	8,270	-0.3%	1.8%
Coastal Other, low income, no solar	2,446	0.0%	1.9%
Coastal Other, not low income, with solar	675	-1.8%	9.0%
Coastal Other, low income, with solar	71	-3.1%	9.4%
Inland, not low income, no solar	10,084	-0.4%	1.9%
Inland, low income, no solar	6,564	0.0%	1.8%
Inland, not low income, with solar	1,606	-2.7%	10.5%
Inland, low income, with solar	282	-2.4%	8.2%

Table 23: Model Out-of-Sample Errors for Modeling COVID Impacts in Commercial Sectors

Segment	Count	% Bias	Norm RMSE
Agriculture, Forestry, Fishing and Hunting	595	-0.8%	3.6%
Mining, Quarrying, and Oil and Gas Extraction	323	2.4%	3.6%
Utilities	2,239	-0.2%	2.0%
Construction	827	0.1%	2.9%
Manufacturing	2,750	-0.1%	2.5%
Wholesale Trade	1,192	-0.5%	2.6%
Retail Trade	6,343	-0.8%	1.3%
Transportation and Warehousing	2,141	0.7%	1.8%
Information	7,418	1.6%	1.9%
Finance and Insurance	1,025	-1.4%	2.6%
Real Estate and Rental and Leasing	4,931	-1.4%	2.3%
Professional, Scientific, and Technical Services	1,433	0.3%	1.3%
Management of Companies and Enterprises	816	0.6%	1.8%
Administrative and Support and Waste Management and Remediation Services	642	-1.9%	3.0%
Educational Services	817	-0.7%	4.1%
Health Care and Social Assistance	2,909	-8.7%	41.2%
Arts, Entertainment, and Recreation	1,332	-0.5%	1.3%
Accommodation and Food Services	5,645	-0.3%	1.0%
Other Services (except Public Administration)	4,101	0.1%	2.1%
Public Administration	1,058	-0.7%	1.6%

Table 24: Residential COVID Impacts by Sector and Month

Segment	Average Daily kWh	Jan-20	Feb-20	Mar-20	Apr-20	May-20	Jun-20	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Jan-21	Feb-21	Mar-21
Coastal Bay Area, not low income, no solar	Observed	13.86	12.61	13.05	12.53	11.68	11.59	11.43	12.18	12.14	11.99	13.59	15.21	14.82	13.80	13.39
	Predicted	13.84	12.69	12.63	11.56	10.87	10.76	10.59	11.14	11.02	11.16	12.44	13.86	13.88	12.98	12.91
	% Change	0.1%	-0.6%	3.3%	8.4%	7.4%	7.6%	7.9%	9.3%	10.2%	7.5%	9.2%	9.7%	6.8%	6.3%	3.8%
Coastal Bay Area, low income, no solar	Observed	14.16	12.61	12.84	12.26	11.38	11.24	11.17	11.78	11.80	11.70	13.75	15.79	15.28	14.04	13.44
	Predicted	14.06	12.66	12.32	11.26	10.51	10.45	10.33	10.73	10.59	10.69	12.14	13.88	13.78	12.85	12.62
	% Change	0.7%	-0.4%	4.2%	8.9%	8.3%	7.6%	8.2%	9.8%	11.4%	9.4%	13.3%	13.8%	10.9%	9.3%	6.5%
Coastal Bay Area, not low income, with solar	Observed	18.08	10.59	11.79	6.67	2.73	1.58	2.35	7.22	10.84	12.31	16.10	20.66	19.82	13.51	9.35
	Predicted	18.15	11.11	10.23	5.41	0.76	0.59	0.81	4.80	5.21	8.21	14.61	17.96	17.84	12.78	9.86
	% Change	-0.4%	-4.6%	15.2%	23.2%	262.1%	169.4%	188.9%	50.4%	108.2%	49.9%	10.2%	15.1%	11.1%	5.7%	-5.1%
Coastal Bay Area, low income, with solar	Observed	13.67	9.13	9.13	6.20	3.34	3.82	4.41	6.69	8.73	9.41	12.92	16.79	15.31	11.43	7.98
	Predicted	13.81	9.47	7.93	4.61	1.71	1.25	2.56	4.93	4.99	6.66	11.19	14.37	13.37	10.47	8.17
	% Change	-1.0%	-3.5%	15.2%	34.4%	95.7%	205.2%	72.1%	35.9%	75.0%	41.2%	15.5%	16.8%	14.5%	9.1%	-2.3%
Coastal Other, not low income, no solar	Observed	16.60	14.95	15.53	15.16	15.71	17.11	16.98	19.92	17.58	15.71	16.48	18.44	17.46	16.21	15.72
	Predicted	16.66	15.08	15.44	14.30	14.40	15.74	15.57	17.82	16.24	14.54	15.68	17.16	16.57	15.62	15.48
	% Change	-0.3%	-0.8%	0.6%	6.1%	9.1%	8.7%	9.1%	11.8%	8.2%	8.0%	5.1%	7.5%	5.4%	3.8%	1.6%
Coastal Other, low income, no solar	Observed	17.86	15.73	16.12	15.21	15.37	16.97	17.12	19.79	17.62	15.93	17.66	20.41	19.21	17.56	16.81
	Predicted	17.80	15.72	15.76	14.09	13.62	15.04	15.09	16.88	15.47	13.84	15.83	17.79	17.33	16.13	15.82
	% Change	0.4%	0.1%	2.3%	7.9%	12.9%	12.8%	13.4%	17.2%	13.9%	15.1%	11.6%	14.7%	10.8%	8.8%	6.3%
Coastal Other, not low income, with solar	Observed	21.17	11.60	13.21	6.80	5.20	5.56	5.88	16.35	17.41	16.00	19.04	24.51	23.43	15.74	10.71
	Predicted	21.17	11.95	11.93	5.62	2.71	4.41	4.56	13.01	11.32	11.01	18.33	21.33	20.19	14.70	10.16
	% Change	0.0%	-2.9%	10.7%	20.9%	92.1%	26.1%	28.9%	25.7%	53.8%	45.3%	3.9%	14.9%	16.1%	7.1%	5.5%
Coastal Other, low income, with solar	Observed	16.20	9.34	10.86	5.77	3.91	4.28	5.56	15.28	15.27	13.37	14.34	18.12	16.80	10.44	6.54
	Predicted	16.41	9.57	9.30	4.04	1.61	3.68	3.96	10.80	9.69	8.49	13.65	16.69	16.09	12.09	7.87
	% Change	-1.3%	-2.4%	16.8%	42.9%	141.9%	16.4%	40.2%	41.5%	57.6%	57.4%	5.1%	8.5%	4.4%	-13.7%	-16.9%
Inland, not low income, no solar	Observed	17.82	16.05	16.23	16.65	19.66	25.15	28.88	31.65	24.28	18.68	17.65	20.23	19.04	17.48	16.71
	Predicted	17.91	16.15	16.35	16.25	18.51	23.80	26.82	29.92	24.02	17.93	16.75	18.85	17.91	16.75	16.31
	% Change	-0.5%	-0.6%	-0.7%	2.5%	6.2%	5.7%	7.7%	5.8%	1.1%	4.2%	5.4%	7.3%	6.3%	4.4%	2.5%
Inland, low income, no solar	Observed	17.39	15.25	15.13	16.19	20.38	27.27	32.32	34.44	26.36	19.33	17.12	19.75	18.48	16.71	15.86
	Predicted	17.38	15.26	14.85	15.50	18.54	24.63	28.83	30.91	24.38	17.79	15.51	17.77	16.94	15.68	15.06
	% Change	0.0%	-0.1%	1.9%	4.5%	10.0%	10.7%	12.1%	11.4%	8.1%	8.6%	10.3%	11.2%	9.1%	6.6%	5.3%
Inland, not low income, with solar	Observed	18.20	5.81	5.84	0.49	2.95	10.75	19.27	29.71	23.21	16.06	14.63	21.44	18.74	9.07	2.20
	Predicted	18.37	6.06	5.26	0.07	0.86	8.89	18.65	25.34	16.79	9.61	13.66	18.24	15.42	9.53	3.28
	% Change	-0.9%	-4.1%	11.0%	641.8%	242.1%	20.9%	3.3%	17.2%	38.3%	67.1%	7.1%	17.5%	21.5%	-4.7%	-33.0%
Inland, low income, with solar	Observed	17.41	6.95	5.99	2.72	6.16	14.82	24.21	33.38	26.48	18.48	15.24	21.88	18.42	10.32	4.80
	Predicted	17.55	7.25	5.54	1.53	2.99	12.76	22.43	27.97	20.08	11.56	13.29	17.88	15.20	10.09	3.86
	% Change	-0.8%	-4.2%	8.2%	78.1%	105.8%	16.1%	7.9%	19.3%	31.9%	59.9%	14.7%	22.3%	21.2%	2.3%	24.5%

Table 25: Commercial COVID Impacts by Sector and Month

Segment	Average Daily kWh	Jan-20	Feb-20	Mar-20	Apr-20	May-20	Jun-20	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20
Agriculture, Forestry, Fishing and Hunting	Observed	119.0	101.7	103.3	95.6	110.2	125.9	134.0	187.4	170.9	154.9	127.4	113.0
	Predicted	120.6	102.6	109.1	114.6	124.9	143.0	171.2	191.2	176.9	160.9	136.5	121.4
	% Change	-1.3%	-0.8%	-5.3%	-16.5%	-11.8%	-12.0%	-21.7%	-2.0%	-3.4%	-3.7%	-6.7%	-6.9%
Mining, Quarrying, and Oil and Gas Extraction	Observed	2408.0	2386.6	2329.1	2039.0	1804.8	1970.7	1925.1	1902.8	1932.9	1789.5	1735.1	1799.9
	Predicted	2385.7	2372.3	2351.9	2369.5	2343.6	2439.1	2474.7	2385.9	2442.5	2431.9	2380.5	2365.2
	% Change	0.9%	0.6%	-1.0%	-13.9%	-23.0%	-19.2%	-22.2%	-20.2%	-20.9%	-26.4%	-27.1%	-23.9%
Utilities	Observed	328.6	341.6	329.4	334.4	385.8	426.8	459.1	482.2	447.2	429.6	394.3	369.6
	Predicted	331.4	342.5	344.2	361.8	407.3	447.2	475.3	474.3	457.2	432.2	368.1	355.4
	% Change	-0.8%	-0.2%	-4.3%	-7.6%	-5.3%	-4.5%	-3.4%	1.7%	-2.2%	-0.6%	7.1%	4.0%
Construction	Observed	133.4	132.7	120.8	111.5	115.7	132.8	132.2	135.0	127.4	123.1	114.0	118.1
	Predicted	132.1	130.0	122.7	125.6	125.6	139.5	140.7	142.2	142.2	131.3	123.2	123.3
	% Change	1.0%	2.1%	-1.6%	-11.3%	-7.9%	-4.8%	-6.1%	-5.1%	-10.4%	-6.3%	-7.5%	-4.2%
Manufacturing	Observed	1105.5	1124.0	1059.4	1053.2	1083.5	1191.4	1172.9	1257.5	1258.8	1211.3	1060.9	1009.6
	Predicted	1121.0	1128.6	1154.5	1202.2	1208.5	1299.6	1298.1	1311.7	1340.1	1276.2	1150.1	1079.7
	% Change	-1.4%	-0.4%	-8.2%	-12.4%	-10.3%	-8.3%	-9.6%	-4.1%	-6.1%	-5.1%	-7.8%	-6.5%
Wholesale Trade	Observed	481.8	464.8	444.7	438.2	451.2	506.8	519.9	536.6	536.3	516.0	459.4	460.8
	Predicted	487.7	470.0	490.1	504.8	505.7	550.5	558.1	566.5	569.4	535.3	492.8	486.2
	% Change	-1.2%	-1.1%	-9.3%	-13.2%	-10.8%	-7.9%	-6.9%	-5.3%	-5.8%	-3.6%	-6.8%	-5.2%
Retail Trade	Observed	300.0	301.9	282.5	274.4	311.4	358.6	374.3	389.6	372.4	344.8	292.0	284.4
	Predicted	303.3	305.2	313.7	337.5	369.3	398.7	405.8	415.5	408.0	375.6	317.8	309.5
	% Change	-1.1%	-1.1%	-9.9%	-18.7%	-15.7%	-10.1%	-7.8%	-6.2%	-8.7%	-8.2%	-8.1%	-8.1%
Transportation and Warehousing	Observed	376.3	370.7	356.3	343.9	355.0	389.3	390.3	401.2	393.5	371.3	332.9	332.3
	Predicted	376.8	368.2	365.1	374.3	389.2	412.3	417.2	424.0	421.6	399.3	371.4	375.2
	% Change	-0.1%	0.7%	-2.4%	-8.1%	-8.8%	-5.6%	-6.4%	-5.4%	-6.7%	-7.0%	-10.4%	-11.4%
Information	Observed	209.1	212.7	208.8	211.2	219.5	227.2	229.3	231.8	225.4	220.7	204.0	199.7
	Predicted	206.1	209.0	198.1	205.0	213.0	220.5	223.0	228.9	224.0	218.7	202.1	200.8
	% Change	1.5%	1.8%	5.4%	3.0%	3.0%	3.0%	3.8%	1.3%	0.6%	0.9%	1.0%	-0.5%
Finance and Insurance	Observed	177.2	178.5	170.6	167.6	185.5	209.3	213.8	223.3	207.8	187.3	156.3	158.3
	Predicted	181.1	181.4	187.8	202.2	215.5	237.4	240.7	248.5	239.4	219.0	186.2	184.0
	% Change	-2.1%	-1.6%	-9.1%	-17.1%	-13.9%	-11.8%	-11.2%	-10.1%	-13.2%	-14.5%	-16.1%	-14.0%

Segment	Average Daily kWh	Jan-20	Feb-20	Mar-20	Apr-20	May-20	Jun-20	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20
Real Estate and Rental and Leasing	Observed	268.9	271.6	243.8	221.9	236.6	268.8	270.9	284.8	280.0	266.5	235.5	232.9
	Predicted	277.2	278.5	293.9	305.6	315.6	336.5	337.4	345.5	340.9	325.9	289.3	282.6
	% Change	-3.0%	-2.5%	-17.0%	-27.4%	-25.0%	-20.1%	-19.7%	-17.6%	-17.9%	-18.2%	-18.6%	-17.6%
Professional, Scientific, and Technical Services	Observed	528.9	541.9	506.9	497.1	532.7	583.7	585.3	613.3	599.4	538.7	457.7	444.3
	Predicted	533.7	545.2	551.1	579.4	605.6	640.6	643.2	667.5	658.3	628.8	554.2	534.4
	% Change	-0.9%	-0.6%	-8.0%	-14.2%	-12.0%	-8.9%	-9.0%	-8.1%	-9.0%	-14.3%	-17.4%	-16.9%
Management of Companies and Enterprises	Observed	463.8	467.2	438.1	423.3	446.7	491.0	493.4	518.6	500.0	483.6	424.3	414.0
	Predicted	464.2	465.3	461.5	478.6	500.0	535.9	542.1	558.4	547.2	517.9	467.4	459.1
	% Change	-0.1%	0.4%	-5.1%	-11.6%	-10.7%	-8.4%	-9.0%	-7.1%	-8.6%	-6.6%	-9.2%	-9.8%
Administrative and Support and Waste Management and	Observed	269.7	259.6	242.8	227.4	225.0	250.7	251.4	259.1	251.6	246.3	226.1	226.6
	Predicted	275.7	265.0	279.8	284.0	278.7	293.5	296.5	290.7	290.2	273.9	262.7	264.4
	% Change	-2.2%	-2.0%	-13.2%	-19.9%	-19.3%	-14.6%	-15.2%	-10.9%	-13.3%	-10.1%	-13.9%	-14.3%
Educational Services	Observed	342.9	322.1	259.3	165.0	179.8	224.1	240.7	320.2	307.3	283.7	249.5	250.8
	Predicted	346.7	323.7	338.3	318.8	325.7	322.1	307.2	412.2	412.9	378.7	320.4	321.0
	% Change	-1.1%	-0.5%	-23.4%	-48.3%	-44.8%	-30.4%	-21.7%	-22.3%	-25.6%	-25.1%	-22.1%	-21.9%
Health Care and Social Assistance	Observed	383.1	379.5	364.5	364.1	402.0	454.0	459.6	487.1	456.3	415.2	357.5	359.5
	Predicted	609.9	571.3	770.4	419.9	442.0	483.7	498.0	515.3	489.7	446.8	403.9	436.2
	% Change	-37.2%	-33.6%	-52.7%	-13.3%	-9.0%	-6.1%	-7.7%	-5.5%	-6.8%	-7.1%	-11.5%	-17.6%
Arts, Entertainment, and Recreation	Observed	377.4	382.5	304.3	220.8	250.8	319.9	335.8	332.2	311.6	315.9	288.9	269.9
	Predicted	381.3	383.3	385.8	407.5	437.7	474.5	491.2	504.1	477.3	436.6	388.4	381.7
	% Change	-1.0%	-0.2%	-21.1%	-45.8%	-42.7%	-32.6%	-31.6%	-34.1%	-34.7%	-27.7%	-25.6%	-29.3%
Accommodation and Food Services	Observed	319.6	322.5	275.1	234.1	273.8	323.7	337.9	356.2	339.5	319.1	276.5	257.8
	Predicted	321.8	324.8	327.8	347.1	377.2	402.1	415.1	431.7	407.5	378.0	329.1	325.7
	% Change	-0.7%	-0.7%	-16.1%	-32.6%	-27.4%	-19.5%	-18.6%	-17.5%	-16.7%	-15.6%	-16.0%	-20.8%
Other Services (except Public Administration)	Observed	112.7	112.0	94.5	78.2	87.2	103.2	106.0	109.9	108.8	103.3	94.7	95.4
	Predicted	112.7	111.4	108.0	111.1	116.4	125.5	127.6	131.0	127.7	119.7	110.6	112.4
	% Change	-0.1%	0.5%	-12.5%	-29.6%	-25.1%	-17.8%	-16.9%	-16.1%	-14.9%	-13.7%	-14.4%	-15.1%
Public Administration	Observed	411.9	411.1	402.6	410.0	453.1	511.9	520.7	543.6	508.1	459.7	377.7	370.8
	Predicted	417.6	415.5	430.4	464.8	503.6	558.3	579.4	590.8	561.9	507.0	431.3	419.9
	% Change	-1.4%	-1.1%	-6.5%	-11.8%	-10.0%	-8.3%	-10.1%	-8.0%	-9.6%	-9.3%	-12.4%	-11.7%

6 REVIEWER COMMENTS: RECURVE

Comments from Recurve were organized in to three main categories. Comments and recommendations for each topic are highlighted below

THE ONGOING CHALLENGES OF DATA ACCESS TO SCALE DISTRIBUTED ENERGY RESOURCES AND ENABLE ROBUST PERFORMANCE ANALYSIS IN CALIFORNIA

- "Technical solutions, like differential privacy which can systematically add noise to protect against individual re-identification, offer innovations for protecting customer privacy balanced against the essential public good derived from analysis of non-participant data. Actual data transfer can become a more replicable process that utilities can operationalize more easily. Overall, investments in technology and protocols to continue to improve these capabilities should be a priority, rather than investing in methodological workarounds that can potentially undermine access to valid performance analysis."
    - **Response**: Demand Side Analytics takes no position on this comment. The analysis performed by DSA on behalf of PG&E had a goal of assessing the accuracy of as wide a variety of methods as possible. Entities that have access to meter data are the only ones that can determine what data transfer is feasible at scale.
    - PG&E adds that it is misleading to characterize synthetic controls as a "methodological workaround." They provide advantages over using actual non-participant data as comparators. They can be shared with third parties readily (including program implementers) without the need for IT security reviews and cybersecurity insurance. Their execution does not require investments in technology that are subject to competing IT needs, rates pressures, and Commission approval of proposed budgets.
- "While the report makes no mention of the GRIDmeter methods nor is it included in the "tournament" results it most closely resembles results for "CalTRACK models with matched controls" which were cited as performing well."
    - **Response**: The GRIDmeter method was not tested in this assessment, however DSA agrees with Recurve's statement that it is similar in approach to CalTRACK with matched control groups. Our understanding of GRIDmeter methods is that the matching variables align quite closely with those that were tested in this study[24], however we found that Euclidian distance matching and propensity score matching both performed substantially better than stratified sampling when testing matching approaches.
    - PG&E adds that GRIDmeter is a term that is trademarked by Recurve.

[24] https://github.com/recurve-methods/comparison_groups/blob/master/Study_Plan_COVID_19_Comparison_Groups.pdf

- "We would like to see comparison groups as a default approach for California. While DSA notes in the report that it presents challenges of non-participant privacy, methodological transparency, and complexity, we find these challenges fully "overcomeable" in the interest of scaling distributed energy resources to contribute to the grid in a meaningful way. With open-source methods and code (like the FLEXmeter suite), it can be done with consistency and transparency. With investment in operationalizing secure data transfer protocols, it can be done with frequency and minimal complexity. Within existing regulations and with new methods of differential privacy it can be completed with full consideration and respect for customer privacy."
    - **Response**: DSA takes no position on this comment as the scope of our study was specific to determining the accuracy and precision of population NMEC methods, rather than the larger regulatory and technological landscape.
    - PG&E notes that FLEXmeter is a term that is trademarked by Recurve. PG&E has used comparison groups to calculate claimable savings for programs using the Population NMEC methodology since their inception. These methods are included in M&V plans approved in 2016 (see Advice Letters for Res P4P and OBF-AP 3698-G-A/4813-E-A and 3697-G/4812-E, respectively). Using synthetic controls overcomes customer data privacy challenges without impacting methodological transparency. It mitigates the risk of data loss and misuse of personally-identifiable information by making it unnecessary to transfer large volumes of customer energy usage data on an ongoing basis. Their applications are well documented and tested in public health and other academic literature.

THE NEED FOR COLLABORATION TO INNOVATE ON METHODS DEVELOPMENT TO DRIVE SCALE

- "While we agree with the academic position that "many models can perform accurately and with a high degree of precision" it is critical that any model is applied consistently and transparently and it ignores the value of standards (which include things like agreed upon conventions like selecting a weather station) to drive more of a weights and standards model for measurement and verification than a customized treatment for each situation."
    - **Response**: The intention of this comment is not clear. It is the position of DSA that if a model is certified to produce accurate and precise estimates of energy savings, and it is transparently documented and replicable, it may be used. It may be of value to provide a standard set of weather stations[25] or other conventions that aid vendors, utilities, regulators and implementers to use, but adoption of these tools are not required. The accuracy assessment report produced by DSA attempted, to the best of its ability, to rely on the weather station mapping method described by the CalTRACK algorithm.
- "A certification process may constrain innovation more than collaboration in an open-source process depending on who "holds the reigns" on a final review. The CPUC and CEC, for example, initiated the development of CalTRACK, and all experts and

[25] Such as CALMAC's historical weather records, http://www.calmac.org/weather.asp

stakeholders are welcome to utilize and contribute to the OpenEEmeter code base. We believe that fostering the collaborative process, which started over ten years ago, is where stakeholders and experts can have the most significant collective impact to continue innovation. However, we wholly support the principles focused on reproducibility and transparency. We find that transparency is best effectuated in an open-source governance process driving products accessible to the market rather than a consultant-led certification process."

- **Response**: We agree that transparency and collaboration are principles that should form the basis of any further work on population NMEC methods. Our recommendations are specifically designed to inform such a collaborative process and suggest that a third party such as CALMAC or the National Labs be responsible for certification.

SPECIFIC RECOMMENDATIONS AND ADDITIONS TO IMPROVE THE FINAL REPORT

- "The report provides detail on the outcomes of many of the options reviewed as part of the "tournament" analysis but did not include results or acknowledgment of the existence of the GRIDmeter methods."
  - **Response:** The GRIDmeter method was not suggested as a matching method in our stakeholder feedback surveys. DSA attempted to replicate the methods described in Recurve's Report 'Comparison Groups for the COVID Era and Beyond', which included various matching methods and matching variables.
  - PG&E adds that GRIDmeter is a term that is trademarked by Recurve.
- "The study included no hourly assessments r K/S T-test for the accuracy for the control groups studied."
  - **Response:** Individual hourly r K/S T-tests were computationally challenging to compute for the many bootstrapped iterations of customers across this assessment. Measures of accuracy (% bias) and precision (normalized root mean squared error) were produced both in and out of sample. Our recommendations and findings are based on out-of-sample statistics as these are the most relevant to the validity of post-treatment savings estimates.
- "As noted in the study, "Synthetic controls may perform well but are subject to extreme bias in many cases," and the report should also specify that cautions that must be taken include:
  - The need to generate a wide variety of load shapes or they will not have good accuracy. "
    - **Response**: We do not consider the number of load shapes developed for this method to be substantially more computationally complex than the analysis required to produce any other method that includes a comparison group of some kind. The benefit of synthetic controls is that individual non-participant data does not have to be shared, and implementers and other stakeholders can validate impact estimates produced by others because they will have access to all data needed to reproduce the estimates. PG&E has operationalized this approach. We

have changed the wording in the report to state that estimates produced using synthetic controls are highly sensitive to the choice of segmentation used.

- "That company policies severely limiting data access, though allowed via Commission rules, may undermine performance assessment"
  - **Response**: DSA is not in a position to draw this conclusion as it was not part of the scope of our assessment.
  - Commission rules governing customer data access are more complex than this statement suggests. California consumers hold additional rights over the use of their data under the provisions of the California Consumer Privacy Act of 2018 (CCPA).

7 REVIEWER COMMENTS: SOUTHERN CALIFORNIA EDISON COMPANY (SCE)

Comments from Galib Rustamov at Southern California Edison

1. "Considering that the suggested methodologies are not applied in the experimental context, it is no surprise that comparison group analysis reduces the magnitude of the bias and confounding effects born out of self-selection, site selection bias, etc."

   a. **Response**: DSA did replicate self-selection. Participants in this study were selected from prior energy efficiency program participants, while the pool of customers for the comparison groups were drawn from the general population. Matching was done on pre-treatment data and no post-treatment matching occurred.

2. "Comparison groups should be considered regardless of COVID or COVID-like incidences in population-level NMEC or wherever feasible."

   a. **Response**: We agree.

3. "One of the primary challenges is that the comparison groups are often identified after customer participation (or program launch) and their similarity to customers in the "treatment" group. Although there are some matching techniques available, which ones to use would be a question for technical staff. For instance, DSA only utilizes only a few of them. Then, the efficacy of the design may depend on the creativity of the evaluator. Therefore, to mitigate some of these challenges, it would be helpful to determine the pool of comparison group sites (non-participant) ahead of time (pre-implementation) and be clear on the approach and assumptions are used to determine the comparison group. The results will only be as good as the sample and data from which they are obtained."

a. **Response**: Matching methods tested were developed in conjunction with stakeholder feedback. Additional methods were not suggested by stakeholders surveyed in February of 2021. All certified methods must meet bias and precision targets, meaning that all can accurately estimate program savings. Vendors and implementers are free to select the most efficacious of method within the pool of certified methods, however all certified methods have demonstrated performance. We will adjust our recommendations (#10) to propose having the NMEC method used for estimation of savings be pre-certified prior to program implementation rather than impact estimation

4. "Evaluators and implementers should be clear on the assumptions when developing the (propensity score, exact, nearest neighbor, radius, kernel) matching approaches. These approaches may provide improvements in estimating the savings impact. However, identifying the variables that determine the likelihood a unit receives the "treatment" or site participates in the program must be clear and consistent."

    a. **Response**: We agree. Our recommendation requires that the code used to produce the matching method and estimation procedures be publicly available (#9) and replicable (#2, #3)

5. "Additionally, we should be careful when we standardize a method over the others to allow creativity. (The methodologies are based on the scientific research and can be updated in the future.) The usefulness may depend on the situation, context, and data quality. However, minimum requirements should be transparent to the implementers, evaluators, and practitioners."

    a. **Response**: We agree. Recommended principles for population NMEC methods, including minimum requirements and transparency practices, are described in Section 5.1 of the report

6. "Finally, I think the research provides limited discussion on the sample sizes, power and precision. In terms of sample sizes what are general observations from the previous studies that DSA reviewed? Does the methodology or its quality varies by the sample sizes (of both comparison and treatment groups)? I would assume small sample cases are subject higher technical scrutiny."

    a. **Response**: A discussion of sample sizes and their effects on each method can be found in Section 4.7 of the report, where we have added some additional discussion of sample size, margins of error, and statistical power. Recommendation #6 includes accuracy and precision of certified methods to meet certain portfolio bias and normalized RMSE targets based on sample size.