Reliability Demonstration of A Novel Randomized Measurement and Verification Method for Switchable Control Retrofit Using Large-scale Public Dataset

Aoyu Zou [[1]](#footnote-20), Paul Raftery 1, Stefano Schiavon 1, Carlos Durate 1

Conventional measurement and verification (M&V) methods relying on pre- and post- retrofit comparison for estimating whole-building energy savings are often time-consuming and unreliable, especially when non-routine events such as reduced occupancy occur during the M&V process. Those events are unrelated to the intervention but significantgly affect building energy consumption and thus when the analyst applies conventiona M&V suggested by industry guidelines, the results can be largely confounded. In this study, we demonstrated that switchable interventions, such as most of the control retrofits, can benefit from random sampling, where the analyst randomly selects whether to implement the baseline or the intervention strategy each day. We tested this novel randomized M&V method on a large public dataset covering various climate zones and commercial building types, using a virtual chilled water supply temperature reset based on outdoor weather as the intervention. Our results show that comparing to the conventional method, the randomized method provides faster and more accurate savings estimation. Additionaly we found that the randomized M&V approach estimates much closer to the true savings when non-routine events are present demonstrating much improved reliability.

# 1 Introduction

## 1.1 Background

### 1.1.1 Conventional M&V

Measurement and Verification (M&V) is the process of quantifying energy savings from energy efficiency projects by comparing actual energy consumption against a baseline, adjusting for factors like weather and occupancy. This process ensures that improvements in energy performance are accurately evaluated. In the United States, practitioners often refer to ASHRAE Guideline 14, the International Performance Measurement and Verification Protocol (IPMVP), and the Federal Energy Management Program (FEMP) for standard guidelines ASHRAE (2014). These guidelines outline standardized methods for quantifying energy savings, whether through calibrated simulations or monitored measurements for specific equipment or systems (isolation methods) or for entire buildings (whole-building methods). In this study, we will focus on the energy savings quantified at the whole-building level where the measurements are obtained from utility bills or whole-building meters. The corresponding M&V process required by ASHRAE Guideline 14, which we refer to as the convention method in this paper. Typically, the process begins with baseline measurements taken over a year before implementing any energy-efficiency retrofit, followed by the same measurement procedure during the intervention period. After collecting two years of data, an M&V analyst fits an energy prediction model, using variables such as outdoor temperature and time (Mathieu et al. 2011) to project baseline energy consumption in the post-retrofit period. The difference between the counter-factual baseline and the measured intervention represents the energy savings. A key drawback of this method is its reliance on a two-year timeline to quantify savings, during which baseline measurements can become outdated due to changes in building performance caused by non-routine events unrelated to the intervention. This limitation reduces the feasibility of rapid M&V and complicates the quantification of estimation uncertainty, thus impacting the accuracy and timeliness of savings assessments.

### 1.1.2 Randomized M&V

To address the limitations of conventional M&V methods and the challenges posed by non-routine events, we propose a novel M&V method that adopts the randomized crossover design, a concept borrowed from medical and agricultural studies. Another improvement is that we proposed a sequential evaluation framework and defined stopping criteria to end the M&V if the target effect is detected. This is to avoid unnecessary measurement collection over the full 2-year M&V cycle. The full framework is detailed in another study with all stopping criteria outlined . In summary, this method provides M&V analysts with a randomized schedule that alternates between baseline and intervention implementation while ensuring balanced sampling across days of the week. For example, given a 10-week M&V period for 1 intervention, the balanced randomized schedule would equally sample 5 Mondays with the baseline and 5 Mondays with the intervention. The limitation of the randomized M&V is that it is only applicable to a subset of retrofit projects such as control interventions. However, for all applicable use cases, it allows analysts to detect energy savings sequentially as the study progresses meaning once the desired savings target is achieved, analysts can terminate the M&V. The key advantage of randomization, which is one of the study objectives, is that if control strategies are sampled with equal probability, the influence of non-routine events is likely to be evenly distributed between the baseline and intervention measurements. This means that the savings estimate, calculated as the difference between the two, effectively ‘cancels out’ the impact of these disturbances, leading to a more accurate and unbiased assessment of the intervention’s impact.

### 1.1.3 BDG2 dataset

The Building Genome Dataset 2 (BGD2) is an extensive open-access dataset designed to advance research and development in building energy efficiency and control strategies acting as a test-bed for modeling, simulation, and algorithm development (Miller et al. 2020). BGD2 contains over 500 buildings’ metadata and realistic operational information from across North America and Europe, making it one of the most comprehensive collections of building-related data available for scientific use. The dataset includes various commercial building types such as offices, education facilities, public, and retail buildings, and provides detailed information on their physical characteristics (e.g. enregy ratings, heating types and floor area) and hourly measurements of chilled and hot water, electricity, gas usage as well as site outdoor weather conditions.

For the purpose of this study, we only queried whole-building hourly electricity usage. All building names and precise locations were erased but the climate zone and city name were provided, which are shown in Figure 2.2 and Figure 2.3. In addition, the measurements were pre-processed by the authors with timestamp already converted to local time. We further described the data filtering process in Section 2.

## 1.2 Literature review

### 1.2.1 Measurement and verification energy-efficient measures

Most research related to M&V for whole-building approach focuses on the accuracy of baseline modeling, exploring model performance from simple regression models to more complex machine learning techniques. One study reviewed various models suitable for M&V applications as well as selected input features (Alrobaie and Krarti 2022) and another study provided a definitive methodology to apply machine learning models for M&V use cases (Gallagher et al. 2018). In addition, a few studies investigated the critical performance metrics to evaluate the developed baseline models Granderson and Price (2014) and compared a variety of models using those metrics (Granderson et al. 2016). These studies made significant contributions by emphasizing the uncertainty associated with the model-fitting process, a key factor in accurately determining energy savings. Furthermore, other researchers addressed this issue by leveraging statistical formulation and inference to improve baseline energy models Walter, Price, and Sohn (2014). However, a gap still remains in the literature regarding the rigorous quantification of uncertainties directly associated with calculated savings, for instance, accounting for the potential bias that baseline model might deteriorate (i.e. becomes ‘stale’) over an extended period of pre- and post-analysis.

### 1.2.2 Impact of non-routine events on building energy usage

A common non-routine event in energy-saving M&V projects is a change in occupancy or a significant shift in occupant behavior. These changes can greatly affect measured energy consumption in buildings and are typically unrelated to the intervention strategy. For instance, during the COVID-19 pandemic in 2020, most commercial buildings were unoccupied, leading to a noticeable drop in energy bills despite no energy-efficiency measures being implemented Kang et al. (2021). Furthermore, subsequent research has shown that hybrid working modes, allowing employees to work remotely, have persisted after the outbreak of the pandemic (Aksoy et al. 2022), adding further complexity to energy consumption patterns due to evolved occupant behaviors with higher flexibility Xie et al. (2021). One study realized the limitation of current M&V methods, which only consider adjusting for outdoor weather, is insufficient and emphasized the importance of requiring matched comparison groups to control for exogenous factors beyond weather differences when comparing between baseline and intervention (Demand Side Analytics 2022). Another type of non-routin event is filter clogging in air handling units due to particle accumulation. This can cause supply fans to gradually consume more energy to maintain required duct static pressure Zhai and Nathaniel Johnson (2017). If M&V analysts are unaware of such changes and lack an appropriate adjustment method (e.g., replacing filters before the intervention begins), the savings could be underestimated as increased energy use is incorrectly attributed to the intervention rather than the mechanical issue.

### 1.2.3 Randomized experimental design in other scientific fields

To address the research gap identified in existing M&V studies, particularly the challenges posed by non-routine events, we applied a randomized experimental design method—an approach commonly used in other scientific disciplines. In clinical trials, for instance, randomization is employed to determine the effect of a medical treatment (such as a new drug) by randomly assigning participants to either a control group (e.g., receiving a placebo) or a treatment group Wiley et al. (2016). The primary purpose of randomization is to block confounding variables such as age, gender, and socioeconomic status when evaluating treatment effects. In the context of buildings, a “treatment” refers to a control retrofit, while the “placebo” represents the existing baseline control. Unlike clinical trials, where participants are randomly assigned, our approach randomizes treatment assignment longitudinally for each individual building—an approach known as n-of-1 trials (Gabler et al. 2011). Although the BDG2 dataset allows for randomization at a population level (e.g., across climate zones), building owners are generally more interested in understanding the effects of control retrofits specific to their own buildings, rather than at a generalized population level.

## 1.3 Objectives

As mentioned, the goal of an M&V project is to determine the effect—typically energy savings—of an energy-efficient intervention. In this study, we focus primarily on switchable interventions, which often involve control retrofits. An example of such an intervention is a control retrofit developed by a software-as-a-service company that adjusts the chilled water plant’s supply water temperature based on outdoor weather conditions Qiu et al. (2022). Therefore, we defined the M&V scenario as follows:

*“A company aims to sell its supply temperature reset control software package to a customer, such as a building owner, with a guarantee that it will reduce the building’s electricity usage. If the building owner decides to purchase the service, the company agrees to charge a service fee based on a percentage of the measured energy savings.”*

As required by the M&V scenario, we assessed the performance of both the conventional and the novel randomized M&V methods by estimating the intervention energy savings for all valid buildings in the dataset. By conducting such analysis, we aim to:

1. Compare the energy saving estimation accuracy between the conventional and the randomized method. This study extends the comparison to a large sample of buildings, covering a variety of types and climate zones. The comparison metrics include both estimation accuracy and M&V finishing timeline.
2. Verify the enhanced robustness of the randomized method. By using realistic measurements from real-world buildings, which include various sources of noise, we aim to reflect the challenges faced by building analysts in real projects. As will be demonstrated in the following sections, the randomized approach is less impacted by non-routine events (i.e., measurement ‘noise’), resulting in more reliable energy savings estimates.
3. Demonstrate the implementation of the proposed randomized M&V method using a public available dataset. We ensured the reproducibility of the method by making the analysis code open source including randomized schedule generation, sequential statistical analysis, energy modeling and normalized saving calculation. Using the available open resources, building analysts should be able to seamlessly integrate and apply them in their own M&V projects.

# 2 Method

We outlined the methodology of the study in Figure 2.1 and extended several key components in this section.

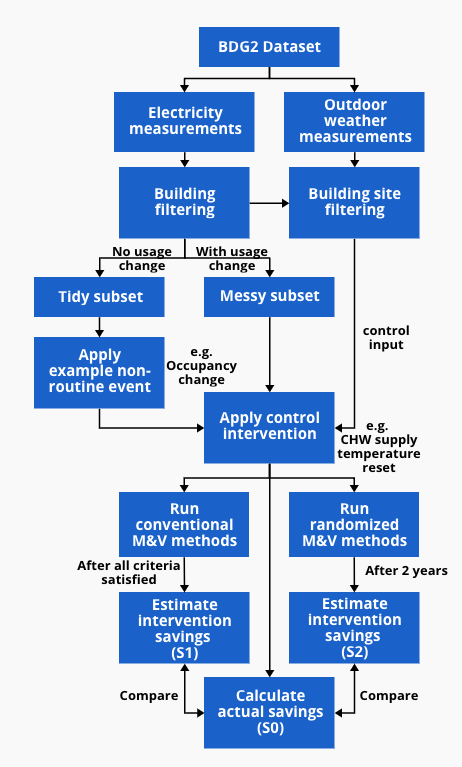


Figure 2.1: Flow chat showing the methodology for comparing the estimated savings of randomized M&V with the conventional M&V

## 2.1 Building filtering

### 2.1.1 ‘Tidy’ subset

In this study, we extracted the electricity measurements from the BDG2 dataset. On a first pass, we filtered out buildings with less noise based on the following criteria:

1. Missing values < 1000: given the hourly resolution of all measurements, this is equivalent to 1.5 months of missing days.
2. Mean electricity usage > 0 kWh: target buildings should have active electricity usage.
3. No statistical significant difference (P-value > 0.05) between the two-year electricity usage: target buildings should have no change in the electricity usage between the two years.
4. Target buildings should have known site location: buildings with anonymous location are excluded due to unavailable typical meteorological weather.
5. Warehouse and parking types are excluded: target buildings have less demand flexibility to implement a chilled water setpoint reset control.

Therefor the resulting subset contains all buildings with ‘tidy’ measurements. Figure 2.2 shows in total, the subset contains 66 buildings of 6 types.

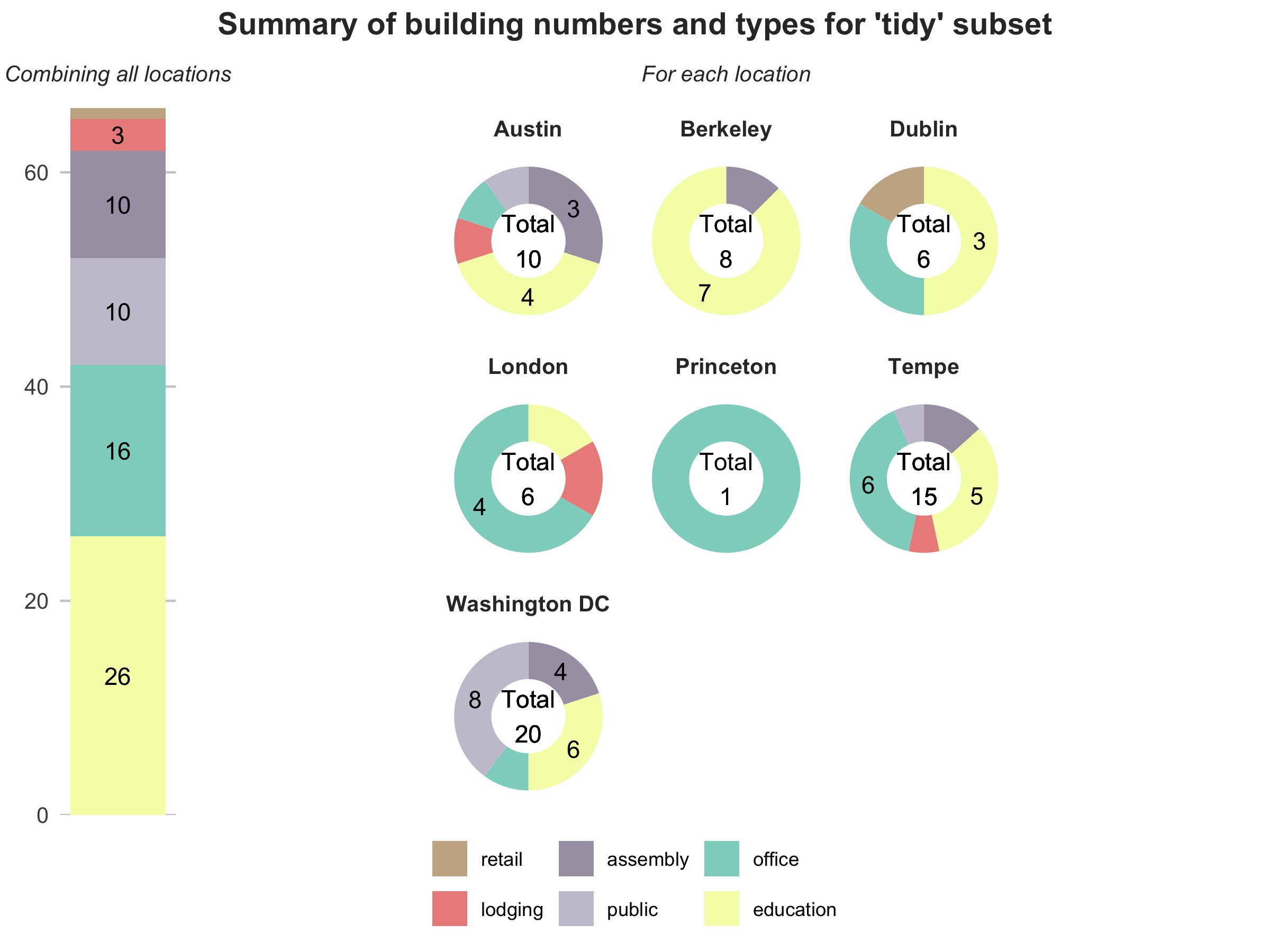


Figure 2.2: Site summary of the tidy building subset (counts < 2 are omitted for visualization; left: aggragated counts of buildings for each type; right: breakdown building counts for each building type at each location)

### 2.1.2 ‘Messy’ subset

In reality, the ‘tidy’ subset is less representative of the measurements typically collected from existing building stock, as whole-building electricity measurements collected over two years often show more variability the the filtering critieron. These variations can stem from issues like sensor calibration errors or inherent changes in the building, which, as previously discussed, can bias M&V results. Therefore, to more realistically assess the robustness of the two M&V method, we included an additional ‘messy’ subset which first exclude the ‘tidy’ subset and then re-apply the filtering rule with one amendment:

1. : any increase or decrease of building electricity usage in the second year should be less than 25% of that in th first year.

Figure 2.3 shows the summary of the ‘messy’ dataset, which contains 573 buildings in 12 types.

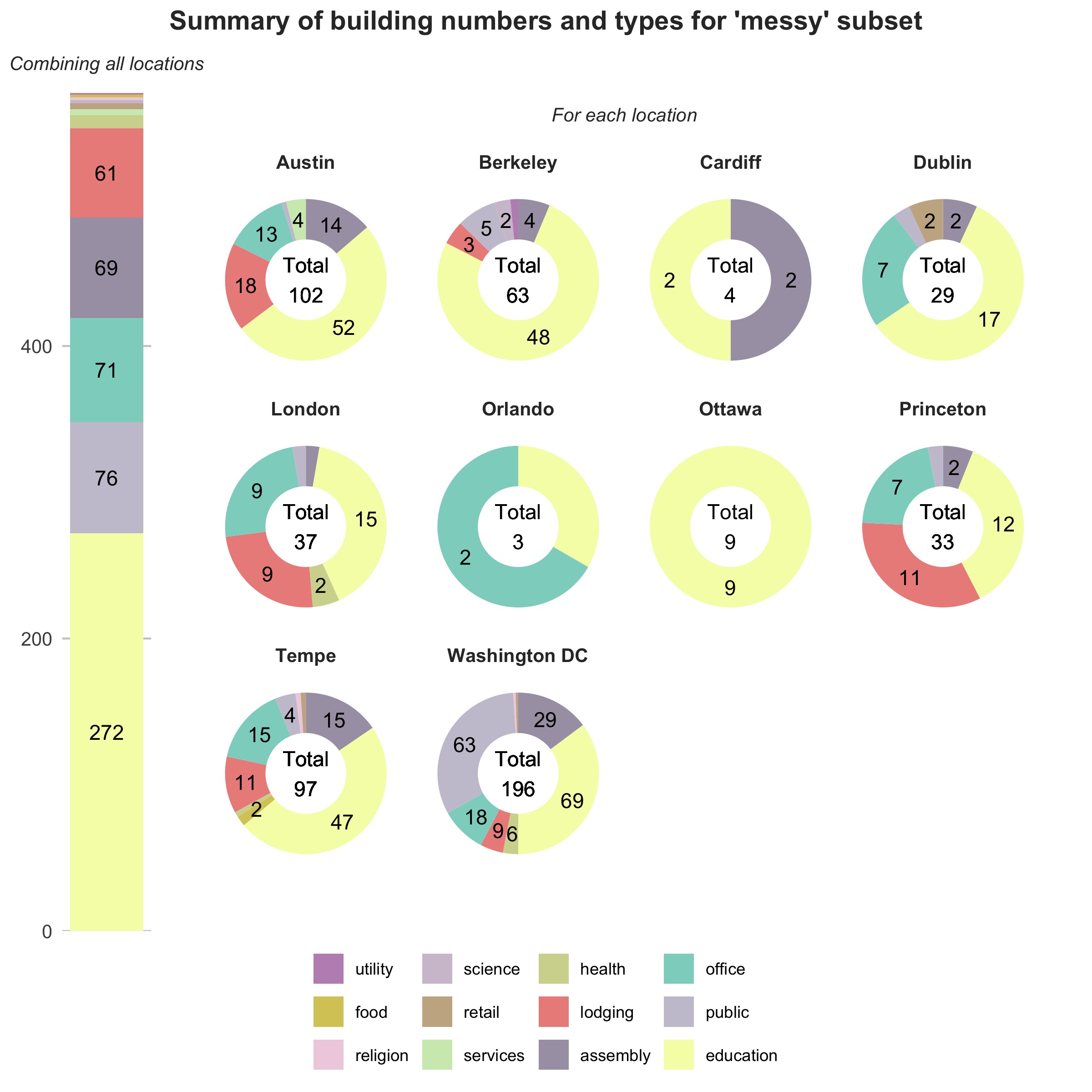


Figure 2.3: Site summary of the messy building subset (counts < 2 are omitted for visualization; left: aggragated counts of buildings for each type; right: breakdown building counts for each building type at each location)

## 2.2 Apply control intervention

Figure 2.4 shows the algorithm for the proposed control intervention that reset the chiller supply temperature based on the outdoor weather conditions, which can be commonly found in the literature Čongradac and Kulić (2012). For both strategies, we assume that the chiller is activated when the outdoor temperature exceeds 10°C. The baseline strategy, representing the existing measurements from the dataset, operates with a constant water supply temperature. The intervention strategy, as illustrated in the figure, adjusts the water supply temperature dynamically, resetting it from 7°C to 12°C.

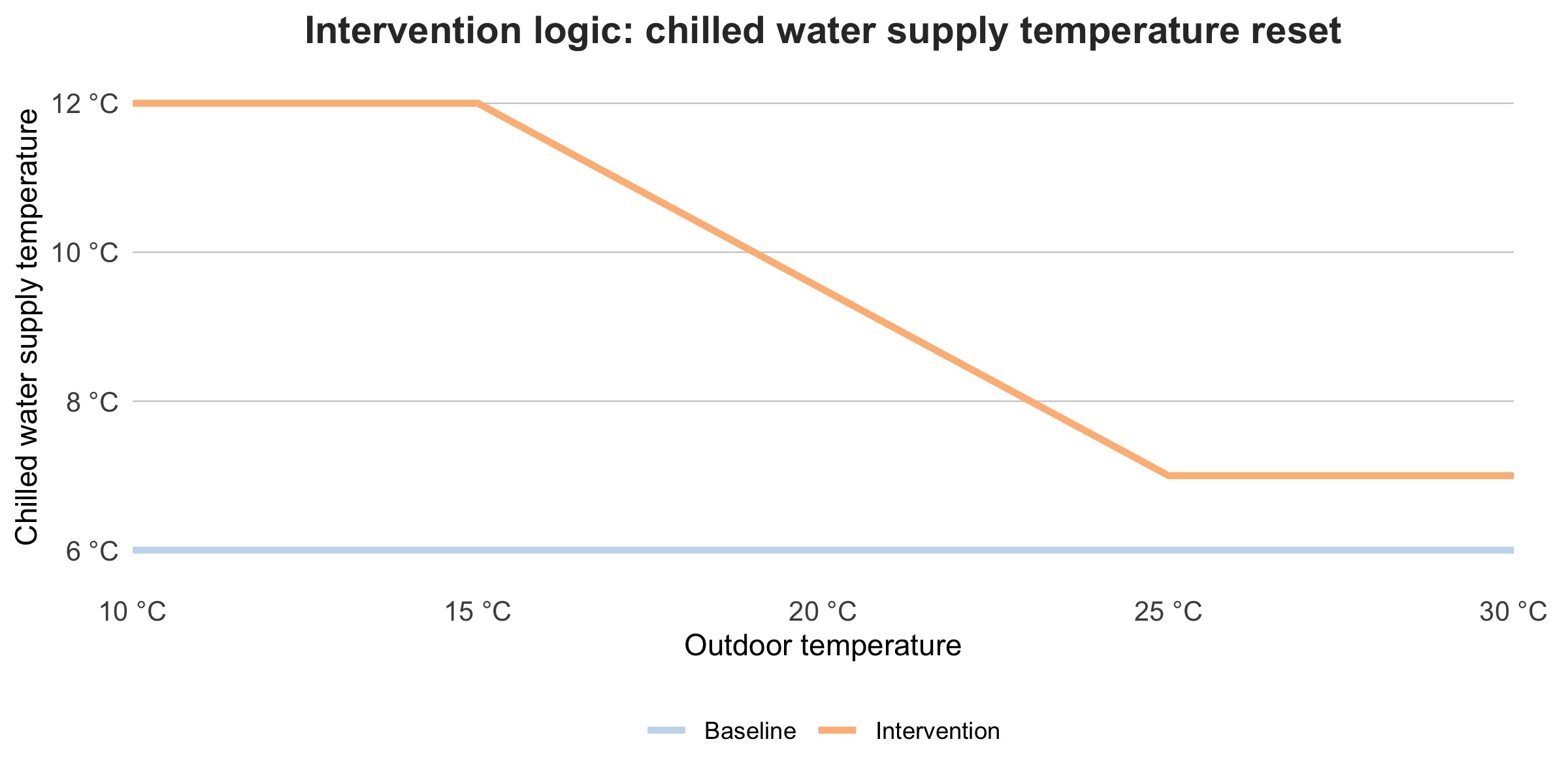


Figure 2.4: Proposed intervention strategy: chilled water supply temperature reset based on outdoor temperature

We mapped the chilled water supply temperature reset to the electrical energy savings as:

We assume on average, HVAC systems account for approximately 50% of a building’s total electricity consumption, and the chilled water plant further consumes around 50% of the HVAC electricity. While this assumption largely simplifies the diverse energy usage across various building types, for the scope of this paper, we assume that 25% of the total building electricity is used by the chilled water plant, (Administration 2012). Typically, the savings from an intervention are not proportional to the building’s hourly electricity usage, which is generally the challenge for M&V. To address this, we mapped the resulting electricity savings as a percentage of the plant’s normal operation, calculated as its mean electricity usage over the two-year period. This percentage is influenced by factors such as outdoor temperature (), intervention supply water temperature (), baseline supply water temperature () and hour of the day (, binary indicator whether during peak hours from 9 AM to 4 PM).

Parameters and their pre-defined values are summarized in the table below. For simplicity, those parameters were not rigorously calibrated for each building, and were used uniformly across the dataset.

Table 1. Parameters for calculating the intervention savings.

|  |  |  |
| --- | --- | --- |
| Parameter | Description | Value |
|  | % savings from setting 1 °C higher than | 0.08 |
|  | % savings adjustment during occupied hours | 1.2 |
|  | % savings adjustment during unoccupied hours | 0.8 |

## 2.3 Apply example non-routine event

To illustrate the impact of non-routine events that are not properly adjusted for, we developed three scenarios to quantify the energy consumption changes associated with a hypothetical increase in occupancy. Specifically, we assume that a rise in occupancy in 2016 (referred to as the “baseline year” when using the conventional method) led to a 20% increase in whole-building measured electricity usage. The three scenarios describe when this increase occurred: January to April (S1), May to August (S2), and September to December (S3). This hypothetical change was applied only to buildings in the “tidy” subset, as these buildings were initially more likely to be free from non-routine events.

## 2.4 Run M&V methods

### 2.4.1 Conventional M&V

As described in Section 1, conventional method for M&V is time-consuming and unable to separate non-routine event impact from measured savings. If using conventional method to estimate intervention energy savings, the result (after 24 months) is calculated by the difference between measured intervention and projected baseline in the post-retrofit period. In this study, we leveraged a piece-wise linear regression considering time-of-week and outdoor temperature (TOWT) as independent variables for projection for the counter-factual baseline in the post-retrofit period and normalization on typical meteorological year (Mathieu et al. 2011).

### 2.4.2 Randomized M&V

Compared to the conventional method, the randomized M&V approach offers a more rapid and reliable estimation of energy savings. To apply this method, analysts first define the target savings and stopping criteria. Then they design a randomized switchback schedule and perform sequential statistical tests, such as the sequential probability ratio test (SPRT), to monitor savings as data is collected. Once the target savings are detected, the analysts fit a prediction model (e.g. TOWT) to adjust for differences in outdoor temperature, following the same adjustment process as the conventional method. We provide example switchback experimental design and stopping criteria thresholds below:

* The HVAC system operates from 06:00 to 22:00 each day, so we use a daily sampling interval with the sampling time at midnight each day.
* Block by day of the week with a block period of 12 weeks.

Stopping criteria are:

* A minimum and maximum of 12 and 108 weeks respectively. The randomized schedule covers the entire two-year period but stopping criteria enables an early stop at the end of satisfied blocking period.
* At least 80% of the drybulb temperature range in the annual TMY data sampled by both strategies.
* Test for no carryover effect using a t-test with a p-value not exceeding a defined significance threshold of 0.05.
* 90% confidence that energy savings exceed or do not exceed 0% using the SPRT test. Medium effect size (d = 0.5) quantified by cohen’s d and calculated SPRT statistics either falls below the lower threshold or exceeds the upper threshold.
* As no baseline data is available, test with an equal sampling ratio (50% baseline, 50% intervention).

To remain consistency, savings normalization on typical meteorological year is also modeled through TOWT.

# 3 Results

## 3.1 M&V methods comparison

In this section, we compare the performance of two M&V methods. The key aspects of the assessment include: 1) time required to reach a saving estimation: in most cases, a shorter M&V timeline reduces associated cost and interruption for the building owner; and 2) saving estimation accuracy: this is particularly important for any software-as-a-service company to set a reasonable price with their customers.

### 3.1.1 Savings estimation time

ASHRAE Guideline 14 offering minimum requirements for whole-building measurement path states that the baseline period is either a full range of all independent variables (typically outdoor weather conditions) under normal facility operation or a 12-month worth of continuous measurements. The same requirements also applies to intervention installed in the post-retrofit period. Thus, the conventional M&V method is likely to run over 24 months or even longer due to missing data. For example if an analyst is asked to determine the energy savings of a chilled water plant retrofit but missed most of the cooling season due to delay in retrofit deployment, he/she needs to wait until the next cooling season to measure the savings. If the randomized M&V method is applicable, delays in retrofit deployment pose less risk to the building owner since baseline are continuously monitored. Sampling at equal probabilities (e.g., 50%/50% between baseline and intervention) helps to balance the distribution of independent variables like outdoor weather conditions across all sampled control strategies. As long as one strategy is not sampled over a large number of consecutive days (e.g., 7 days or more), it is unlikely that the measurements of independent variables will differ significantly between the control strategies.

Figure 3.1 shows the average estimated timeline for each building across all sites and climate conditions if using randomized M&V for the applied chilled water supply temperature reset intervention. The timeline figures for each individual building are attached in the supplementary material. The figure shows all buildings can detect a saving statistically from sampled measurements within 12 months, which is equivalent to the minimum test criterion. Similar to the findings from our pervious study, covering a sufficient range of outdoor weather condition is the most stringent requirement. If target buildings are located in a climate zone with more preferable outdoor weather conditions such as California, covering 80% of the TMY range should only take 3 ~ 4 months. In this case, the analyst can conclude the M&V simultaneously if the blocking period ends at the same time. However, given most buildings requires 6 months to achieve 80% of the TMY range, all buildings can determine the savings, including associated uncertainties, within 9 months—significantly shorter than the baseline measurement period required by the conventional method.

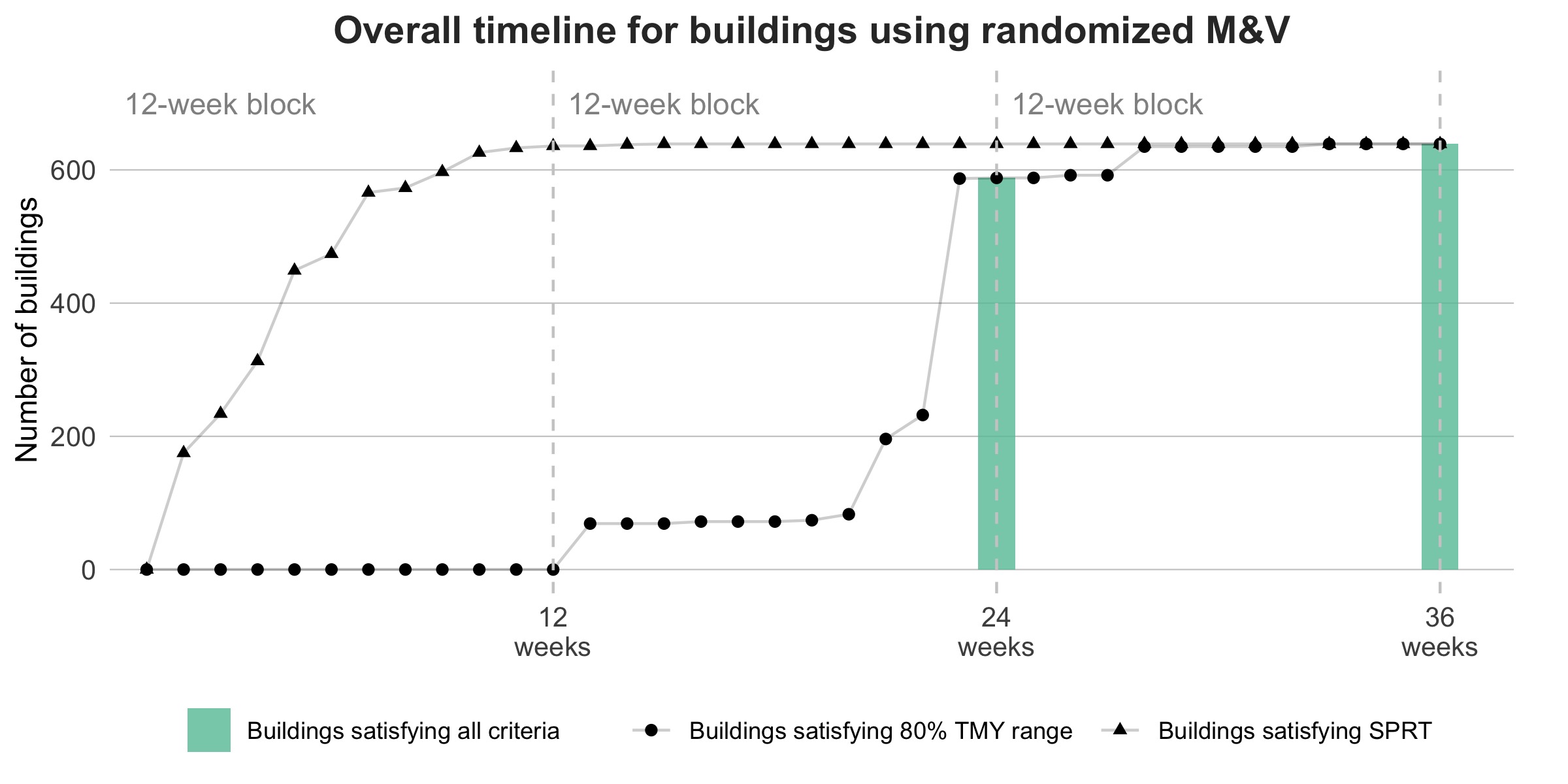


Figure 3.1: Randomized M&V timeline of satisfying key stopping criteria for all buildings

### 3.1.2 Savings estimation accuracy

In this study, we define the overall target savings (i.e., the ground truth) that an M&V analyst aims to detect as the mean reduction in electricity consumption over a two-year period resulting from the implementation of the chilled water supply temperature reset control. This target can be expressed in terms of normalized savings, fractional savings, or simply the measured difference between baseline and intervention measurements. Due to different timeline required by the two methods, we compare the conventional M&V savings at the end of the two-year period with the randomized M&V savings after all criteria are satisfied. Since those are the time in reality, an M&V analyst report estimated savings.

Figure 3.2 shows the overall comparison of the savings estimated for all buildings in both subsets normalized on the typical meteorological year weather conditions of each site. In subplot a), the narrower range and clustering of the true savings indicates the dependence of the intervention effect on the outdoor weather condition, which is intended after weather normalization. Sites with mild climate all year round such as locations in California shows higher savings potential above 10%, while locations with more extreme climate such as Washington DC shows only 6% savings annually. In subplot b), the results indicate that the conventional M&V method tends to estimate savings with greater uncertainty, and its distribution median deviates from the true savings. In comparison, the randomized method, which requires much less time, shows more accurate and precise estimation results.

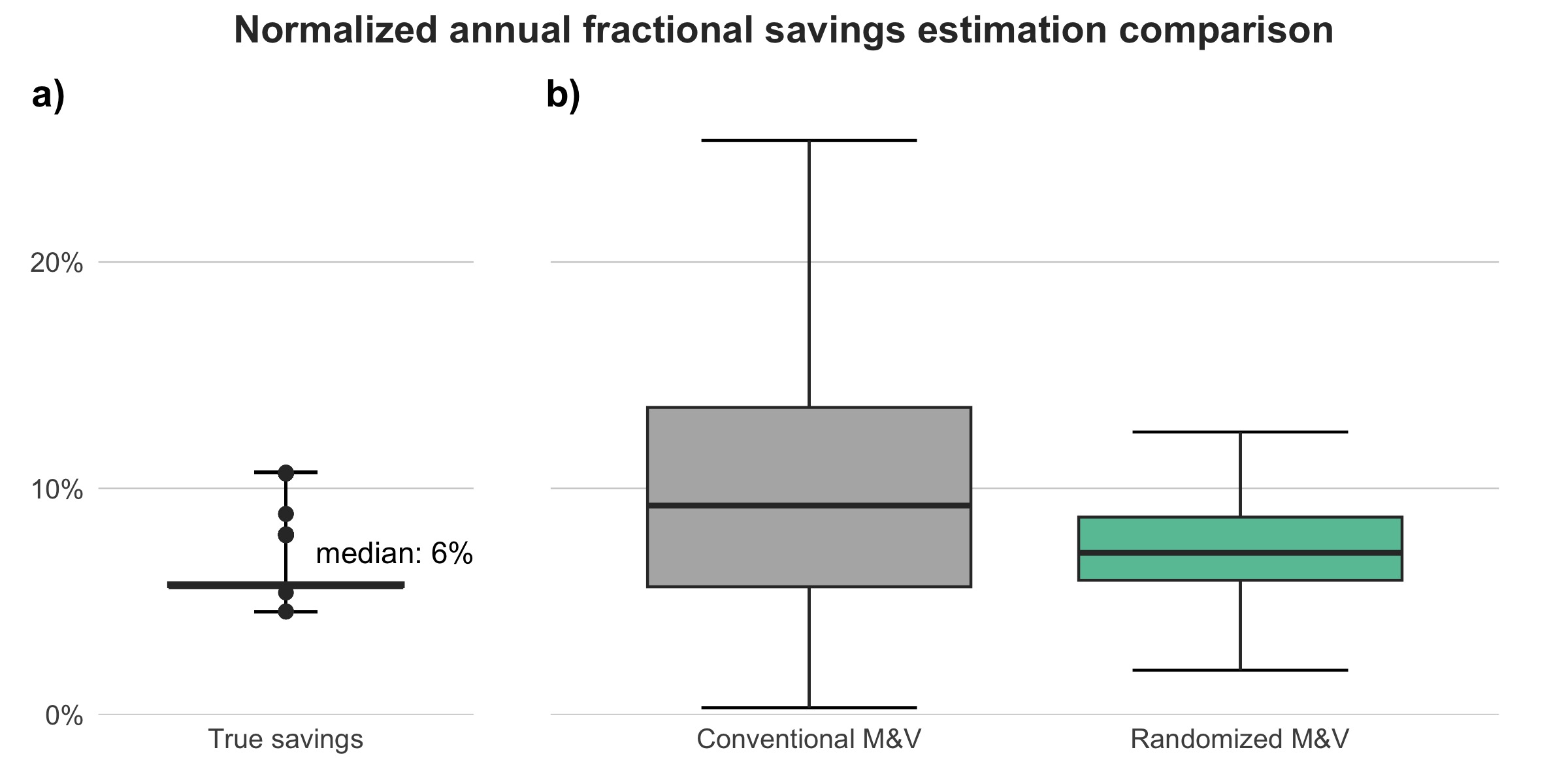


Figure 3.2: Comparison of annual fractional savings normalized on TMY weather conditions between conventional and randomized M&V method

To assess accuracy at the individual building level, we calculated the mean absolute difference between the true savings and the estimated savings and normalized the results as a fraction of measured baseline for generalized comparison. Figure 3.3 shows the comparison between the two methods using the tidy subset. In this comparison, we focused on the median and the 95th percentile of the plotted distribution to minimize the influence of outliers. The results indicate that the conventional method outperforms the randomized method in this subset. This is mostly because the tidy subset includes buildings with consistent usage patterns throughout the two-year period, meaning the data used for regression model fitting closely aligns with the conditions during the model prediction phase. In our case, this justifies the TOWT model selection when there is no additional adjustment needed for such pre- and post-analysis.

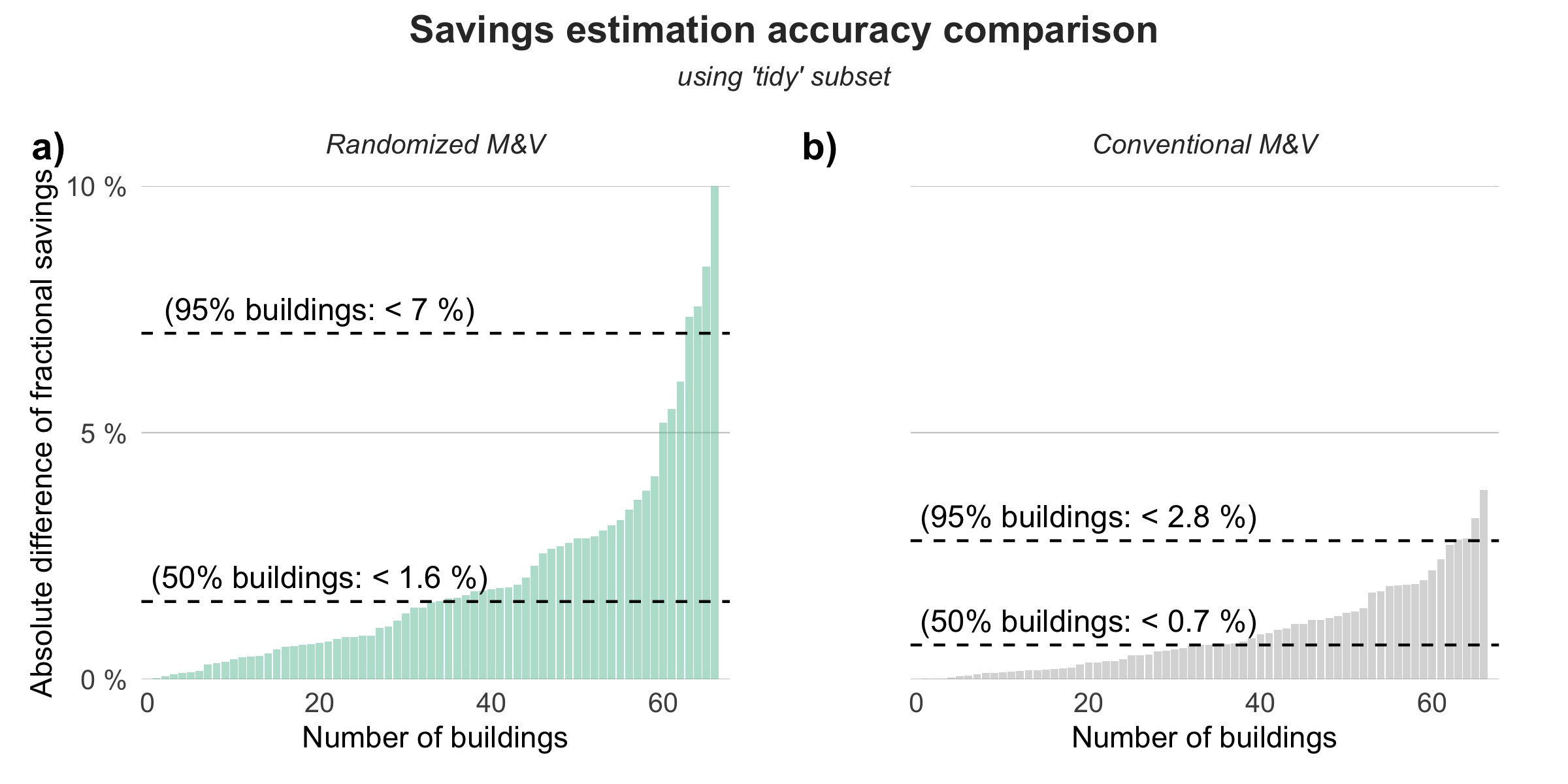


Figure 3.3: Absolute deviation of fractional savings from true target savings shown for the ‘tidy’ subset

While these findings suggest that the conventional M&V method is advantageous when a building maintains stable electricity usage over time, such stability is rarely guaranteed in real-world scenarios. A more realistic condition is shown in the messy subset in Figure 3.4. In this case, the randomized method demonstrates consistent accuracy in savings estimation, even in the presence of random measurement noise, implying strong robustness. In contrast, the conventional method shows deviations exceeding 5% in approximately 50% of all cases. Non-normalized mean absolute difference comparison plots are provided in the supplementary material. The much deteriorated performance of the conventional method highlights its vulnerability when changes in electricity usage occur. This is because baseline projections based on regression models become unreliable when non-routine events are present. Non-routine events encompass all influential factors not accounted for in the regression model, either due to their impracticality to measure—such as hourly occupancy rates—or their unpredictability—such as a sudden change in building use where an office being converted into a warehouse. We dedicate the following section to continue the discussion of the impact.

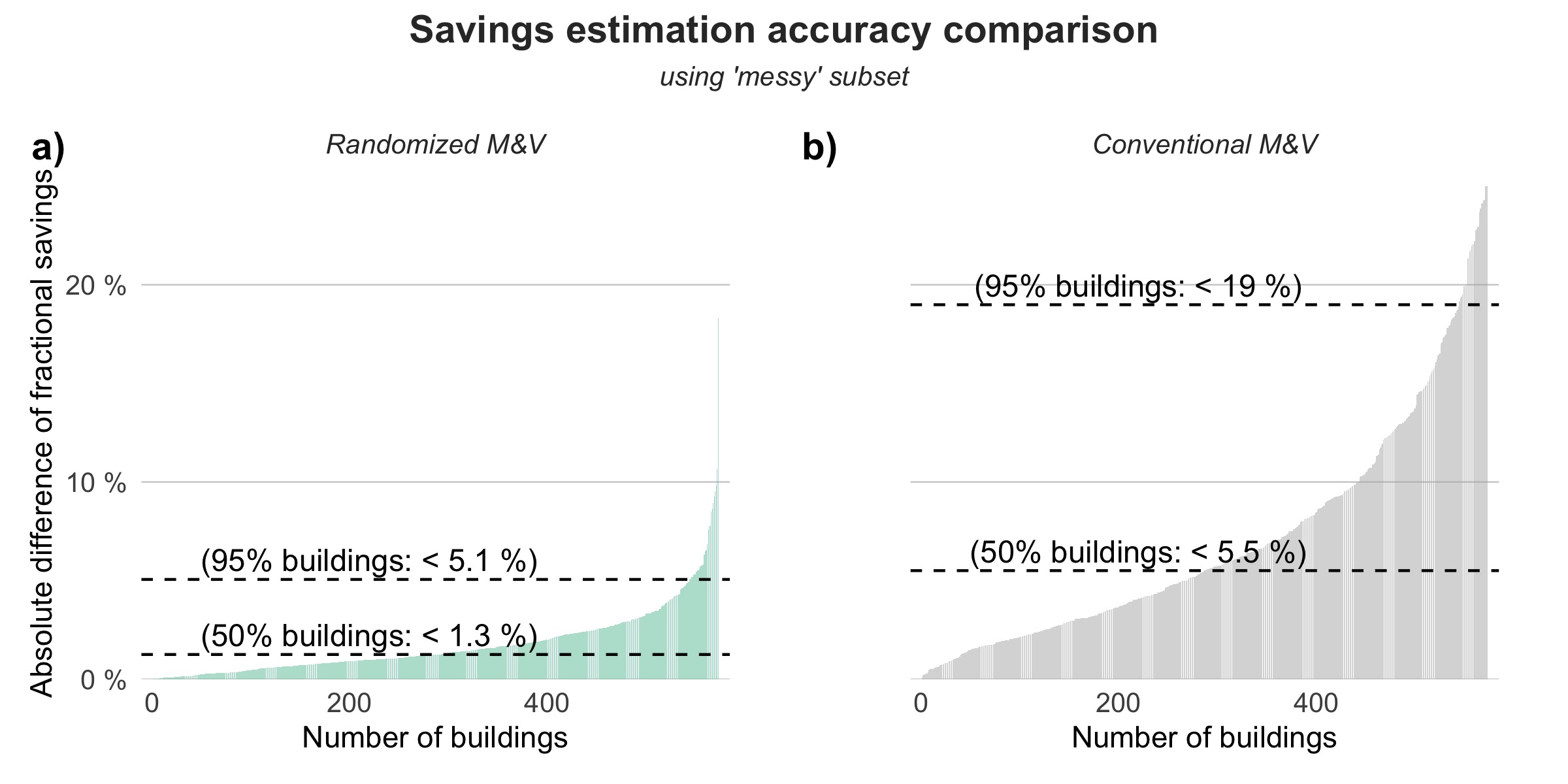


Figure 3.4: Absolute deviation of fractional savings from true target savings shown for the ‘messy’ subset

## 3.2 Non-routine events impact on savings estimation

We included two examples in this section to illustrate the impact of non-routine events on M&V savings estimation. In the first scenario, we manually subtracted a portion of the electricity usage for a specific period in the tidy subset to simulate occupancy-confounded measurements, which were not originally present in the dataset. In the second scenario, focusing on the messy subset, which already contains underlying changes, we removed the intervention effect and applied both M&V methods to assess which method could more accurately detect the resulting zero ‘savings.’

### 3.2.1 Occupancy change

Although occupancy can be estimated in various ways, such as counting check-ins, monitoring WiFi connections, or using indoor CO2 concentrations as a proxy, these methods are often not cost-effective for routine operations in most buildings and may raise privacy concerns. In the proposed scenario, we hypothesize that during the baseline measurement period, one floor of tenants vacated the building, leaving the space unoccupied for four months. This led to a fixed reduction in electricity usage, approximated as 20% of the yearly average. We showed the impact on TOWT model fitting in Figure 3.5 where from May 1st 2016 to August 31st 2016, the target building electricity dropped. Subplot (a) displays the measured baseline data, including the occupancy-related change, used as input for model training. Subplot (b) shows the model’s prediction results for the post-retrofit period. For reference, we also plotted the correctly adjusted baseline (labeled as ‘adjusted baseline’) in the post-retrofit period. Since occupancy was not considered an independent variable in the model, the baseline change was incorrectly attributed to temperature and time variations shown as the ‘projected baseline’, leading to underestimated savings estimation.

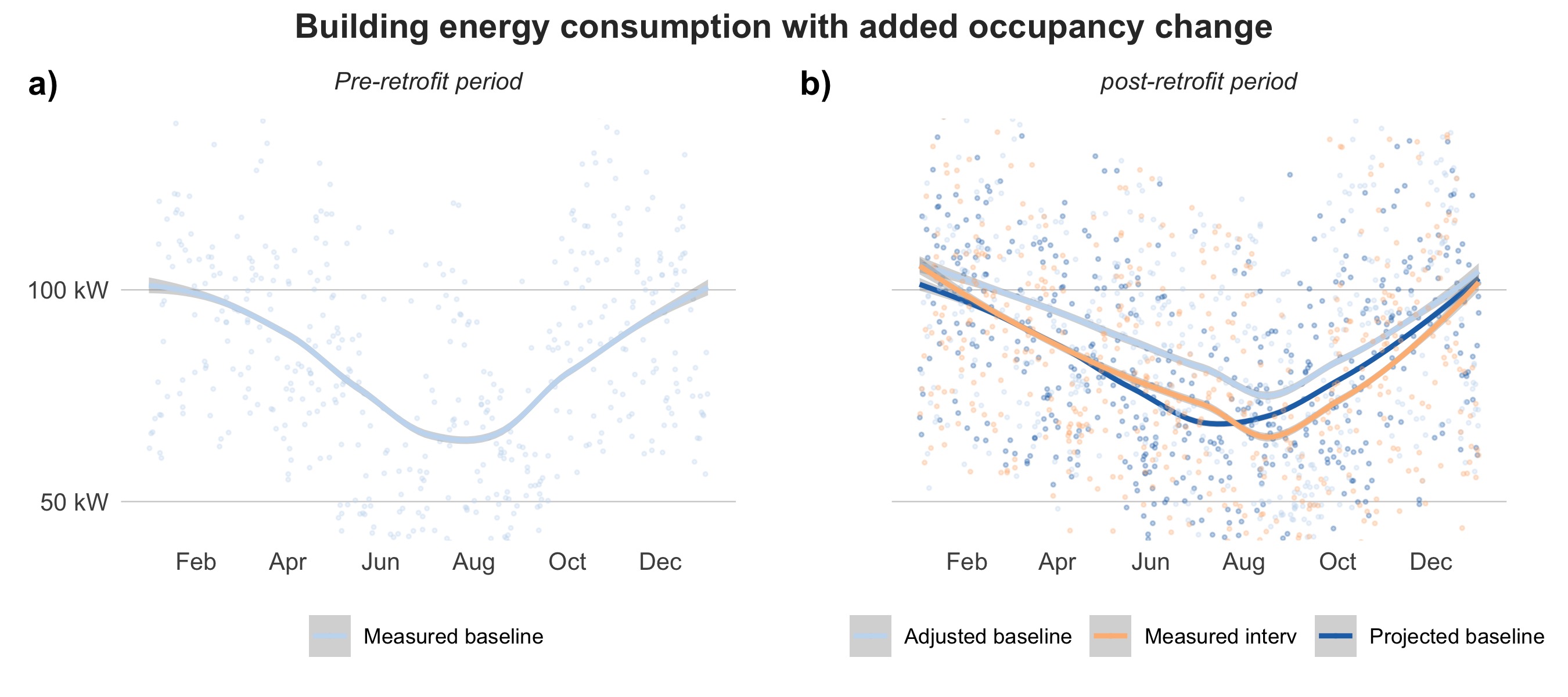


Figure 3.5: Illustration of occupancy change impact on TOWT model fitting and baseline projection in the post-retrofit period

Figure 3.6 shows the overall estimation accuracy between the two M&V methods where the accuracy is indicated as a difference-in-difference value:

Therefore, a positive value indicates that the randomized method provides an estimation more closely aligned with the true target savings. The bar plot shows the difference-in-difference of the calculated fractional savings () for each target building and the dotted line shows the absolute deviation of randomized method (i.e. ) as a reference. In addition, we showed the averaged difference-in-difference metric for each site in the figure, quantifying how much more accurate the randomized method is compared to the conventional method. As a result, across all sites, the randomized method consistently outperforms the conventional method and the values of dotted line show similar results compared to Figure 3.3 further implying that the non-routine event has negligible influence on the randomized method.

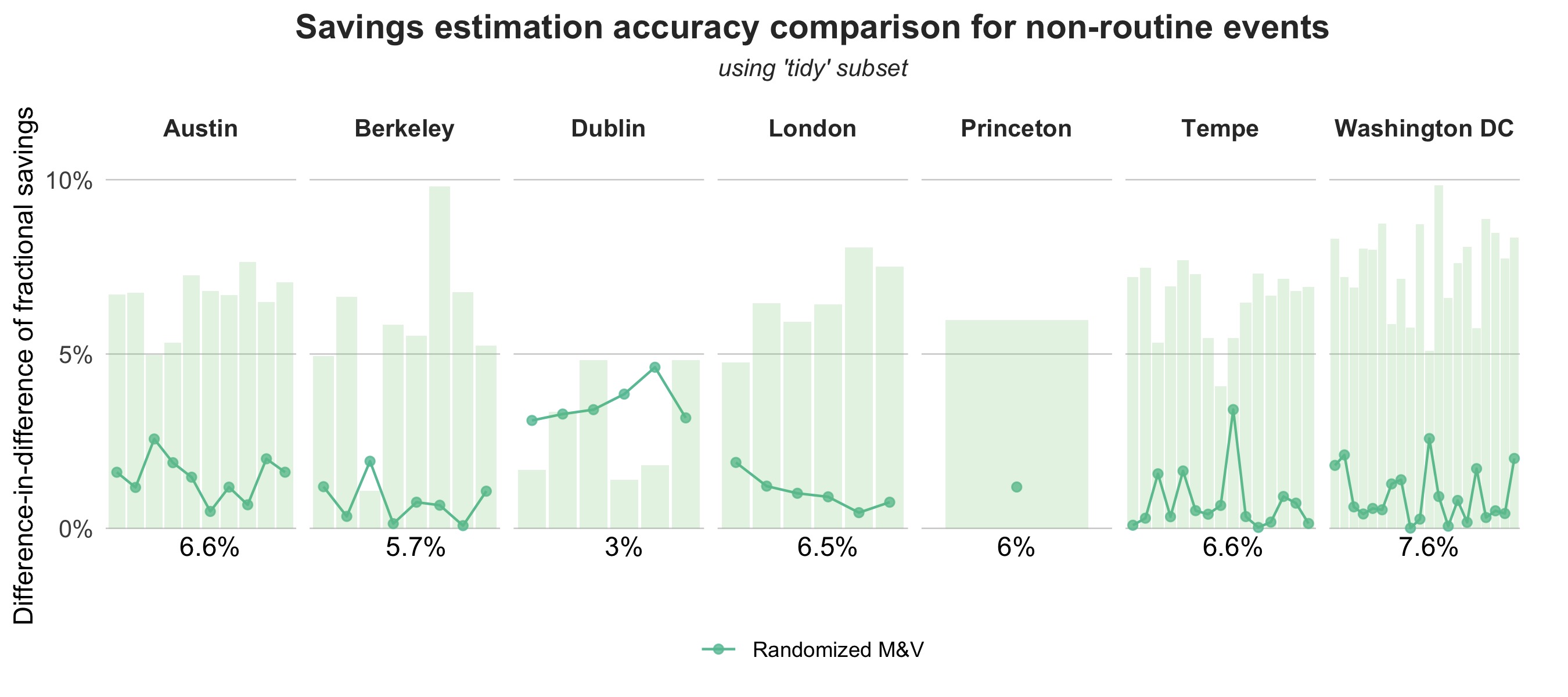


Figure 3.6: Savings estimation accuracy comparison between two M&V methods with added occupancy change (with the average difference-in-difference average on each location displayed at the bottom).

### 3.2.2 No-saving detection

As previously mentioned, the ‘savings’ detected in commercial buildings during the pandemic were largely confounded by null occupancy. In other words, if no intervention is applied or the intervention effect is known to be null (e.g., constantly overridden by the baseline), an ideal M&V method should detect no savings. Given that the messy dataset allows for up to a 25% annual usage change in the original dataset, a more reliable M&V method should inform the analyst that no savings occurred prior to the addition of the intervention. To test this, we removed the chilled water plant reset effect from the sampled days and assessed whether the M&V methods could correctly detect zero savings.

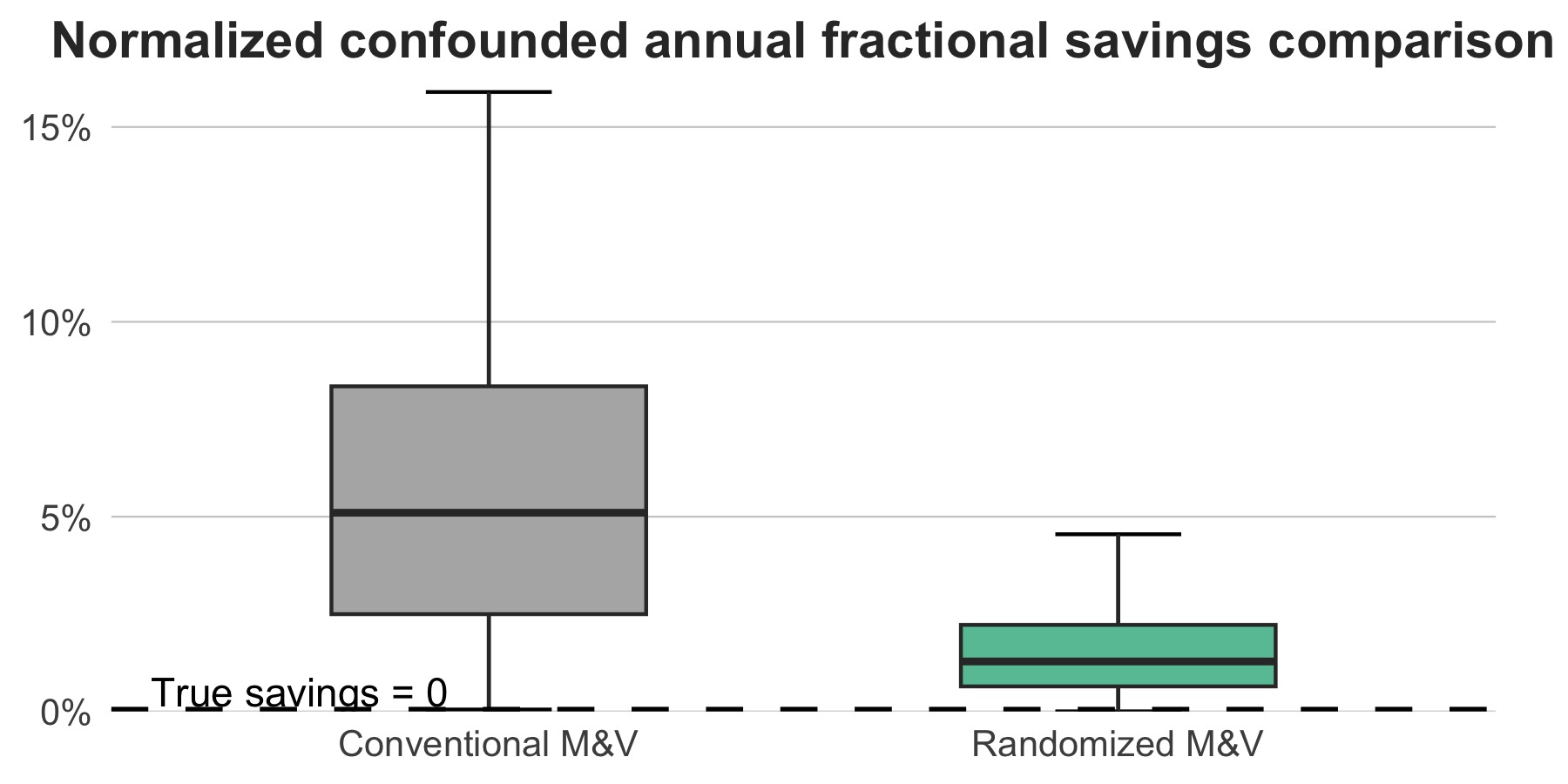


Figure 3.7: Comparison of no intervention effect detection results (with 0% as true annual fractional savings) between conventional and randomized M&V method.

Figure 3.7 shows that the conventional M&V method in average detected 5% savings when there should be none. The randomized method on the other hand is able to provide an estimation much closer to 0% with reduced uncertainty. This improvement is largely due to the randomized sampling approach, where the two strategies (baseline and intervention) are sampled at a 50%/50% ratio. This ensures that any changes or disturbances additional to the intervention implementation, such as concurrent lighting retrofits, are evenly distributed across both strategies, balancing their effects.

# 4 Discussion

## 4.1 TOWT modeling accuracy

To assess modeling accuracy, we used the Coefficient of Variation of Root-Mean Squared Error, or CV(RMSE) as the error metric. Since this metric is calculated as a normalized value, it is useful to compare different model fitting results. Guideline 14 requires that whole-building baseline model fitting accuracy should maintain a CV(RMSE) lower than 30% (ASHRAE 2014). In addition one study focusing on the baseline energy data-driven model fitting indicates that TOWT performs as accurate as other more advanced machine learning models (Granderson et al. 2016) and the calculated CV(RMSE) distribution for a large sample of commercial buildings indicates a median of 20%. Figure 4.1shows the distribution of model fitting accuracy, calculated separately for the two measurement sets. The box plot provides a statistical summary of the accuracy data points, with error bars highlighting any points beyond the range, which should be considered outliers. The results show that the TOWT model’s performance in this study aligns with, and is even slightly better than, the performance reported in the literature. The improved results can be attributed to relatively strict data filtering criteria for pre-processing described in Section 2.

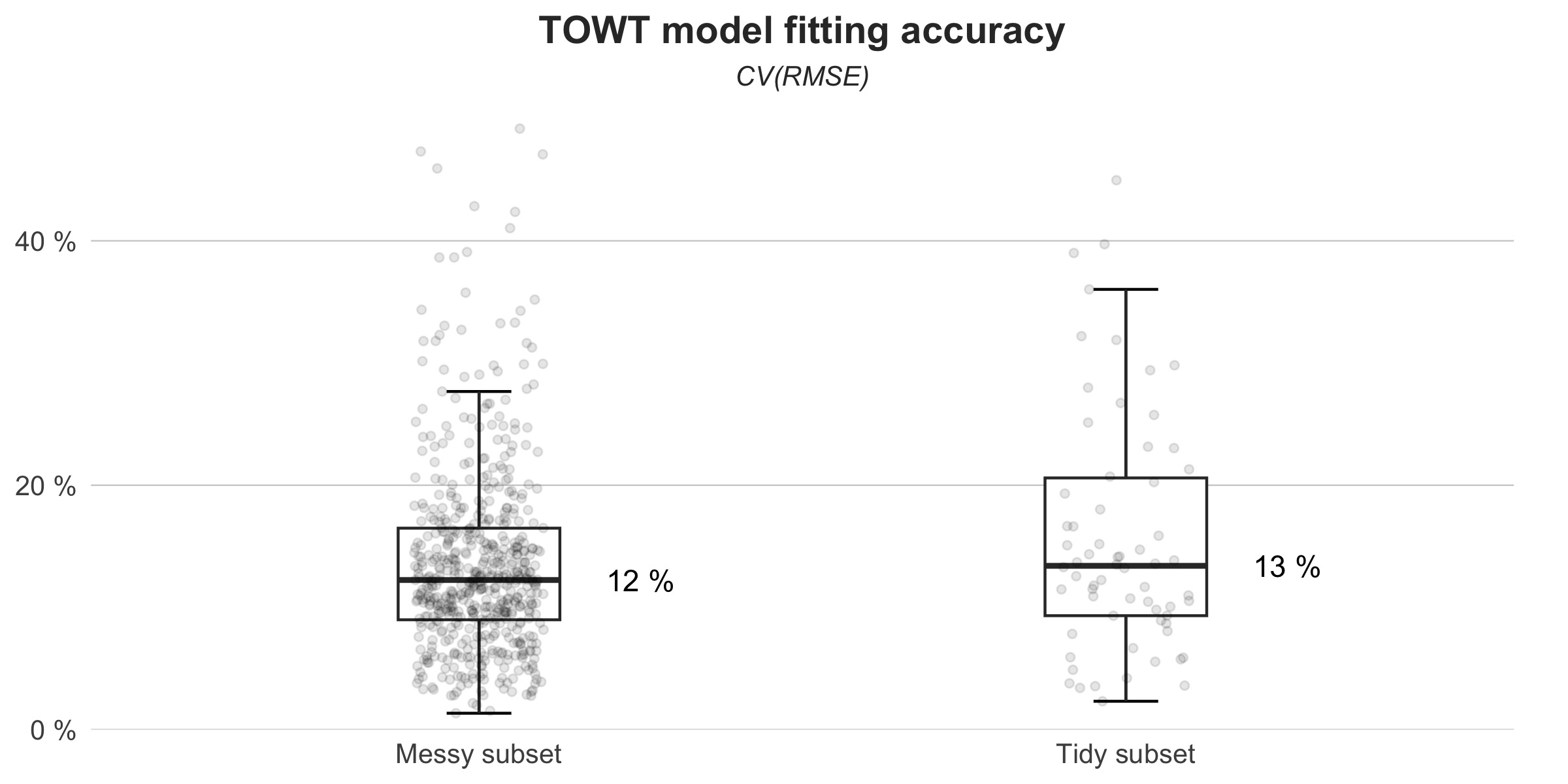


Figure 4.1: TOWT model fitting accuracy distribution for all buildings included (with each data point representing one building)

The limitation of regression models is that despite they capture well on the mean building energy consumption but can underestimate the 15-min daily peak load (Granderson et al. 2021), which is not useful for assessing savings for demand response events.

## 4.2 Sampling ratio for randomized M&V

Another advantage of using randomized M&V is the flexibility of changing sampling ratio after the target savings detected. For instance, the building owner can continue sampling at a 50%/50% ratio between the baseline and the intervention to further reduce the uncertainty associated with the savings. Alternatively, they could switch to 100% intervention to maximize savings on the utility bill, though this approach risks the baseline becoming outdated. A middle-ground approach would be to sample at an 20%/80% ratio between the baseline and the intervention. Through this way, the software-as-a-service company can accurately model future baseline changes and adjust customers’ bill accordinly with minimum baseline days sampled. To demonstrate, figure 4.2 shows after the analyst reports the randomized M&V results indicated in Figure 3.1, a new schedule sampling at 20%/80% was implemented till the end of the year. Similarly, we calculated the accuracy metric as the absolute deviation from the annual true savings.

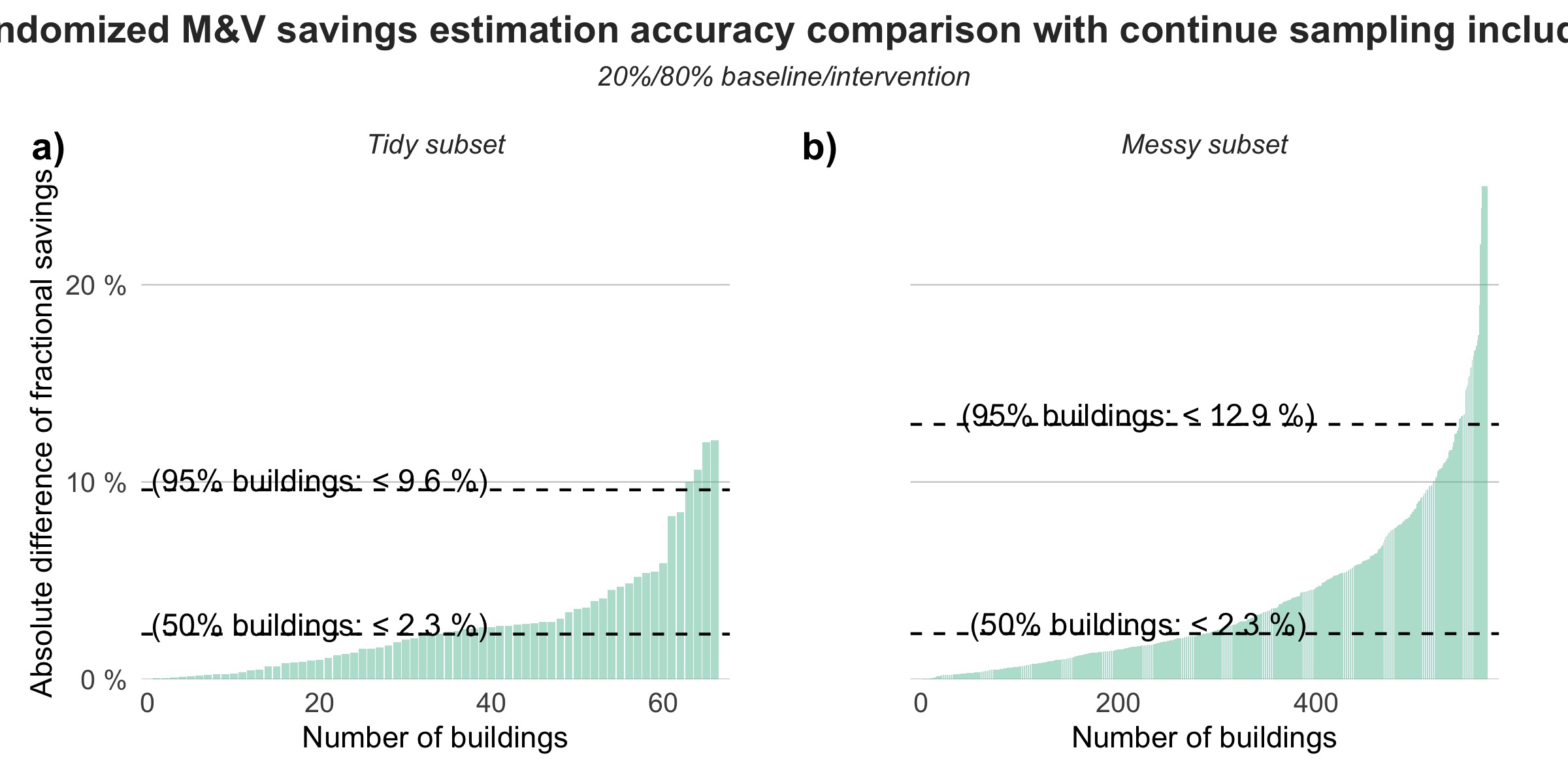


Figure 4.2: Randomized M&V absolute deviation of fractional savings from true target savings for a whole year with more intervention days sampled after all stopping criteria satisfied

The plot shows the deviation remains consistent in the tidy subset comapred to Figure 3.3 but increased slightly in the messy subset compare to to Figure 3.4. With unbalanced sampling ratio, there is a trade-off for estimation accuracy. But compared with conventional method, this is still a preferred approach. Most importantly, this approach allows customers to realize a proportion of expected savings while documenting baseline measurement. For example, opting for an early stop at 24 or 36 weeks with 50% intervention, followed by re-sampling at 80% until the year’s end (~another 20 weeks), would enable customers to capture about 65% of the full-range savings.

## 4.3 Limitations and future study

We identify two key limitations in this study:

1. Application of control intervention: Although the primary focus is on accurately detecting intervention effects rather than validating their broader impact, the simulated intervention remains somewhat generic considering the diverse building types and climate zones in the BDG2 dataset. For simplicity, we applied the parameters listed in Table 1 uniformly across all buildings. Yet, in some cases, raising the water temperature by 1°C might yield more or less than 8% in electricity savings due to target buildings’ demand flexibility.
2. Design of the randomized switchback experiment: In this study, we assumed a daily sampling interval would suffice for most commercial buildings, but exceptions exist. Buildings with significant thermal lag, such as those with heavy concrete construction, hot water tanks, or Thermally Active Building Systems (TABS), may experience carryover effects. In these cases, the impact of one day’s control strategy can influence subsequent measurements due to thermal storage. For example, if chilled water is pre-charged in the thermal mass at the end of a sampled intervention day and the control swicthes to baseline at 12 AM the following day, the analyst would observe lower energy consumption in that sampled baseline day. While this study does not account for carryover effects, we recommend using a 3-day sampling interval and excluding non-consecutive days in practice to ‘wash out’ residual effects from previous strategies.

For future work, our main objective is to extend the application of the proposed randomized M&V approach. For instance, a customer may want to know whether Model Predictive Control (MPC) can reduce energy bills under a dynamic pricing structure or estimate the Return on Investment (ROI) for a retrofit. Or whether a load shift control can save operational cost by shifting load out of a given time window such grid peak price/emissions period. We also believe that this approach could be adapted to analyze the effects of other retrofit interventions at a population level. Additionally, we aim to demonstrate that the randomized method and sequential analysis framework can be applied to various performance metrics, including indoor air quality, operational carbon emissions, and thermal comfort.

# 5 Conclusion

This research demonstrated the application of a novel whole-building measurement and verification (M&V) method, comparing its performance to the conventional approach outlined in ASHRAE Guideline 14 using a large, open-source commercial building dataset. The proposed M&V method leverages the randomized experimental design concept from other scientific fields, along with statistical sequential inference techniques, to determine when target savings are detected. We used a virtual control retrofit case—resetting the chilled water setpoint based on outdoor weather conditions—and applied it to over 500 filtered commercial buildings. By comparing the savings estimations of the conventional method with the novel randomized method, we found that the randomized approach provides faster and more robust savings estimations.

Specifically, we showed that throughout 7 climate zones assessed in this study, the randomized M&V can provide a saving estimation by 36 weeks (with the majority finishes by 24 weeks) once all stopping criteria satisfied. In contrast, the conventional method requires a full range of baseline and intervention measurements under normal operating conditions, typically taking 6-9 months for each phase. Most importantly, Furthermore, we verified that with reduced M&V timeline the randomized method can estimate savings more accurately than the two-year conventional method.

We also evaluated the impact of non-routine events on the proposed M&V method by: 1) introducing a known change, such as occupancy-induced energy reductions, and 2) detecting when no intervention was applied in buildings with marginal energy consumption differences. In both cases, we demonstrated that baseline model fitting could be biased, while randomization effectively blocked confounding effects, ensuring the robustness of savings estimations.

Although the limitation of the method is that it only applies to a subset of all M&V use cases (i.e., strategies that can be switched on and off), we believe its true value lies in the usefulness and conviencience for most control retrofit validation in the field test.

# 6 Acknowledgements

# References

Administration, US Energy Information. 2012. “Commercial Buildings Energy Consumption Survey (CBECS).” US EIA Washington, DC. <https://www.eia.gov/consumption/commercial/data/2003/>.

Aksoy, Cevat Giray, Jose Maria Barrero, Nicholas Bloom, Steven J. Davis, Mathias Dolls, and Pablo Zarate. 2022. “Working from Home Around the World.” *Brookings Papers on Economic Activity* 2022 (2): 281–360. <https://muse.jhu.edu/pub/1/article/901274>.

Alrobaie, Abdurahman, and Moncef Krarti. 2022. “A Review of Data-Driven Approaches for Measurement and Verification Analysis of Building Energy Retrofits.” *Energies* 15 (21): 7824. <https://doi.org/10.3390/en15217824>.

Angus, Derek C., Brian M. Alexander, Scott Berry, Meredith Buxton, Roger Lewis, Melissa Paoloni, Steven A. R. Webb, et al. 2019. “Adaptive Platform Trials: Definition, Design, Conduct and Reporting Considerations.” *Nature Reviews Drug Discovery* 18 (10): 797–807. <https://doi.org/10.1038/s41573-019-0034-3>.

ASHRAE. 2014. “ASHRAE Guideline 14 -2014: Measurement of Energy, Demand and Water Savings.” Atlanta, GA: American Society of Heating, Refrigeration; Air Conditioning Engineers.

Burger, Bram, Marc Vaudel, and Harald Barsnes. 2021. “Importance of Block Randomization When Designing Proteomics Experiments.” *Journal of Proteome Research* 20 (1): 122–28. <https://doi.org/10.1021/acs.jproteome.0c00536>.

Burkhart, Michael C., Yeonsook Heo, and Victor M. Zavala. 2014. “Measurement and Verification of Building Systems Under Uncertain Data: A Gaussian Process Modeling Approach.” *Energy and Buildings* 75 (June): 189–98. <https://doi.org/10.1016/j.enbuild.2014.01.048>.

Cai, Senhong, and Zhonghua Gou. 2024. “Impact of COVID-19 on the Energy Consumption of Commercial Buildings: A Case Study in Singapore.” *Energy and Built Environment* 5 (3): 364–73. <https://doi.org/10.1016/j.enbenv.2022.11.004>.

Chihib, Mehdi, Esther Salmerón-Manzano, Mimoun Chourak, Alberto-Jesus Perea-Moreno, and Francisco Manzano-Agugliaro. 2021. “Impact of the COVID-19 Pandemic on the Energy Use at the University of Almeria (Spain).” *Sustainability* 13 (11): 5843. <https://doi.org/10.3390/su13115843>.

Čongradac, Velimir, and Filip Kulić. 2012. “Recognition of the Importance of Using Artificial Neural Networks and Genetic Algorithms to Optimize Chiller Operation.” *Energy and Buildings* 47 (April): 651–58. <https://doi.org/10.1016/j.enbuild.2012.01.007>.

Demand Side Analytics. 2022. “Population NMEC Control Group Accuracy Assessment.” CALMAC ID: PGE0476.01. <https://www.calmac.org/publications/PGE0476.01.pdf>.

DOE, US. 2008. “M&v Guidelines: Measurement and Verification for Federal Energy Projects Version 3.0.” US Department Of Energy.

Feng, Zhuangbo, and Shi-Jie Cao. 2019. “A Newly Developed Electrostatic Enhanced Pleated Air Filters Towards the Improvement of Energy and Filtration Efficiency.” *Sustainable Cities and Society* 49 (August): 101569. <https://doi.org/10.1016/j.scs.2019.101569>.

Gabler, Nicole B., Naihua Duan, Sunita Vohra, and Richard L. Kravitz. 2011. “N-of-1 Trials in the Medical Literature: A Systematic Review.” *Medical Care* 49 (8): 761768.

Gallagher, Colm V., Kevin Leahy, Peter O’Donovan, Ken Bruton, and Dominic T. J. O’Sullivan. 2018. “Development and Application of a Machine Learning Supported Methodology for Measurement and Verification (m&v) 2.0.” *Energy and Buildings* 167 (May): 8–22. <https://doi.org/10.1016/j.enbuild.2018.02.023>.

Gaspar, K., M. Gangolells, M. Casals, P. Pujadas, N. Forcada, M. Macarulla, and B. Tejedor. 2022. “Assessing the Impact of the COVID-19 Lockdown on the Energy Consumption of University Buildings.” *Energy and Buildings* 257 (February): 111783. <https://doi.org/10.1016/j.enbuild.2021.111783>.

Granderson, Jessica, and Phillip N. Price. 2014. “Development and Application of a Statistical Methodology to Evaluate the Predictive Accuracy of Building Energy Baseline Models.” *Energy* 66 (March): 981–90. <https://doi.org/10.1016/j.energy.2014.01.074>.

Granderson, Jessica, Phillip N. Price, David Jump, Nathan Addy, and Michael D. Sohn. 2015. “Automated Measurement and Verification: Performance of Public Domain Whole-Building Electric Baseline Models.” *Applied Energy* 144 (April): 106–13. <https://doi.org/10.1016/j.apenergy.2015.01.026>.

Granderson, Jessica, Mrinalini Sharma, Eliot Crowe, David Jump, Samuel Fernandes, Samir Touzani, and Devan Johnson. 2021. “Assessment of Model-Based Peak Electric Consumption Prediction for Commercial Buildings.” *Energy and Buildings* 245 (August): 111031. <https://doi.org/10.1016/j.enbuild.2021.111031>.

Granderson, Jessica, Samir Touzani, Claudine Custodio, Michael D. Sohn, David Jump, and Samuel Fernandes. 2016. “Accuracy of Automated Measurement and Verification (m&v) Techniques for Energy Savings in Commercial Buildings.” *Applied Energy* 173 (July): 296–308. <https://doi.org/10.1016/j.apenergy.2016.04.049>.

Gui, Xuechen, Zhonghua Gou, Fan Zhang, and Rongrong Yu. 2021. “The Impact of COVID-19 on Higher Education Building Energy Use and Implications for Future Education Building Energy Studies.” *Energy and Buildings* 251 (November): 111346. <https://doi.org/10.1016/j.enbuild.2021.111346>.

Heo, Y., R. Choudhary, and G. A. Augenbroe. 2012. “Calibration of Building Energy Models for Retrofit Analysis Under Uncertainty.” *Energy and Buildings* 47 (April): 550–60. <https://doi.org/10.1016/j.enbuild.2011.12.029>.

Kang, Hyuna, Jongbaek An, Hakpyeong Kim, Changyoon Ji, Taehoon Hong, and Seunghye Lee. 2021. “Changes in Energy Consumption According to Building Use Type Under COVID-19 Pandemic in South Korea.” *Renewable and Sustainable Energy Reviews* 148 (September): 111294. <https://doi.org/10.1016/j.rser.2021.111294>.

Lee, Kuei-Peng, and Te-Ang Cheng. 2012. “A Simulation–Optimization Approach for Energy Efficiency of Chilled Water System.” *Energy and Buildings* 54 (November): 290–96. <https://doi.org/10.1016/j.enbuild.2012.06.028>.

Lee, Young Jun, Sung Hyup Hong, Jong Man Lee, Yeo Beom Yoon, Jong Min Choi, and Kwang Ho Lee. 2022. “Chilled Water Temperature Set-Point Reset Based on Outdoor Air Temperature and Its Cooling Energy Performance in an Office Building.” *Journal of Mechanical Science and Technology* 36 (3): 1557–68. <https://doi.org/10.1007/s12206-022-0241-4>.

Lim, Chi-Yeon, and Junyong In. 2019. “Randomization in Clinical Studies.” *Korean Journal of Anesthesiology* 72 (3): 221–32. <https://doi.org/10.4097/kja.19049>.

Mantesi, Eirini, Ksenia Chmutina, and Chris Goodier. 2022. “The Office of the Future: Operational Energy Consumption in the Post-Pandemic Era.” *Energy Research & Social Science* 87 (May): 102472. <https://doi.org/10.1016/j.erss.2021.102472>.

Mathieu, Johanna L., Phillip N. Price, Sila Kiliccote, and Mary Ann Piette. 2011. “Quantifying Changes in Building Electricity Use, with Application to Demand Response.” *IEEE Transactions on Smart Grid* 2 (3): 507–18. <https://doi.org/10.1109/TSG.2011.2145010>.

Miller, Clayton, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W. Hobson, Zixiao Shi, and Forrest Meggers. 2020. “The Building Data Genome Project 2, Energy Meter Data from the ASHRAE Great Energy Predictor III Competition.” *Scientific Data* 7 (1): 368. <https://doi.org/10.1038/s41597-020-00712-x>.

Organisation, Efficiency Valuation. 2007. “International Performance Measurement and Verification Protocol.” Efficiency Valuation Organisation.

Qiu, Shunian, Zhenhai Li, Dalian Fan, Ruikai He, Xinghui Dai, and Zhengwei Li. 2022. “Chilled Water Temperature Resetting Using Model-Free Reinforcement Learning: Engineering Application.” *Energy and Buildings* 255 (January): 111694. <https://doi.org/10.1016/j.enbuild.2021.111694>.

Roa, Carlos Duarte, Paul Raftery, Anand Prakash, and Therese Peffer. 2023. “Field Demonstration of the Brick Ontology to Scale up the Deployment of ASHRAE Guideline 36 Control Sequences,” 9.

Walter, Travis, Phillip N. Price, and Michael D. Sohn. 2014. “Uncertainty Estimation Improves Energy Measurement and Verification Procedures.” *Applied Energy* 130 (October): 230–36. <https://doi.org/10.1016/j.apenergy.2014.05.030>.

Wiley, Henry E., Darby J. S. Thompson, Clare Bailey, Emily Y. Chew, Catherine A. Cukras, Glenn J. Jaffe, Richard W. J. Lee, et al. 2016. “A Crossover Design for Comparative Efficacy: A 36-Week Randomized Trial of Bevacizumab and Ranibizumab for Diabetic Macular Edema.” *Ophthalmology* 123 (4): 841–49. <https://doi.org/10.1016/j.ophtha.2015.11.021>.

Xie, Xiang, Qiuchen Lu, Manuel Herrera, Qiaojun Yu, Ajith Kumar Parlikad, and Jennifer Mary Schooling. 2021. “Does Historical Data Still Count? Exploring the Applicability of Smart Building Applications in the Post-Pandemic Period.” *Sustainable Cities and Society* 69 (June): 102804. <https://doi.org/10.1016/j.scs.2021.102804>.

Zhai, Zhiqiang (John), and Stephen Nathaniel Johnson. 2017. “Full-Scale Laboratory Test on Energy Dependence on Pressure Drops in HVAC Systems.” *Procedia Engineering*, 10th international symposium on heating, ventilation and air conditioning, ISHVAC2017, 19-22 october 2017, jinan, china, 205 (January): 2133–40. <https://doi.org/10.1016/j.proeng.2017.10.138>.

1. Center for the Built Environment, University of California Berkeley, USA [↑](#footnote-ref-20)